User's Guide III

Regression and Curve Fitting

NCSS
Statistical System

Published by
NCSS
Dr. Jerry L. Hintze
Kaysville, Utah

NCSS User’s Guide III

Direct inquiries to:

NCSS
329 North 1000 East
Kaysville, Utah 84037
Phone (801) 546-0445
Fax (801) 546-3907
Email: support@ncss.com

Warning:

This software and manual are both protected by U.S. Copyright Law (Title 17 United States Code). Unauthorized reproduction and/or sales may result in imprisonment of up to one year and fines of up to $10,000 (17 USC 506). Copyright infringers may also be subject to civil liability.

NCSS License Agreement

Important: The enclosed Number Cruncher Statistical System (NCSS) is licensed by Dr. Jerry L. Hintze to customers for their use only on the terms set forth below. Purchasing the system indicates your acceptance of these terms.

1. **LICENSE.** Dr. Jerry L. Hintze hereby agrees to grant you a non-exclusive license to use the accompanying NCSS program subject to the terms and restrictions set forth in this License Agreement.

2. **COPYRIGHT.** NCSS and its documentation are copyrighted. You may not copy or otherwise reproduce any part of NCSS or its documentation, except that you may load NCSS into a computer as an essential step in executing it on the computer and make backup copies for your use on the same computer.

3. **BACKUP POLICY.** NCSS may be backed up by you for your use on the same machine for which NCSS was purchased.

4. **RESTRICTIONS ON USE AND TRANSFER.** The original and any backup copies of NCSS and its documentation are to be used only in connection with a single computer. You may physically transfer NCSS from one computer to another, provided that NCSS is used in connection with only one computer at a time. You may not transfer NCSS electronically from one computer to another over a network. You may not distribute copies of NCSS or its documentation to others. You may transfer this license together with the original and all backup copies of NCSS and its documentation, provided that the transferee agrees to be bound by the terms of this License Agreement. Neither NCSS nor its documentation may be modified or translated without written permission from Dr. Jerry L. Hintze.

   *You may not use, copy, modify, or transfer **NCSS**, or any copy, modification, or merged portion, in whole or in part, except as expressly provided for in this license.*

5. **NO WARRANTY OF PERFORMANCE.** Dr. Jerry L. Hintze does not and cannot warrant the performance or results that may be obtained by using NCSS. Accordingly, NCSS and its documentation are licensed "as is" without warranty as to their performance, merchantability, or fitness for any particular purpose. The entire risk as to the results and performance of NCSS is assumed by you. Should NCSS prove defective, you (and not Dr. Jerry L. Hintze nor his dealers) assume the entire cost of all necessary servicing, repair, or correction.

6. **LIMITED WARRANTY ON CD.** To the original licensee only, Dr. Jerry L. Hintze warrants the medium on which NCSS is recorded to be free from defects in materials and faulty workmanship under normal use and service for a period of ninety days from the date NCSS is delivered. If, during this ninety-day period, a defect in a CD should occur, the CD may be returned to Dr. Jerry L. Hintze at his address, or to the dealer from which NCSS was purchased, and NCSS will replace the CD without charge to you, provided that you have sent a copy of your receipt for NCSS. Your sole and exclusive remedy in the event of a defect is expressly limited to the replacement of the CD as provided above.

   Any implied warranties of merchantability and fitness for a particular purpose are limited in duration to a period of ninety (90) days from the date of delivery. If the failure of a CD has resulted from accident, abuse, or misapplication of the CD, Dr. Jerry L. Hintze shall have no responsibility to replace the CD under the terms of this limited warranty. This limited warranty gives you specific legal rights, and you may also have other rights, which vary from state to state.

7. **LIMITATION OF LIABILITY.** Neither Dr. Jerry L. Hintze nor anyone else who has been involved in the creation, production, or delivery of NCSS shall be liable for any direct, incidental, or consequential damages, such as, but not limited to, loss of anticipated profits or benefits, resulting from the use of NCSS or arising out of any breach of any warranty. Some states do not allow the exclusion or limitation of direct, incidental, or consequential damages, so the above limitation may not apply to you.

8. **TERM.** The license is effective until terminated. You may terminate it at any time by destroying NCSS and documentation together with all copies, modifications, and merged portions in any form. It will also terminate if you fail to comply with any term or condition of this License Agreement. You agree upon such termination to destroy NCSS and documentation together with all copies, modifications, and merged portions in any form.

9. **YOUR USE OF NCSS ACKNOWLEDGES** that you have read this customer license agreement and agree to its terms. You further agree that the license agreement is the complete and exclusive statement of the agreement between us and supersedes any proposal or prior agreement, oral or written, and any other communications between us relating to the subject matter of this agreement.

Dr. Jerry L. Hintze & **NCSS**, Kaysville, Utah

Preface

Number Cruncher Statistical System (**NCSS**) is an advanced, easy-to-use statistical analysis software package. The system was designed and written by Dr. Jerry L. Hintze over the last several years. Dr. Hintze drew upon his experience both in teaching statistics at the university level and in various types of statistical consulting.

The present version, written for 32-bit versions of Microsoft Windows (95, 98, ME, 2000, NT, etc.) computer systems, is the result of several iterations. Experience over the years with several different types of users has helped the program evolve into its present form.

Statistics is a broad, rapidly developing field. Updates and additions are constantly being made to the program. If you would like to be kept informed about updates, additions, and corrections, please send your name, address, and phone number to:

User Registration
NCSS
329 North 1000 East
Kaysville, Utah 84037

or Email you name, address, and phone number to:

Sales@NCSS.COM

NCSS maintains a website at **WWW.NCSS.COM** where we make the latest edition of NCSS available for free downloading. The software is password protected, so only users with valid serial numbers may use this downloaded edition. We hope that you will download the latest edition routinely and thus avoid any bugs that have been corrected since you purchased your copy.

NCSS maintains the following program and documentation copying policy. Copies are limited to a one person / one machine basis for “BACKUP” purposes only. You may make as many backup copies as you wish. Further distribution constitutes a violation of this copy agreement and will be prosecuted to the fullest extent of the law.

NCSS is not “copy protected.” You may freely load the program onto your hard disk. We have avoided copy protection in order to make the system more convenient for you. Please honor our good faith (and low price) by avoiding the temptation to distribute copies to friends and associates.

We believe this to be an accurate, exciting, easy-to-use system. If you find any portion that you feel needs to be changed, please let us know. Also, we openly welcome suggestions for additions to the system.

User’s Guide III
Table of Contents

Regression

Linear and Multiple Regression

Variable Selection

Other Regression Routines

Cox Regression is found in User’s Guide V in the Survival/Reliability section

Curve Fitting

Curve Fitting

Ratio of Polynomials

Other

References and Indices

User's Guide I
Table of Contents

Quick Start & Self Help

Introduction

Data

Tools

Graphics

Introduction

Single-Variable Charts

Two-Variable Charts (Discrete / Continuous)

Two-Variable Charts (Both Continuous)

Three-Variable Charts

Settings Windows

References and Indices

User’s Guide II
Table of Contents

Descriptive Statistics

Means

T-Tests

Analysis of Variance

Mixed Models

Other

Quality Control

Design of Experiments

References and Indices

User’s Guide IV
Table of Contents

Multivariate Analysis

Clustering

Meta-Analysis

Forecasting / Time Series

Exponential Smoothing

Time Series Analysis

Operations Research

Mass Appraisal

References and Indices

User's Guide V

Table of Contents

Tabulation

Item Analysis

Proportions

Diagnostic Tests

Binary Diagnostic Tests

ROC Curves

Survival / Reliability

References and Indices

Chapter 300

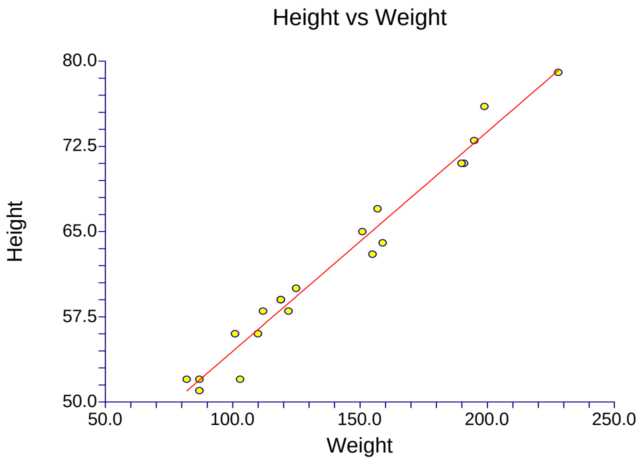
Linear Regression and Correlation

Introduction

Linear Regression refers to a group of techniques for fitting and studying the straight-line relationship between two variables. Linear regression estimates the regression coefficients β_0 and β_1 in the equation

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j$$

where X is the independent variable, Y is the dependent variable, β_0 is the Y *intercept*, β_1 is the *slope*, and ε is the error.



In order to calculate confidence intervals and hypothesis tests, it is assumed that the errors are independent and normally distributed with mean zero and variance σ^2.

Given a sample of N observations on X and Y, the method of least squares estimates β_0 and β_1 as well as various other quantities that describe the precision of the estimates and the goodness-of-fit of the straight line to the data. Since the estimated line will seldom fit the data exactly, a term for the discrepancy between the actual and fitted data values must be added. The equation then becomes

$$\begin{aligned} y_j &= b_0 + b_1 x_j + e_j \\ &= \hat{y}_j + e_j \end{aligned}$$

where *j* is the observation (row) number, b_0 estimates β_0, b_1 estimates β_1, and e_j is the discrepancy between the actual data value y_j and the fitted value given by the regression equation, which is often referred to as $\hat{y}_j$. This discrepancy is usually referred to as the *residual*.

Note that the linear regression equation is a mathematical model describing the relationship between *X* and *Y*. In most cases, we do not believe that the model defines the exact relationship between the two variables. Rather, we use it as an approximation to the exact relationship. Part of the analysis will be to determine how close the approximation is.

Also note that the equation predicts *Y* from *X*. The value of *Y* depends on the value of *X*. The influence of all other variables on the value of *Y* is lumped into the residual.

Correlation

Once the intercept and slope have been estimated using least squares, various indices are studied to determine the reliability of these estimates. One of the most popular of these reliability indices is the *correlation coefficient*. The correlation coefficient, or simply the *correlation*, is an index that ranges from -1 to 1. When the value is near zero, there is no linear relationship. As the correlation gets closer to plus or minus one, the relationship is stronger. A value of one (or negative one) indicates a perfect linear relationship between two variables.

Actually, the strict interpretation of the correlation is different from that given in the last paragraph. The correlation is a parameter of the bivariate normal distribution. This distribution is used to describe the association between two variables. This association does not include a cause and effect statement. That is, the variables are not labeled as dependent and independent. One does not depend on the other. Rather, they are considered as two random variables that seem to vary together. The important point is that in linear regression, *Y* is assumed to be a random variable and *X* is assumed to be a fixed variable. In correlation analysis, both *Y* and *X* are assumed to be random variables.

Possible Uses of Linear Regression Analysis

Montgomery (1982) outlines the following four purposes for running a regression analysis.

Description

The analyst is seeking to find an equation that describes or summarizes the relationship between two variables. This purpose makes the fewest assumptions.

Coefficient Estimation

This is a popular reason for doing regression analysis. The analyst may have a theoretical relationship in mind, and the regression analysis will confirm this theory. Most likely, there is specific interest in the magnitudes and signs of the coefficients. Frequently, this purpose for regression overlaps with others.

Prediction

The prime concern here is to predict the response variable, such as sales, delivery time, efficiency, occupancy rate in a hospital, reaction yield in some chemical process, or strength of some metal. These predictions may be very crucial in planning, monitoring, or evaluating some process or system. There are many assumptions and qualifications that must be made in this case. For instance, you must not extrapolate beyond the range of the data. Also, interval estimates require that normality assumptions to hold.

Control

Regression models may be used for monitoring and controlling a system. For example, you might want to calibrate a measurement system or keep a response variable within certain guidelines. When a regression model is used for control purposes, the independent variable must be related to the dependent variable in a causal way. Furthermore, this functional relationship must continue over time. If it does not, continual modification of the model must occur.

Assumptions

The following assumptions must be considered when using linear regression analysis.

Linearity

Linear regression models the straight-line relationship between *Y* and *X*. Any curvilinear relationship is ignored. This assumption is most easily evaluated by using a scatter plot. This should be done early on in your analysis. Nonlinear patterns can also show up in residual plot. A lack of fit test is also provided.

Constant Variance

The variance of the residuals is assumed to be constant for all values of *X*. This assumption can be detected by plotting the residuals versus the independent variable. If these residual plots show a rectangular shape, we can assume constant variance. On the other hand, if a residual plot shows an increasing or decreasing wedge or bowtie shape, nonconstant variance (*heteroscedasticity*) exists and must be corrected.

The corrective action for nonconstant variance is to use weighted linear regression or to transform either *Y* or *X* in such a way that variance is more nearly constant. The most popular *variance stabilizing transformation* is the to take the logarithm of *Y*.

Special Causes

It is assumed that all special causes, outliers due to one-time situations, have been removed from the data. If not, they may cause nonconstant variance, nonnormality, or other problems with the regression model. The existence of outliers is detected by considering scatter plots of *Y* and *X* as well as the residuals versus *X*. Outliers show up as points that do not follow the general pattern.

Normality

When hypothesis tests and confidence limits are to be used, the residuals are assumed to follow the normal distribution.

Independence

The residuals are assumed to be uncorrelated with one another, which implies that the *Y's* are also uncorrelated. This assumption can be violated in two ways: model misspecification or time-sequenced data.

1. *Model misspecification.* If an important independent variable is omitted or if an incorrect functional form is used, the residuals may not be independent. The solution to this dilemma is to find the proper functional form or to include the proper independent variables and use multiple regression.

2. *Time-sequenced data*. Whenever regression analysis is performed on data taken over time, the residuals may be correlated. This correlation among residuals is called *serial correlation*. Positive serial correlation means that the residual in time period *j* tends to have the same sign as the residual in time period (*j* - *k*), where *k* is the lag in time periods. On the other hand, negative serial correlation means that the residual in time period *j* tends to have the opposite sign as the residual in time period (*j* - *k*).

The presence of serial correlation among the residuals has several negative impacts.

1. The regression coefficients remain unbiased, but they are no longer efficient, i.e., minimum variance estimates.

2. With positive serial correlation, the mean square error may be seriously underestimated. The impact of this is that the standard errors are underestimated, the *t*-tests are inflated (show significance when there is none), and the confidence intervals are shorter than they should be.

3. Any hypothesis tests or confidence limits that require the use of the *t* or *F* distribution are invalid.

You could try to identify these serial correlation patterns informally, with the residual plots versus time. A better analytical way would be to use the Durbin-Watson test to assess the amount of serial correlation.

Technical Details

Regression Analysis

This section presents the technical details of least squares regression analysis using a mixture of summation and matrix notation. Because this module also calculates weighted linear regression, the formulas will include the weights, w_j. When weights are not used, the w_j are set to one.

Define the following vectors and matrices.

$$\mathbf{Y}=\begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_N \end{bmatrix},\ \mathbf{X}=\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_j \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix},\ \mathbf{e}=\begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_N \end{bmatrix},\ \mathbf{1}=\begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix},\ \mathbf{b}=\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & \vdots \\ 0 & 0 & w_j & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & w_N \end{bmatrix}$$

Least Squares

Using this notation, the least squares estimates are found using the equation.

$$\mathbf{b} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WY}$$

Note that when the weights are not used, this reduces to

$$\mathbf{b} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'Y}$$

The predicted values of the dependent variable are given by

$$\hat{\mathbf{Y}} = \mathbf{b'X}$$

The residuals are calculated using

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Estimated Variances

An estimate of the variance of the residuals is computed using

$$s^2 = \frac{\mathbf{e'We}}{N-2}$$

An estimate of the variance of the regression coefficients is calculated using

$$\begin{aligned} \mathrm{V}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} &= \begin{pmatrix} s_{b_0}^2 & s_{b_0 b_1} \\ s_{b_0 b_1} & s_{b_1}^2 \end{pmatrix} \\ &= s^2\left(\mathbf{X'WX}\right)^{-1} \end{aligned}$$

An estimate of the variance of the predicted mean of Y at a specific value of X, say X_0, is given by

$$s_{Y_m|X_0}^2 = s^2\left(1, X_0\right)\left(\mathbf{X'WX}\right)^{-1}\begin{pmatrix} 1 \\ X_0 \end{pmatrix}$$

An estimate of the variance of the predicted value of Y for an individual for a specific value of X, say X_0, is given by

$$s_{Y_I|X_0}^2 = s^2 + s_{Y_m|X_0}^2$$

Hypothesis Tests of the Intercept and Slope

Using these variance estimates and assuming the residuals are normally distributed, hypothesis tests may be constructed using the Student's *t* distribution with *N* - 2 degrees of freedom using

$$t_{b_0} = \frac{b_0 - \mathrm{B}_0}{s_{b_0}}$$

and

$$t_{b_1} = \frac{b_1 - \mathrm{B}_1}{s_{b_1}}$$

Usually, the hypothesized values of B_0 and B_1 are zero, but this does not have to be the case.

Confidence Intervals of the Intercept and Slope

A $100(1-\alpha)\%$ confidence interval for the intercept, β_0, is given by

$$b_0 \pm t_{1-\alpha/2,N-2} s_{b_0}$$

A $100(1-\alpha)\%$ confidence interval for the slope, β_1, is given by

$$b_1 \pm t_{1-\alpha/2,N-2} s_{b_1}$$

Confidence Interval of Y for Given X

A $100(1-\alpha)\%$ confidence interval for the mean of *Y* at a specific value of *X*, say X_0, is given by

$$b_0 + b_1 X_0 \pm t_{1-\alpha/2,N-2} s_{Y_m|X_0}$$

Note that this confidence interval assumes that the sample size at *X* is *N*.

A $100(1-\alpha)\%$ prediction interval for the value of *Y* for an individual at a specific value of *X*, say X_0, is given by

$$b_0 + b_1 X_0 \pm t_{1-\alpha/2,N-2} s_{Y_I|X_0}$$

Working-Hotelling Confidence Band for the Mean of Y

A $100(1-\alpha)\%$ simultaneous confidence band for the mean of *Y* at all values of *X* is given by

$$b_0 + b_1 X \pm s_{Y_m|X} \sqrt{2F_{1-\alpha/2,2,N-2}}$$

This confidence band applies to all possible values of *X*. The confidence coefficient, $100(1-\alpha)\%$, is the percent of a long series of samples for which this band covers the entire line for all values of *X* from negativity infinity to positive infinity.

Confidence Interval of X for Given Y

This type of analysis is called *inverse prediction* or *calibration*. A $100(1-\alpha)\%$ confidence interval for the mean value of *X* for a given value of *Y* is calculated as follows. First, calculate *X* from *Y* using

$$\hat{X} = \frac{Y - b_0}{b_1}$$

Then, calculate the interval using

$$\frac{\left(\hat{X} - g\overline{X}\right) \pm A\sqrt{\frac{(1-g)}{N} + \frac{\left(\hat{X} - \overline{X}\right)^2}{\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)}}}{1-g}$$

where

$$A = \frac{t_{1-\alpha/2,N-2}s}{b_1}$$

$$g = \frac{A^2}{\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)}$$

A $100(1-\alpha)\%$ confidence interval for an individual value of *X* for a given value of *Y* is

$$\frac{\left(\hat{X} - g\overline{X}\right) \pm A\sqrt{\frac{(N+1)(1-g)}{N} + \frac{\left(\hat{X} - \overline{X}\right)^2}{\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)}}}{1-g}$$

R-Squared (Percent of Variation Explained)

Several measures of the goodness-of-fit of the regression model to the data have been proposed, but by far the most popular is R^2. R^2 is the square of the correlation coefficient. It is the proportion of the variation in *Y* that is accounted by the variation in *X*. R^2 varies between zero (no linear relationship) and one (perfect linear relationship).

R^2, officially known as the coefficient of determination, is defined as the sum of squares due to the regression divided by the adjusted total sum of squares of *Y*. The formula for R^2 is

$$R^2 = 1 - \left(\frac{\mathbf{e'We}}{\mathbf{Y'WY} - \frac{(\mathbf{1'WY})^2}{\mathbf{1'W1}}} \right)$$

$$= \frac{SS_{Model}}{SS_{Total}}$$

R^2 is probably the most popular measure of how well a regression model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Additional independent variables*. It is possible to increase R^2 by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This usually happens when the sample size is small.

2. *Range of the independent variable*. R^2 is influenced by the range of the independent variable. R^2 increases as the range of *X* increases and decreases as the range of the *X* decreases.

3. *Slope magnitudes*. R^2 does not measure the magnitude of the slopes.

4. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *X* and *Y* was a perfect circle. Although there is a perfect relationship between the variables, the R^2 value would be zero.

5. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.

6. *No-intercept model*. The definition of R^2 assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of R^2 changes dramatically. The fact that your R^2 value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying definition of R^2.

7. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Rbar-Squared (Adjusted R-Squared)

R^2 varies directly with *N*, the sample size. In fact, when *N* = 2, $R^2 = 1$. Because R^2 is so closely tied to the sample size, an adjusted R^2 value, called $\overline{R}^2$, has been developed. $\overline{R}^2$ was developed to minimize the impact of sample size. The formula for $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \left[\frac{\left(N - \left(p-1\right)\right)\left(1 - R^2\right)}{N - p} \right]$$

where *p* is 2 if the intercept is included in the model and 1 if not.

Probability Ellipse

When both variables are random variables and they follow the bivariate normal distribution, it is possible to construct a probability ellipse for them (see Jackson (1991) page 342). The equation of the $100(1-\alpha)\%$ probability ellipse is given by those values of *X* and *Y* that are solutions of

$$T^2_{2,N-2,\alpha} = \frac{s_{YY}s_{XX}}{s_{YY}s_{XX} - s^2_{XY}}\left[\frac{\left(X - \overline{X}\right)^2}{s_{XX}} + \frac{\left(Y - \overline{Y}\right)^2}{s_{YY}} - \frac{2s_{XY}\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{s_{XX}s_{YY}}\right]$$

Orthogonal Regression Line

The least squares estimates discussed above minimize the sum of the squared distances between the *Y*'s and there predicted values. In some situations, both variables are random variables and it is arbitrary which is designated as the dependent variable and which is the independent variable. When the choice of which variable is the dependent variable is arbitrary, you may want to use the *orthogonal regression line* rather than the least squares regression line. The orthogonal regression line minimizes the sum of the squared perpendicular distances from the each observation to the regression line. The orthogonal regression line is the first principal component when a principal components analysis is run on the two variables.

Jackson (1991) page 343 gives a formula for computing the orthogonal regression line without computing a principal components analysis. The slope is given by

$$b_{ortho,1} = \frac{s_{YY} - s_{XX} + \sqrt{s_{YY} - s_{XX} + 4s^2_{XY}}}{2s_{XY}}$$

where

$$s_{XY} = \frac{\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)\left(Y_j - \overline{Y}\right)}{N - 1}$$

The estimate of the intercept is then computed using

$$b_{ortho,y} = \overline{Y} - b_{ortho,1}\overline{X}$$

Although Jackson gives formulas for a confidence interval on the slope and intercept, we do not provide them in ***NCSS*** because their properties are not well understood and the require certain bivariate normal assumptions. Instead, ***NCSS*** provides bootstrap confidence intervals for the slope and intercept.

The Correlation Coefficient

The correlation coefficient can be interpreted in several ways. Here are some of the interpretations.

1. If both *Y* and *X* are standardized by subtracting their means and dividing by their standard deviations, the correlation is the slope of the regression of the standardized *Y* on the standardized *X*.
2. The correlation is the standardized covariance between *Y* and *X*.
3. The correlation is the geometric average of the slopes of the regressions of *Y* on *X* and of *X* on *Y*.
4. The correlation is the square root of *R*-squared, using the sign from the slope of the regression of *Y* on *X*.

The corresponding formulas for the calculation of the correlation coefficient are

$$
\begin{aligned}
r &= \frac{\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)\left(Y_j - \overline{Y}\right)}{\sqrt{\left[\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)^2\right]\left[\sum_{j=1}^{N} w_j\left(Y_j - \overline{Y}\right)^2\right]}} \\
&= \frac{s_{XY}}{\sqrt{s_{XX}s_{YY}}} \\
&= \pm\sqrt{b_{YX}b_{XY}} \\
&= \operatorname{sign}\left(b_{YX}\right)\sqrt{R^2}
\end{aligned}
$$

where s_{XY} is the covariance between *X* and *Y*, b_{XY} is the slope from the regression of *X* on *Y*, and b_{YX} is the slope from the regression of *Y* on *X*. s_{XY} is calculated using the formula

$$
s_{XY} = \frac{\sum_{j=1}^{N} w_j\left(X_j - \overline{X}\right)\left(Y_j - \overline{Y}\right)}{N-1}
$$

The *population correlation coefficient,* ρ, is defined for two random variables, *U* and *W*, as follows

$$
\begin{aligned}
\rho &= \frac{\sigma_{UW}}{\sqrt{\sigma_U^2\sigma_W^2}} \\
&= \frac{\mathrm{E}\left[\left(U-\mu_U\right)\left(W-\mu_W\right)\right]}{\sqrt{\mathrm{Var}(U)\mathrm{Var}(W)}}
\end{aligned}
$$

Note that this definition does not refer to one variable as dependent and the other as independent. Rather, it simply refers to two random variables.

Facts about the Correlation Coefficient

The correlation coefficient has the following characteristics.

1. The range of *r* is between -1 and 1, inclusive.
2. If *r* = 1, the observations fall on a straight line with positive slope.
3. If *r* = -1, the observations fall on a straight line with negative slope.
4. If *r* = 0, there is no linear relationship between the two variables.
5. *r* is a measure of the linear (straight-line) association between two variables.
6. The value of *r* is unchanged if either *X* or *Y* is multiplied by a constant or if a constant is added.

7. The physical meaning of *r* is mathematically abstract and may not be very help. However, we provide it for completeness. The correlation is the cosine of the angle formed by the intersection of two vectors in *N*-dimensional space. The components of the first vector are the values of *X* while the components of the second vector are the corresponding values of *Y*. These components are arranged so that the first dimension corresponds to the first observation, the second dimension corresponds to the second observation, and so on.

Hypothesis Tests for the Correlation

You may be interested in testing hypotheses about the population correlation coefficient, such as $\rho = \rho_0$. When $\rho_0 = 0$, the test is identical to the *t*-test used to test the hypothesis that the slope is zero. The test statistic is calculated using

$$t_{N-2} = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$

However, when $\rho_0 \neq 0$, the test is different from the corresponding test that the slope is a specified, nonzero, value.

NCSS provides two methods for testing whether the correlation is equal to a specified, nonzero, value.

Method 1. This method uses the distribution of the correlation coefficient. Under the null hypothesis that $\rho = \rho_0$ and using the distribution of the sample correlation coefficient, the likelihood of obtaining the sample correlation coefficient, *r*, can be computed. This likelihood is the statistical significance of the test. This method requires the assumption that the two variables follow the bivariate normal distribution.

Method 2. This method uses the fact that Fisher's *z* transformation, given by

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

is closely approximated by a normal distribution with mean

$$\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

and variance

$$\frac{1}{N-3}$$

To test the hypothesis that $\rho = \rho_0$, you calculate *z* using

$$z = \frac{F(r) - F(\rho_0)}{\sqrt{\frac{1}{N-3}}}$$

$$= \frac{\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{2\sqrt{\frac{1}{N-3}}}$$

and use the fact that *z* is approximately distributed as the standard normal distribution with mean equal to zero and variance equal to one. This method requires two assumptions. First, that the two variables follow the bivariate normal distribution. Second, that the distribution of *z* is approximated by the standard normal distribution.

This method has become popular because it uses the commonly available normal distribution rather than the obscure correlation distribution. However, because it makes an additional assumption, it is not as accurate as is method 1. In fact, we have included in for completeness, but recommend the use of Method 1.

Confidence Intervals for the Correlation

A $100(1-\alpha)\%$ confidence interval for ρ may be constructed using either of the two hypothesis methods described above. The confidence interval is calculated by finding, either directly using Method 2 or by a search using Method 1, all those values of ρ_0 for which the hypothesis test is not rejected. This set of values becomes the confidence interval.

Be careful not to make the common mistake in assuming that this confidence interval is related to a transformation of the confidence interval on the slope β_1. The two confidence intervals are not simple transformations of each other.

Spearman Rank Correlation Coefficient

The *Spearman rank correlation coefficient* is a popular nonparametric analog of the usual correlation coefficient. This statistic is calculated by replacing the data values with their ranks and calculating the correlation coefficient of the ranks. Tied values are replaced with the average rank of the ties. This coefficient is really a measure of association rather than correlation, since the ranks are unchanged by a monotonic transformation of the original data.

When *N* is greater than 10, the distribution of the Spearman rank correlation coefficient can be approximated by the distribution of the regular correlation coefficient.

Note that when weights are specified, the calculation of the Spearman rank correlation coefficient uses the weights.

Smoothing with Loess

The *loess* (locally weighted regression scatter plot smoothing) method is used to obtain a smooth curve representing the relationship between *X* and *Y*. Unlike linear regression, loess does not have a simple mathematical model. Rather, it is an algorithm that, given a value of *X*, computes an appropriate value of *Y*. The algorithm was designed so that the loess curve travels through the middle of the data, summarizing the relationship between *X* and *Y*.

The loess algorithm works as follows.

1. Select a value for *X*. Call it *X*0.
2. Select a neighborhood of points close to *X*0.
3. Fit a weighted regression of *Y* on *X* using only the points in this neighborhood. In the regression, the weights are inversely proportional to the distance between *X* and *X*0.
4. To make the procedure robust to outliers, a robust regression may be substituted for the weighted regression in step 3. This robust procedure modifies the weights so that observations with large residuals receive smaller weights.
5. Use the regression coefficients from the weighted regression in step 3 to obtained a predicted value for *Y* at *X*0.
6. Repeat steps 1 - 5 for a set of *X*’s between the minimum and maximum of *X*.

Mathematical Details of Loess

This section presents the mathematical details of the loess method of scatter plot smoothing. Note that implicit in the discussion below is the assumption that *Y* is the dependent variable and *X* is the independent variable.

Loess gives the value of *Y* for a given value of *X*, say *X*0. For each observation, define the distance between *X* and *X*0 as

$$d_j = \left| X_j - X0 \right|$$

Let *q* be the number of observations in the neighborhood of *X*0. Define *q* as [*fN*] where *f* is the user-supplied fraction of the sample. Here, [*Z*] is the largest integer in *Z*. Often *f* = 0.40 is a good choice. The neighborhood is defined as the observations with the *q* smallest values of d_j. Define d_q as the largest distance in the neighborhood of observations close to *X*0.

The tricube weight function is defined as

$$T(u) = \begin{cases} \left(1 - |u|^3\right)^3 & |u| < 1 \\ 0 & |u| \geq 1 \end{cases}$$

The weight for each observation is defined as

$$w_j = T\left(\frac{\left| X_j - X0 \right|}{d_q} \right)$$

The weighted regression for *X*0 is defined by the value of *b*0, *b*1, and *b*2 that minimize the sum of squares

$$\sum_{j=1}^{N} T\left(\frac{X_j - X0}{d_q}\right)\left(Y_j - b0 - b1\left(X_j\right) - b2\left(X_j\right)^2\right)^2$$

Note the if *b*2 is zero, a linear regression is fit. Otherwise, a quadratic regression is fit. The choice of linear or quadratic is an option in the procedure. The linear option is quicker, while the quadratic option fits peaks and valleys better. In most cases, there is little difference except at the extremes in the *X* space.

Once *b*0, *b*1, and *b*2 have be estimated using weighted least squares, the loess value is computed using

$$\hat{Y}_{loess}\left(X0\right) = b0 - b1\left(X0\right) - b2\left(X0\right)^2$$

Note that a separate weighted regression must be run for each value of *X*0.

Robust Loess

Outliers often have a large impact on least squares impact. A robust weighted regression procedure may be used to lessen the influence of outliers on the loess curve. This is done as follows.

The q loess residuals are computed using the loess regression coefficients using the formula

$$r_j = Y_j - \hat{Y}_{loess}\left(X_j\right)$$

New weights are defined as

$$w_j = w_{last,j} B\left(\frac{\left|r_j\right|}{6M}\right)$$

where $w_{last,j}$ is the previous weight for this observation, M is the median of the q absolute values of the residuals, and B(u) is the bisquare weight function defined as

$$B(u) = \begin{cases} \left(1 - u^2\right)^2 & |u| < 1 \\ 0 & |u| \geq 1 \end{cases}$$

This robust procedure may be iterated up to five items, but we have seen little difference in the appearance of the loess curve after two iterations.

Note that it is not always necessary to create the robust weights. If you are not going to remove the outliers from you final results, you probably should not remove them from the loess curve by setting the number of robust iterations to zero.

Testing Assumptions Using Residual Diagnostics

Evaluating the amount of departure in your data from each linear regression assumption is necessary to see if any remedial action is necessary before the fitted results can be used. First, the types of plots and statistical analyses the are used to evaluate each assumption will be given. Second, each of the diagnostic values will be defined.

Notation – Use of (j) and p

Several of these residual diagnostic statistics are based on the concept of studying what happens to various aspects of the regression analysis when each row is removed from the analysis. In what follows, we use the notation (j) to mean that observation j has been omitted from the analysis. Thus, $b(j)$ means the value of b calculated without using observation j.

Some of the formulas depend on whether the intercept is fitted or not. We use p to indicate the number of regression parameters. When the intercept is fit, p will be two. Otherwise, p will be one.

1 – No Outliers

Outliers are observations that are poorly fit by the regression model. If outliers are influential, they will cause serious distortions in the regression calculations. Once an observation has been determined to be an outlier, it must be checked to see if it resulted from a mistake. If so, it must be corrected or omitted. However, if no mistake can be found, the outlier should not be discarded just because it is an outlier. Many scientific discoveries have been made because outliers, data points that were different from the norm, were studied more closely. Besides being caused by simple data-entry mistakes, outliers often suggest the presence of an important independent variable that has been ignored.

Outliers are easy to spot on bar charts or box plots of the residuals and RStudent. RStudent is the preferred statistic for finding outliers because each observation is omitted from the calculation making it less likely that the outlier can mask its presence. Scatter plots of the residuals and RStudent against the X variable are also helpful because they may show other problems as well.

2 – Linear Regression Function - No Curvature

The relationship between Y and X is assumed to be linear (straight-line). No mechanism for curvature is included in the model. Although a scatter plot of Y versus X can show curvature in the relationship, the best diagnostic tool is the scatter plot of the residual versus X. If curvature is detected, the model must be modified to account for the curvature. This may mean adding a quadratic terms, taking logarithms of Y or X, or some other appropriate transformation.

Loess Curve

A loess curve should be plotted between X and Y to see if any curvature is present.

Lack of Fit Test

When the data include repeat observations at one or more *X* values (*replicates*), the adequacy of the linear model can evaluated numerically by performing a *lack of fit* test. This test procedure detects nonlinearities.

The lack of fit test is constructed as follows. First, the sum of squares for error is partitioned into two quantities: *lack of fit* and *pure error*. The pure error sum of squares is found by considering only those observations that are replicates. The *X* values are treated as the levels of the factor in a one-way analysis of variance. The sum of squares error from this analysis measures the underlying variation in *Y* that occurs when the value of *X* is held constant. Thus it is called *pure error*. When the pure error sum of squares is subtracted from the error sum of squares of the linear regression, the result is measure of the amount of nonlinearity in the data. An *F*-ratio can be constructed from these two values that will test the statistical significant of the lack of fit. The *F*-ratio is constructed using the following equation.

$$F_{DF1,DF2} = \frac{\frac{SS_{Lack\ of\ \ fit}}{DF1}}{\frac{SS_{Pure\ Error}}{DF2}}$$

where *DF*2 is the degrees of freedom for the error term in the one-way analysis of variance and *DF*1 is *N* - *DF2* - 2.

3 – Constant Variance

The errors are assumed to have constant variance across all values of *X*. If there are a lot of data (N > 100), nonconstant variance can be detected on a scatter plot of the residuals versus *X*. However, the most direct diagnostic tool to evaluate this assumption is a scatter plot of the absolute values of the residuals versus *X*. Often, the assumption is violated because the variance increases with *X*. This will show up as a ‘megaphone’ pattern to this plot.

When nonconstant variance is detected, a variance-stabilizing transformation such as the square-root or logarithm may be used. However, the best solution is probably to use weighted regression, with weights inversely proportional to the magnitude of the residuals.

Modified Levene Test

The *modified Levene test* can be used to evaluate the validity of the assumption of constant variance. It has been shown to be reliable even when the residuals do not follow a normal distribution.

The test is constructed by grouping the residuals according to the values of *X*. The number of groups is arbitrary, but usually, two groups are used. In this case, the absolute residuals of observations with low values of *X* are compared against those with high values of *X*. If the variability is constant, the variability in these two groups of residuals should be equal. The test is computed using the formula

$$L = \frac{\bar{d}_1 - \bar{d}_2}{s_L\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_L = \sqrt{\frac{\sum_{j=1}^{n_1}\left(d_{j1} - \bar{d}_1\right) + \sum_{j=1}^{n_2}\left(d_{j2} - \bar{d}_2\right)}{n_1 + n_2 - 2}}$$

$$d_{j1} = \left|e_{j1} - \tilde{e}_1\right|$$

$$d_{j2} = \left|e_{j2} - \tilde{e}_2\right|$$

and $\tilde{e}_1$ is the median of the group of residuals for low values of X and $\tilde{e}_2$ is the median of the group of residuals for high values of X. The test statistic L is approximately distributed as a t statistic with N - 2 degrees of freedom.

4 – Independent Errors

The Y's, and thus the errors, are assumed to be independent. This assumption is usually ignored unless there is a reason to think that it has been violated, such as when the observations were taken across time. An easy way to evaluate this assumption is a scatter plot of the residuals versus their sequence number (assuming that the data are arranged in time sequence order). This plot should show a relative random pattern.

The Durbin-Watson statistic is used as a formal test for the presence of first-order serial correlation. A more comprehensive method of evaluation is to look at the autocorrelations of the residuals at various lags. Large autocorrelations are found by testing each using Fisher's z transformation. Although Fisher's z transformation is only approximate in the case of autocorrelations, it does provide a reasonable measuring stick with which to judge the size of the autocorrelations.

If independence is violated, confidence intervals and hypothesis tests are erroneous. Some remedial method that accounts for the lack of independence must be adopted, such as using first differences or the Cochrane-Orcutt procedure.

Durbin-Watson Test

The Durbin-Watson test is often used to test for positive or negative, first-order, serial correlation. It is calculated as follows

$$DW = \frac{\sum_{j=2}^{N}\left(e_j - e_{j-1}\right)^2}{\sum_{j=1}^{N} e_j^2}$$

The distribution of this test is difficult because it involves the X values. Originally, Durbin-Watson (1950, 1951) gave a pair of bounds to be used. However, there is a large range of 'inclusion' found when using these bounds. Instead of using these bounds, we calculate the exact probability using the beta distribution approximation suggested by Durbin-Watson (1951). This approximation has been shown to be accurate to three decimal places in most cases which is all that are needed for practical work.

5 – Normality of Residuals

The residuals are assumed to follow the normal probability distribution with zero mean and constant variance. This can be evaluated using a normal probability plot of the residuals. Also, normality tests are used to evaluate this assumption. The most popular of the five normality tests provided is the Shapiro-Wilk test.

Unfortunately, a breakdown in any of the other assumptions results in a departure from this assumption as well. Hence, you should investigate the other assumptions first, leaving this assumption until last.

Influential Observations

Part of the evaluation of the assumptions includes an analysis to determine if any of the observations have an extra large influence on the estimated regression coefficients, on the fit of the model, or on the value of Cook's distance. By looking at how much removing an observation changes the results, an observation's influence can be determined.

Five statistics are used to investigate influence. These are Hat diagonal, DFFITS, DFBETAS, Cook's D, and COVARATIO.

Definitions Used in Residual Diagnostics

Residual

The residual is the difference between the actual *Y* value and the *Y* value predicted by the estimated regression model. It is also called the *error*, the *deviate*, or the *discrepancy*.

$$e_j = y_j - \hat{y}_j$$

Although the true errors, ε_j, are assumed to be independent, the computed residuals, e_j, are not. Although the lack of independence among the residuals is a concern in developing theoretical tests, it is not a concern on the plots and graphs.

The variance of the ε_j is σ^2. However, the variance of the e_j is not σ^2. In vector notation, the covariance matrix of **e** is given by

$$\begin{aligned} \mathrm{V}(\mathbf{e}) &= \sigma^2\left(\mathbf{I} - \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'W}^{\frac{1}{2}}\right) \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

The matrix **H** is called the *hat matrix* since it puts the 'hat' on *y* as is shown in the unweighted case.

$$\begin{aligned} \hat{Y} &= \mathbf{Xb} \\ &= \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y} \\ &= \mathbf{HY} \end{aligned}$$

Hence, the variance of e_j is given by

$$\mathrm{V}\left(e_j\right) = \sigma^2\left(1 - h_{jj}\right)$$

where h_{jj} is the jth diagonal element of **H**. This variance is estimated using

$$\hat{\mathrm{V}}\left(e_j\right) = s^2\left(1 - h_{jj}\right)$$

Hat Diagonal

The hat diagonal, h_{jj}, is the jth diagonal element of the hat matrix, H where

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$$

H captures an observation's remoteness in the *X*-space. Some authors refer to the hat diagonal as a measure of *leverage* in the *X*-space. As a rule of thumb, hat diagonals greater than 4/*N* are considered influential and are called high-leverage observations.

Note that a high-leverage observation is not a bad observation. Rather, high-leverage observations exert extra influence on the final results, so care should be taken to insure that they are correct. You should not delete an observation just because it has a high-influence. However, when you interpret the regression equation, you should bear in mind that the results may be due to a few, high-leverage observations.

Standardized Residual

As shown above, the variance of the observed residuals is not constant. This makes comparisons among the residuals difficult. One solution is to standardize the residuals by dividing by their standard deviations. This will give a set of residuals with constant variance.

The formula for this residual is

$$r_j = \frac{e_j}{s\sqrt{1 - h_{jj}}}$$

s(j) or MSEi

This is the value of the mean squared error calculated without observation *j*. The formula for *s*(*j*) is given by

$$\begin{aligned} s(j)^2 &= \frac{1}{N - p - 1}\sum_{i=1, i \neq j}^{N} w_i\left(y_i - \mathbf{x}_i\mathbf{b}(j)\right) \\ &= \frac{(N - p)s^2 - \dfrac{w_j e_j^2}{1 - h_{jj}}}{N - p - 1} \end{aligned}$$

RStudent

Rstudent is similar to the studentized residual. The difference is the *s*(*j*) is used rather than *s* in the denominator. The quantity *s*(*j*) is calculated using the same formula as *s*, except that observation *j* is omitted. The hope is that be excluding this observation, a better estimate of σ^2 will be obtained. Some statisticians refer to these as the *studentized deleted residuals*.

$$t_j = \frac{e_j}{s(j)\sqrt{1-h_{jj}}}$$

If the regression assumptions of normality are valid, a single value of the RStudent has a *t* distribution with *N* - 2 degrees of freedom. It is reasonable to consider |RStudent| > 2 as outliers.

DFFITS

DFFITS is the standardized difference between the predicted value with and without that observation. The formula for *DFFITS* is

$$DFFITS_j = \frac{\hat{y}_j - \hat{y}_j(j)}{s(j)\sqrt{h_{jj}}}$$

$$= t_j\sqrt{\frac{h_{jj}}{1-h_{jj}}}$$

The values of $\hat{y}_j(j)$ and $s^2(j)$ are found by removing observation *j* before the doing the calculations. It represents the number of estimated standard errors that the fitted value changes if the j^{th} observation is omitted from the data set. If |*DFFITS*| > 1, the observation should be considered to be influential with regards to prediction.

Cook's D

The DFFITS statistic attempts to measure the influence of a single observation on its fitted value. Cook's distance (Cook's *D*) attempts to measure the influence each observation on all *N* fitted values. The formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j\left[\hat{y}_j - \hat{y}_j(i)\right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. Rather than go to all the time of recalculating the regression coefficients *N* times, we use the following approximation

$$D_j = \frac{w_j e_j^2 h_{jj}}{ps^2\left(1-h_{jj}\right)^2}$$

This approximation is exact when no weight variable is used.

A Cook's *D* value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

CovRatio

This diagnostic flags observations that have a major impact on the generalized variance of the regression coefficients. A value exceeding 1.0 implies that the i^{th} observation provides an improvement, i.e., a reduction in the generalized variance of the coefficients. A value of CovRatio less than 1.0 flags an observation that increases the estimated generalized variance. This is not a favorable condition.

The general formula for the CovRatio is

$$\begin{aligned}\text{CovRatio}_j &= \frac{\det\left[s(j)^2\left(\mathbf{X}(j)'\mathbf{W}\mathbf{X}(j)\right)^{-1}\right]}{\det\left[s^2\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\right]} \\ &= \frac{1}{1-h_{jj}}\left[\frac{s(j)^2}{s^2}\right]^p\end{aligned}$$

where $p = 2$ if the intercept is fit or 1 if not.

Belsley, Kuh, and Welsch (1980) give the following guidelines for the CovRatio.

If CovRatio > $1 + 3p / N$ then omitting this observation significantly damages the precision of at least some of the regression estimates.

If CovRatio < $1 - 3p / N$ then omitting this observation significantly improves the precision of at least some of the regression estimates.

DFBETAS

The *DFBETAS* criterion measures the standardized change in a regression coefficient when an observation is omitted. The formula for this criterion is

$$DFBETAS_{kj} = \frac{b_k - b_k(j)}{s(j)\sqrt{c_{kk}}}$$

where c_{kk} is a diagonal element of the inverse matrix $\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$.

Belsley, Kuh, and Welsch (1980) recommend using a cutoff of $2/\sqrt{N}$ when N is greater than 100. When N is less than 100, others have suggested using a cutoff of 1.0 or 2.0 for the absolute value of *DFBETAS*.

Press Value

PRESS is an acronym for prediction sum of squares. It was developed for use in variable selection to validate a regression model. To calculate *PRESS*, each observation is individually omitted. The remaining N - 1 observations are used to calculate a regression and estimate the value of the omitted observation. This is done N times, once for each observation. The difference between the actual Y value and the predicted Y with the observation deleted is called the prediction error or *PRESS* residual. The sum of the squared prediction errors is the *PRESS* value. The smaller *PRESS* is, the better the predictability of the model.

The formula for PRESS is

$$PRESS = \sum_{j=1}^{N} w_j \left[y_j - \hat{y}_j(j) \right]^2$$

Press R-Squared

The PRESS value above can be used to compute an R^2-like statistic, called *R2Predict*, which reflects the prediction ability of the model. This is a good way to validate the prediction of a regression model without selecting another sample or splitting your data. It is very possible to have a high R^2 and a very low *R2Predict*. When this occurs, it implies that the fitted model is data dependent. This *R2Predict* ranges from below zero to above one. When outside the range of zero to one, it is truncated to stay within this range.

$$R^2_{Predict} = 1 - \frac{PRESS}{SS_{Total}}$$

Sum |Press residuals|

This is the sum of the absolute value of the *PRESS* residuals or prediction errors. If a large value for the *PRESS* is due to one or a few large *PRESS* residuals, this statistic may be a more accurate way to evaluate predictability. This quantity is computed as

$$\sum \left| PRESS \right| = \sum_{j=1}^{N} w_j \left| y_j - \hat{y}_j(j) \right|$$

Bootstrapping

Bootstrapping was developed to provide standard errors and confidence intervals for regression coefficients and predicted values in situations in which the standard assumptions are not valid. In these nonstandard situations, bootstrapping is a viable alternative to the corrective action suggested earlier. The method is simple in concept, but it requires extensive computation time.

The bootstrap is simple to describe. You assume that your sample is actually the population and you draw *B* samples (*B* is over 1000) of size *N* from your original sample with replacement. With replacement means that each observation may be selected more than once. For each bootstrap sample, the regression results are computed and stored.

Suppose that you want the standard error and a confidence interval of the slope. The bootstrap sampling process has provided *B* estimates of the slope. The standard deviation of these *B* estimates of the slope is the bootstrap estimate of the standard error of the slope. The bootstrap confidence interval is found by arranging the *B* values in sorted order and selecting the appropriate percentiles from the list. For example, a 90% bootstrap confidence interval for the slope is given by fifth and ninety-fifth percentiles of the bootstrap slope values. The bootstrap method can be applied to many of the statistics that are computed in regression analysis.

The main assumption made when using the bootstrap method is that your sample approximates the population fairly well. Because of this assumption, bootstrapping does not work well for small samples in which there is little likelihood that the sample is representative of the population. Bootstrapping should only be used in medium to large samples.

When applied to linear regression, there are two types of bootstrapping that can be used.

Modified Residuals

Davison and Hinkley (1999) page 279 recommend the use of a special rescaling of the residuals when bootstrapping to keep results unbiased. These modified residuals are calculated using

$$e_j^* = \frac{e_j}{\sqrt{\frac{1-h_{jj}}{w_j}}} - \overline{e}^*$$

where

$$\overline{e}^* = \frac{\sum_{j=1}^{N} w_j e_j^*}{\sum_{j=1}^{N} w_j}$$

Bootstrap the Observations

The bootstrap samples are selected from the original sample of *X* and *Y* pairs. This method is appropriate for data in which both *X* and *Y* have been selected at random. That is, the *X* values were not predetermined, but came in as measurements just as the *Y* values.

An example of this situation would be if a population of individuals is sampled and both *Y* and *X* are measured on those individuals only after the sample is selected. That is, the value of *X* was not used in the selection of the sample.

Bootstrap the Residuals

The bootstrap samples are constructed using the modified residuals. In each bootstrap sample, the randomly sampled modified residuals are added to the original fitted values forming new values of *Y*. This method forces the original structure of the *X* values to be retained in every bootstrap sample.

This method is appropriate for data obtained from a designed experiment in which the values of *X* are preset by the experimental design.

Because the residuals are sampled and added back at random, the method must assume that the variance of the residuals is constant. **If the sizes of the residuals are proportional to *X*, this method should not be used.**

Bootstrap Prediction Intervals

Bootstrap confidence intervals for the mean of *Y* given *X* are generated from the bootstrap sample in the usual way. To calculate prediction intervals for the predicted value (not the mean) of *Y* given *X* requires a modification to the predicted value of *Y* to be made to account for the variation of *Y* about its mean. This modification of the predicted *Y* values in the bootstrap sample, suggested by Davison and Hinkley, is as follows.

$$\hat{y}_+ = \hat{y} - x\left(b_1^* - b_1\right) + e_+^*$$

where e_{+}^{*} is a randomly selected modified residual. By adding the randomly sample residual we have added an appropriate amount of variation to represent the variance of individual *Y*'s about their mean value.

Randomization Test

Because of the strict assumptions that must be made when using this procedure to test hypotheses about the slope, ***NCSS*** also includes a randomization test as outlined by Edgington (1987). Randomization tests are becoming more and more popular as the speed of computers allows them to be computed in seconds rather than hours.

A randomization test is conducted by enumerating all possible permutations of the dependent variable while leaving the independent variable in the original order. The slope is calculated for each permutation and the number of permutations that result in a slope with a magnitude greater than or equal to the actual slope is counted. Dividing this count by the number of permutations tried gives the significance level of the test.

For even moderate sample sizes, the total number of permutations is in the trillions, so a Monte Carlo approach is used in which the permutations are found by random selection rather than complete enumeration. Edgington suggests that at least 1,000 permutations by selected. We suggest that this be increased to 10,000.

Data Structure

The data are entered as two variables. If weights or frequencies are available, they are entered separately in other variables. An example of data appropriate for this procedure is shown below. These data are the heights and weights of twenty individuals. The data are contained in the LINREG1 database. We suggest that you open this database now so that you can follow along with the examples.

LINREG1 dataset (subset)

Height	Weight
64	159
63	155
67	157
60	125
52	103
58	122
56	101
52	82
79	228
76	199
73	195

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present on a row for all but the dependent variable, a predicted value and confidence limits are generated for that row.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable(s)

Specifies a dependent (*Y*) variable. This variable should contain only numeric values. If more than one variable is specified, a separate analysis is run for each.

Independent Variable

X: Independent Variable

Specifies the variable to be used as independent (*X*) variable. This variable should contain only numeric values.

Frequency Variable

Frequency Variable

Specify an optional frequency (count) variable. This variable contains integers that represent the number of observations (frequency) associated with each observation. If left blank, each observation has a frequency of one. This variable lets you modify that frequency. This is especially useful when your data are already tabulated and you want to enter the counts.

Weight Variable

Weight Variable

A weight variable may be specified to set the (non-negative) weight given to each observation in a weighted regression. By default, each observation receives an equal weight of 1 / *N* (where *N* is the sample size). This variable allows you to specify different weights for different observations.

NCSS automatically scales the weights so that they sum to one. Hence, you can enter integer numbers and ***NCSS*** will scale them to appropriate fractions.

The weight variable is commonly created in the Robust Regression procedure.

Model Specification

Remove Intercept

Specifies whether to remove the *Y*-intercept term from the regression model. In most cases, you will want to keep the intercept term by leaving this option unchecked.

Note that removing the *Y*-intercept from the regression equation distorts many of the common regression measures such as *R*-Squared, mean square error, and *t*-tests. You should not use these measures when the intercept has been omitted.

Resampling

Calculate Bootstrap C.I.'s

This option causes bootstrapping to be done and all associated bootstrap reports and plots to be generated. Bootstrapping may be very time consuming when the sample size is large (say > 1000).

Run randomization tests

Check this option to run the randomization test. Note that this test is computer-intensive and may require a great deal of time to run.

Alpha Levels

Alpha for C.I.'s and Tests

Alpha is the significance level used in the hypothesis tests. One minus alpha is the confidence level (confidence coefficient) of the confidence intervals.

A value of 0.05 is commonly used. This corresponds to a chance of 1 out of 20. You should not be afraid to use other values since 0.05 became popular in pre-computer days when it was the only value available. Typical values range from 0.001 to 0.20.

Alpha for Assumptions

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests use a larger alpha such as 0.15 or 0.20.

We recommend 0.20.

Reports Tab

The following options control which reports and plots are displayed. Since over 25 reports are available, you may want to spend some time deciding which reports you want to display on a routine basis and create a template that saves your favorite choices.

Select Report / Plot Group

Select a Group of Reports and Plots

This option allows you to specify a group of reports and plots without checking them individually. The checking of individual reports and plots is only useful when this option is set to *Display only those items that are CHECKED BELOW*. Otherwise, the checking of individual reports and plots is ignored.

Report Options

Show Notes

This option controls whether the available notes and comments that are displayed at the bottom of each report. This option lets you omit these notes to reduce the length of the output.

Show All Rows

This option makes it possible to display predicted values for only a few designated rows.

When checked predicted values, residuals, and other row-by-row statistics, will be displayed for all rows used in the analysis.

When not checked, predicted values and other row-by-row statistics will be displayed for only those rows in which the dependent variable's value is missing.

Select Reports – Summaries

Run Summary ... Summary Matrices

Each of these options specifies whether the indicated report is calculated and displayed. Note that since some of these reports provide results for each row, they may be too long for normal use when requested on large databases.

Select Reports – Estimation

Regression Estimation

Indicate whether to display this report.

Select Reports – ANOVA

ANOVA

Indicate whether to display this report.

Select Reports – Assumptions

Assumptions

Indicate whether to display this report.

Levene Groups

This option sets the number of groups used in Levene's constant-variance of residuals test. In most cases, a '2' should be used. In all cases, the number of groups should be small enough so that you have at least 25 observations in each group.

Durbin-Watson

Indicate whether to display this report.

PRESS

Indicate whether to display this report.

Select Reports – Prediction

Predict Y at these X Values

Enter an optional list of *X* values at which to report predicted values of *Y* and confidence intervals. Note that these values are also reported on in the bootstrap reports.

You can enter a single number or a list of numbers. The list can be separated with commas or spaces. The list can also be of the form *XX:YY(ZZ)* which means *XX* to *YY* by *ZZ*.

Examples:

10

10 20 30 40 50

0:100(10)

0:90(10) 100:900(100) 1000 5000

Predicted Y – C.L.

Indicate whether to display the confidence limits for the mean of Y at a specific X.

Predicted Y – P.L.

Indicate whether to display the prediction limits for Y at a specific X.

Select Reports – Row-by-Row Lists

Original Data ... Predicted X Individuals

Indicate whether to display these reports. Note that since these reports provide results for each row, they may be too long for normal use when requested on large databases.

Select Reports – Regression Diagnostics

Residuals ... Outlier-Influence Chart

Indicate whether to display these reports.

Select Plots

Y vs X Plot ... Probability Plot

Indicate whether to display these plots.

Format Tab

These options specify the number of decimal places shown when the indicated value is displayed in a report. The number of decimal places shown in plots is controlled by the Tick Labels buttons on the Axis Setup window.

Report Options

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimal Places

Probability ... Matrix Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'All' to display all digits available. The number of digits displayed by this option is controlled by whether the PRECISION option is SINGLE or DOUBLE.

Plot Options

Y vs X Plot Size and All Other Plot Sizes

These options control the size of the plots. Possible choices are shown below.

- **Small**

  Each plot is about 2.5 inches wide. Two plots are shown per line. Six plots fit on a page.

- **Medium**

  Each plot is about 4.5 inches wide. One plot is shown per line. Two plots fit on a page.

- **Large**

  Each plot is about 5.5 inches wide. One plot is shown per line. One plot fits on a page.]

Resampling Tab

This panel controls the bootstrapping and randomization test. Note that bootstrapping and the randomization test are only used when Calculate Bootstrap C.I.'s and Run Randomization Tests are checked, respectively.

Bootstrap Calculation Options

The following options control the calculation of bootstrap confidence intervals.

Bootstrap Calculation Options – Sampling

Samples (N)

This is the number of bootstrap samples used. A general rule of thumb is that you use at least 100 when standard errors are your focus or at least 1000 when confidence intervals are your focus. If computing time is available, it does not hurt to do 4000 or 5000.

We recommend setting this value to at least 3000.

Sampling Method

Specify which of the two sampling methods are to be used in forming the bootstrap sample.

- **Observations**

  Each bootstrap sample is obtained as a random sample with replacement from the original *X*-*Y* pairs. This method is appropriate when the *X* values were not set before the original sample was taken.

- **Residuals**

  Each bootstrap sample is obtained as a random sample with replacement from the original set of residuals. These residuals are added to the predicted values to form the bootstrap sample. The original *X* structure is maintained by each bootstrap sample. This method is appropriate when a limited number of X values were selected by the experimental design.

  We recommend setting this value to at least 3000.

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. We recommend setting the parameter to at least 50.

Bootstrap Calculation Options – Estimation

Percentile Type

The method used to create the percentiles when forming bootstrap confidence limits. You can read more about the various types of percentiles in the Descriptive Statistics chapter. We suggest you use the Ave *X*(p[n+1]) option.

C.I. Method

This option specifies the method used to calculate the bootstrap confidence intervals. The reflection method is recommended.

- **Percentile**

  The confidence limits are the corresponding percentiles of the bootstrap values.

- **Reflection**

  The confidence limits are formed by reflecting the percentile limits. If *X*0 is the original value of the parameter estimate and *XL* and *XU* are the percentile confidence limits, the Reflection interval is (2 *X*0 - *XU*, 2 *X*0 - *XL*).

Bootstrap Confidence Coefficients

These are the confidence coefficients of the bootstrap confidence intervals. Since bootstrapping calculations may take several minutes, it may be useful to obtain confidence intervals using several different confidence coefficients.

All values must be between 0.50 and 1.00. You may enter several values, separated by blanks or commas. A separate confidence interval is given for each value entered.

Examples:

0.90 0.95 0.99

0.90:.99(0.01)

0.90.

Bootstrap Histogram Options

Vertical Axis Label

This is the label of the vertical axis of a bootstrap histogram.

Horizontal Axis Label

This is the label of the horizontal axis of a bootstrap histogram.

Plot Style File

This is the histogram style file. We have provided several different style files to choose from, or you can create your own in the Histogram procedure.

Histogram Title

This is the title used on the bootstrap histograms.

Number of Bars

The number of bars shown in a bootstrap histogram. We recommend setting this value to at least 25 when the number of bootstrap samples is over 1000.

Randomization Test Options

Monte Carlo Samples

Specify the number of Monte Carlo samples used when conducting randomization tests. You also need to check the ‘Run Randomization Tests’ box to run this test.

Somewhere between 1,000 and 100,000 Monte Carlo samples are usually necessary. Although the default is 1,000, we suggest the use of 10,000 when using this test.

Axis Setup Tab

The options on this panel control the appearance of the *X* variable, the *Y* variable, and the residuals whenever they are included on a plot. This makes it easy to give a consistent look to all of your plots without modifying them individually.

Y-Variable, X-Variable, and Residuals Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by the names of the corresponding variables. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the axis associated with this variable. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the tick labels along each axis.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Y vs X Plot Tab

This panel controls which objects are shown on the scatterplot of *Y* and *X*. The labels, minimums, maximums, and number of tickmarks can be modified on the Axis Setup window.

Y vs X Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. We have provided several style files to choose from, or you can create your own. Style files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Plot Contents

Y on X Line

Check this box to cause the regression line of *Y* on *X* to be displayed. The color and thickness of the line can be changed using the button just below this check box.

X on Y Line

Check this box to cause the regression line *X* on *Y* to be displayed. The color and thickness of the line can be changed using the button just below this check box.

Y and X Line

Check this box to cause the orthogonal regression line to be displayed. The color and thickness of the line can be changed using the button just below this check box.

Prediction Limits

Check this box to cause the prediction limits to be displayed. The confidence level used to create these limits is set on the Format window. The color and thickness of the line can be changed using the button just below this check box.

Confidence Limits

Check this box to cause the confidence limits about the mean to be displayed. The confidence level used to create these limits is set on the Format window. The color and thickness of the line can be changed using the button just below this check box.

Confidence Band

Check this box to cause a confidence band to be displayed. The confidence level used to create the band is set on the Format window. The color and thickness of the line can be changed using the button just below this check box.

Probability Ellipse

Check this box to cause a probability ellipse to be displayed. The confidence level used to create the ellipse is set on the Format window. The color and thickness of the line can be changed using the button just below this check box.

LOESS Curve

Check this box to cause a LOESS curve to be displayed. The color and thickness of the line can be changed using the button just below this check box.

Plot Contents – Number of Evaluation Points

Number of Points

Specify the number of points at which the regression lines, confidence limits, prediction limits, confidence bands, and probability ellipse are evaluated. This effects the granularity of these objects. Although this value can range from 20 to over 2000, we have found that 200 works well.

Plot Contents – Loess Options

LOESS Order

The order of the polynomial fit in the LOESS procedure. Select ‘1” for a linear fit or ‘2’ for a quadratic fit. The quadratic fit tends to pick up sudden changes better than the linear fit.

LOESS %N

This specifies the percent of the dataset to be used at each LOESS calculation. Although the allowable range is 1 to 99, a value from 25 to 40 is usually optimal. The large this value is, the smoother the LOESS curve will be.

LOESS Robust

This specifies the number of robust iterations used in the LOESS algorithm to downplay the influence of outliers. Select ‘0’ if you do not want robust iterations. This will greatly reduce the execution time in large datasets. Select ‘1’ for one iteration. Select ‘2’ for two iterations.

We recommend selecting ‘2’ for datasets with $N < 100$, ‘1’ for datasets with $N < 500$, and ‘0’ otherwise.

Note that using ‘1’ or ‘2’ will cause the algorithm to try find and ignore outliers. If you want the effect of outliers to be shown in the LOESS curve, set this value to ‘0’.

Resid vs X Plot ... Serial Corr Plot Tabs

Various residual plots may be displayed to help you validate the assumptions of your regression analysis as well as investigate the fit of your estimated equation. The options on these panels control the appearance of the corresponding residual scatter plot. The appearance of the residual axis and the *Y*-axis (when used) is controlled on the Axis Setup tab (described above).

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Histogram Tab

The options on this panel control the appearance of the histogram of the residuals. Note that the residual axis is controlled on the Axis Setup window. The Vertical (Count) Axis options are described under the Axis Setup Tab heading above.

Plot Settings

Plot Style File

Designate a histogram style file. This file sets all histogram options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Histogram procedure.

Number of Bars

Specify the number of intervals, bins, or bars used in the histogram.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot. Note that the residual axis is controlled on the Axis Setup tab. The Horizontal (Expected) Axis options are described under the Axis Setup Tab heading above.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful.

Data Storage Options – Select Items to Store

Predicted Y ... LOESS Values

Indicate whether to store these row-by-row values, beginning at the variable indicated by the *Store First Variable In* option.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Running a Linear Regression Analysis

This section presents an example of how to run a linear regression analysis of the data in the LINREG1 database. In this example, we will run a regression of *Height* on *Weight*. Predicted values of Height are wanted at Weight values equal to 90, 100, 150, 200, and 250.

This regression program outputs over thirty different reports and plots, many of which contain duplicate information. For the purposes of annotating the output, we will output all of the reports. Normally, you would only select a few these reports.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Linear Regression and Correlation window.

1 Open the LINREG1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LinReg1.s0**.
- Click **Open**.

2 Open the Linear Regression and Correlation window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Linear Regression and Correlation**. The Linear Regression and Correlation procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Linear Regression and Correlation window, select the **Variables tab**.
- Set the ***Y*: Dependent Variable** box to **Height**.
- Set the ***X*: Independent Variable** box to **Weight**.
- Check the **Calculate Bootstrap C.I.'s** and **Run Randomization Tests** boxes.

4 Specify the randomization test options.

- Select the **Resampling tab**.
- Set the **Monte Carlo Samples** to **1000**.

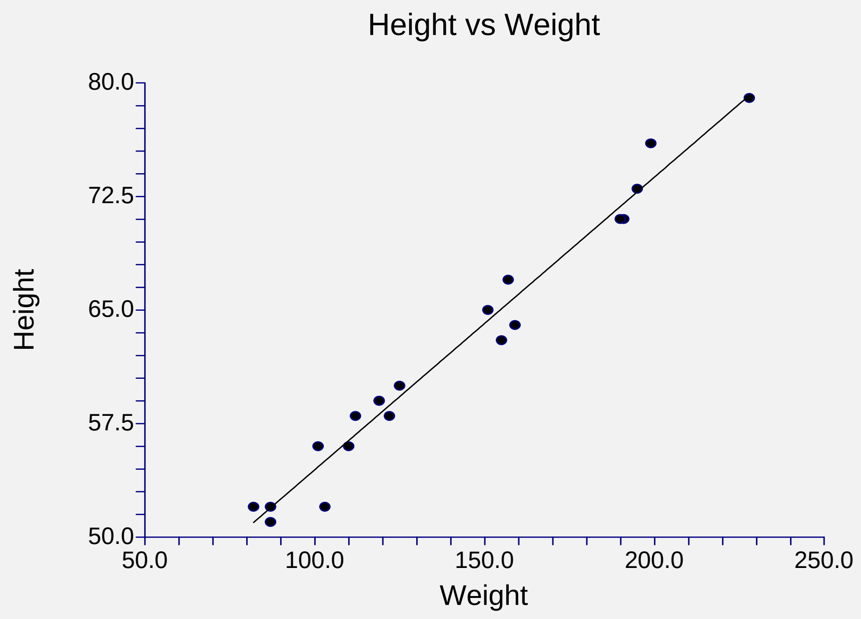
5 Specify the reports.

- Select the **Reports tab**.
- Set the Predict Y at these X Values box to **90 100 150 200 250**.
- Under **Select a Group of Reports and Plots**, select **Display ALL reports & plots**. As we mentioned above, normally you would only view a few of these reports, but we are selecting them all so that we can document them.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Linear Regression Plot Section



The plot shows the data and the linear regression line. This plot is very useful for finding outliers and nonlinearities. It gives you a good feel for how well the linear regression model fits the data.

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	Height	Rows Processed	20
Independent Variable	Weight	Rows Used in Estimation	20
Frequency Variable	None	Rows with X Missing	0
Weight Variable	None	Rows with Freq Missing	0
Intercept	35.1337	Rows Prediction Only	0
Slope	0.1932	Sum of Frequencies	20
R-Squared	0.9738	Sum of Weights	20.0000
Correlation	0.9868	Coefficient of Variation	0.0226
Mean Square Error	1.970176	Square Root of MSE	1.40363

This report summarizes the linear regression results. It presents the variables used, the number of rows used, and the basic least squares results. These values are repeated later in specific reports, so they will not be discussed further here.

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, computed by dividing the square root of the mean square error by the mean of *Y*. By itself, it has little value, but it can be useful in comparative studies.

$$CV = \frac{\sqrt{MSE}}{\overline{Y}}$$

Summary Statement

```
The equation of the straight line relating Height and Weight is estimated as: Height =
(35.1337) + (0.1932) Weight using the 20 observations in this dataset. The y-intercept, the
estimated value of Height when Weight is zero, is 35.1337 with a standard error of 1.0887. The
slope, the estimated change in Height per unit change in Weight, is 0.1932 with a standard
error of 0.0075. The value of R-Squared, the proportion of the variation in Height that can be
accounted for by variation in Weight, is 0.9738. The correlation between Height and Weight is
0.9868.
A significance test that the slope is zero resulted in a t-value of 25.8679. The significance
level of this t-test is 0.0000. Since 0.0000 < 0.0500, the hypothesis that the slope is zero is
rejected.
The estimated slope is 0.1932. The lower limit of the 95% confidence interval for the slope is
0.1775 and the upper limit is 0.2089. The estimated intercept is 35.1337. The lower limit of
the 95% confidence interval for the intercept is 32.8464 and the upper limit is 37.4209.
```

This report gives an explanation of the results in text format.

Descriptive Statistics Section

```
Parameter                 Dependent     Independent
Variable                     Height          Weight
Count                            20              20
Mean                        62.1000        139.6000
Standard Deviation           8.4411         43.1221
Minimum                     51.0000         82.0000
Maximum                     79.0000        228.0000
```

This report presents the mean, standard deviation, minimum, and maximum of the two variables. It is particularly useful for checking that the correct variables were selected.

Regression Estimation Section

```
                                          Intercept          Slope
Parameter                                      B(0)           B(1)
Regression Coefficients                     35.1337         0.1932
Lower 95% Confidence Limit                  32.8464         0.1775
Upper 95% Confidence Limit                  37.4209         0.2089
Standard Error                               1.0887         0.0075
Standardized Coefficient                     0.0000         0.9868

T Value                                     32.2716        25.8679
Prob Level                                   0.0000         0.0000
Prob Level (Randomization Test N =1000)                     0.0010
Reject H0 (Alpha = 0.0500)                      Yes            Yes
Power (Alpha = 0.0500)                       1.0000         1.0000

Regression of Y on X                        35.1337         0.1932
Inverse Regression from X on Y              34.4083         0.1984
Orthogonal Regression of Y and X            35.1076         0.1934

Estimated Model
( 35.1336680743148) + ( .193168566802902) * (Weight)
```

This section reports the values and significance tests of the regression coefficients. Before using this report, check that the assumptions are reasonable by looking at the tests of assumptions report.

Regression Coefficients

The regression coefficients are the least-squares estimates of the *Y*-intercept and the slope. The slope indicates how much of a change in *Y* occurs for a one-unit change in *X*.

Lower - Upper 95% Confidence Limits

These are the lower and upper values of a 100(1-α)% interval estimate for β_j based on a *t*-distribution with *N* - 2 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed.

The formulas for the lower and upper confidence limits are

$$b_j \pm t_{1-\alpha/2,n-2} s_{b_j}$$

Standard Error

The standard error of the regression coefficient, s_{b_j} , is the standard deviation of the estimate. It provides a measure of the precision of the estimated regression coefficient. It is used in hypothesis tests or confidence limits.

Standardized Coefficient

Standardized regression coefficients are the coefficients that would be obtained if you standardized both variables. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

The formula for the standardized regression coefficient is:

$$b_{1,\, std} \;=\; b_1\left(\frac{s_X}{s_Y}\right)$$

where s_Y and s_X are the standard deviations for the dependent and independent variables, respectively.

Note that in the case of linear regression, the standardized coefficient is equal to the correlation between the two variables.

T-Value

These are the *t*-test values for testing the hypotheses that the intercept and the slope are zero versus the alternative that they are nonzero. These *t*-values have *N - 2* degrees of freedom.

To test that the slope is equal to a hypothesized value other than zero, inspect the confidence limits. If the hypothesized value is outside the confidence limits, the hypothesis is rejected. Otherwise, it is not rejected.

Prob Level

This is the two-sided *p*-value for the significance test of the regression coefficient. The *p*-value is the probability that this *t*-statistic will take on a value at least as extreme as the actually observed value, assuming that the null hypothesis is true (i.e., the regression estimate is equal to zero). If the *p*-value is less than alpha, say 0.05, the null hypothesis is rejected.

Prob Level (Randomization Test)

This is the two-sided *p*-value for the randomization test of whether the slope is zero. Since this value is based on a randomization test, it does not require all of the assumptions that the *t*-test does. The number of Monte Carlo samples of the permutation distribution of the slope is shown in parentheses.

Reject H0 (Alpha = 0.05)

This value indicates whether the null hypothesis was reject. Note that the level of significance was specified as the value of *Alpha*.

Power (Alpha = 0.05)

Power is the probability of rejecting the null hypothesis that the regression coefficient is zero when in truth, the regression coefficient is some value other than zero. The power is calculated for the case when the estimate coefficient is the actual coefficient, the estimate variance is the true variance, and Alpha is the given value.

High power is desirable. High power means that there is a high probability of rejecting the null hypothesis when the null hypothesis is false. This is a critical measure of sensitivity in hypothesis testing. This estimate of power is based upon the assumption that the residuals are normally distributed.

Regression of Y on X

These are the usual least squares estimates of the intercept and slope from a linear regression of *Y* on *X*. These quantities were given earlier and are reproduced here to allow easy comparisons.

Regression of X on Y

These are the estimated intercept and slope derived from the coefficients of linear regression of *X* on *Y*. These quantities may be useful in calibration and inverse prediction.

Orthogonal Regression of Y and X

The are the estimates of the intercept and slope from an orthogonal regression of *Y* on *X*. This equation minimizes the sum of the squared perpendicular distances between the points and the regression line.

Estimated Model

This is the least squares regression line presented in double precision. Besides showing the regression model in long form, it may be used as a transformation by copying and pasting it into the Transformation portion of the spreadsheet.

Bootstrap Section

```
---  Estimation Results    ------  | --- Bootstrap Confidence Limits----
Parameter              Estimate  | Conf. Level Lower          Upper
Intercept
Original Value          35.1337  | 0.9000     33.5138       36.8691
Bootstrap Mean          35.1391  | 0.9500     33.1520       37.2812
Bias (BM - OV)           0.0055  | 0.9900     32.6492       38.1285
Bias Corrected          35.1282
Standard Error           1.0178
Slope
Original Value           0.1932  | 0.9000      0.1815        0.2047
Bootstrap Mean           0.1931  | 0.9500      0.1785        0.2069
Bias (BM - OV)           0.0000  | 0.9900      0.1729        0.2118
Bias Corrected           0.1932
Standard Error           0.0071
Correlation
Original Value           0.9868  | 0.9000      0.9799        0.9973
Bootstrap Mean           0.9865  | 0.9500      0.9789        1.0000
Bias (BM - OV)          -0.0003  | 0.9900      0.9772        1.0000
Bias Corrected           0.9871
Standard Error           0.0056
R-Squared
Original Value           0.9738  | 0.9000      0.9601        0.9943
Bootstrap Mean           0.9733  | 0.9500      0.9582        0.9996
Bias (BM - OV)          -0.0005  | 0.9900      0.9548        1.0000
Bias Corrected           0.9743
Standard Error           0.0109
Standard Error of Estimate
Original Value           1.4036  | 0.9000      1.1710        1.8446
Bootstrap Mean           1.3241  | 0.9500      1.1225        1.9071
Bias (BM - OV)          -0.0795  | 0.9900      1.0355        2.0552
Bias Corrected           1.4832
Standard Error           0.2046
Orthogonal Intercept
Original Value          35.1076  | 0.9000     33.4855       36.8576
Bootstrap Mean          35.1123  | 0.9500     33.1251       37.2581
Bias (BM - OV)           0.0047  | 0.9900     32.6179       38.1223
Bias Corrected          35.1028
Standard Error           1.0231
Orthogonal Slope
Original Value           0.1934  | 0.9000      0.1816        0.2048
Bootstrap Mean           0.1933  | 0.9500      0.1786        0.2071
Bias (BM - OV)           0.0000  | 0.9900      0.1731        0.2120
Bias Corrected           0.1934
Standard Error           0.0071
Predicted Mean and Confidence Limits of Height when Weight = 90.0000
Original Value          52.5188  | 0.9000     51.8172       53.2993
Bootstrap Mean          52.5220  | 0.9500     51.6895       53.4913
Bias (BM - OV)           0.0032  | 0.9900     51.4648       53.8741
Bias Corrected          52.5157
Standard Error           0.4549
(Report continues for the other values of Weight)
Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.
```

This report provides bootstrap estimates of the slope and intercept of the least squares regression line and the orthogonal regression line, the correlation coefficient, and other linear regression quantities. Note that bootstrap confidence intervals and prediction intervals are provided for each of the *X* (Weight) values. Details of the bootstrap method were presented earlier in this chapter.

Note that since these results are based on 3000 random bootstrap samples, they will differ slightly from the results you obtain when you run this report.

Original Value

This is the parameter estimate obtained from the complete sample without bootstrapping.

Bootstrap Mean

This is the average of the parameter estimates of the bootstrap samples.

Bias (BM - OV)

This is an estimate of the bias in the original estimate. It is computed by subtracting the original value from the bootstrap mean.

Bias Corrected

This is an estimated of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate.

Standard Error

This is the bootstrap method’s estimate of the standard error of the parameter estimate. It is simply the standard deviation of the parameter estimate computed from the bootstrap estimates.

Conf. Level

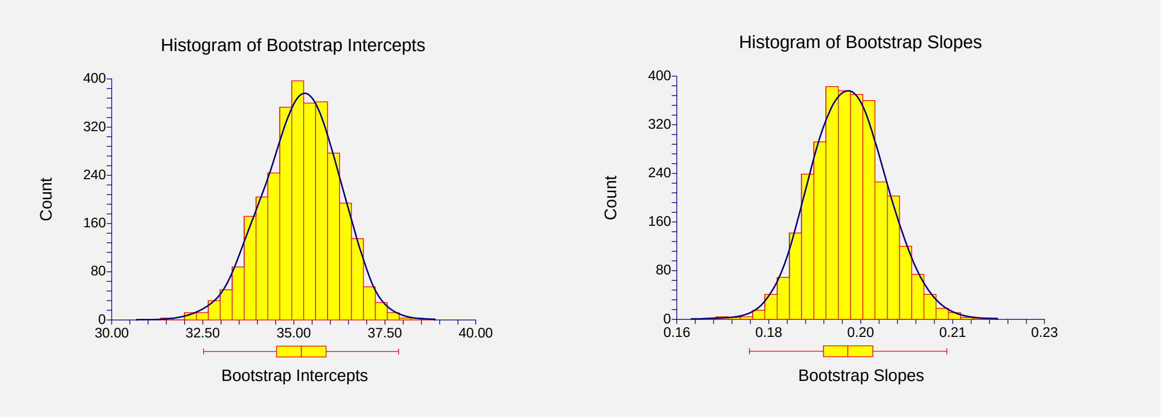
This is the confidence coefficient of the bootstrap confidence interval given to the right.

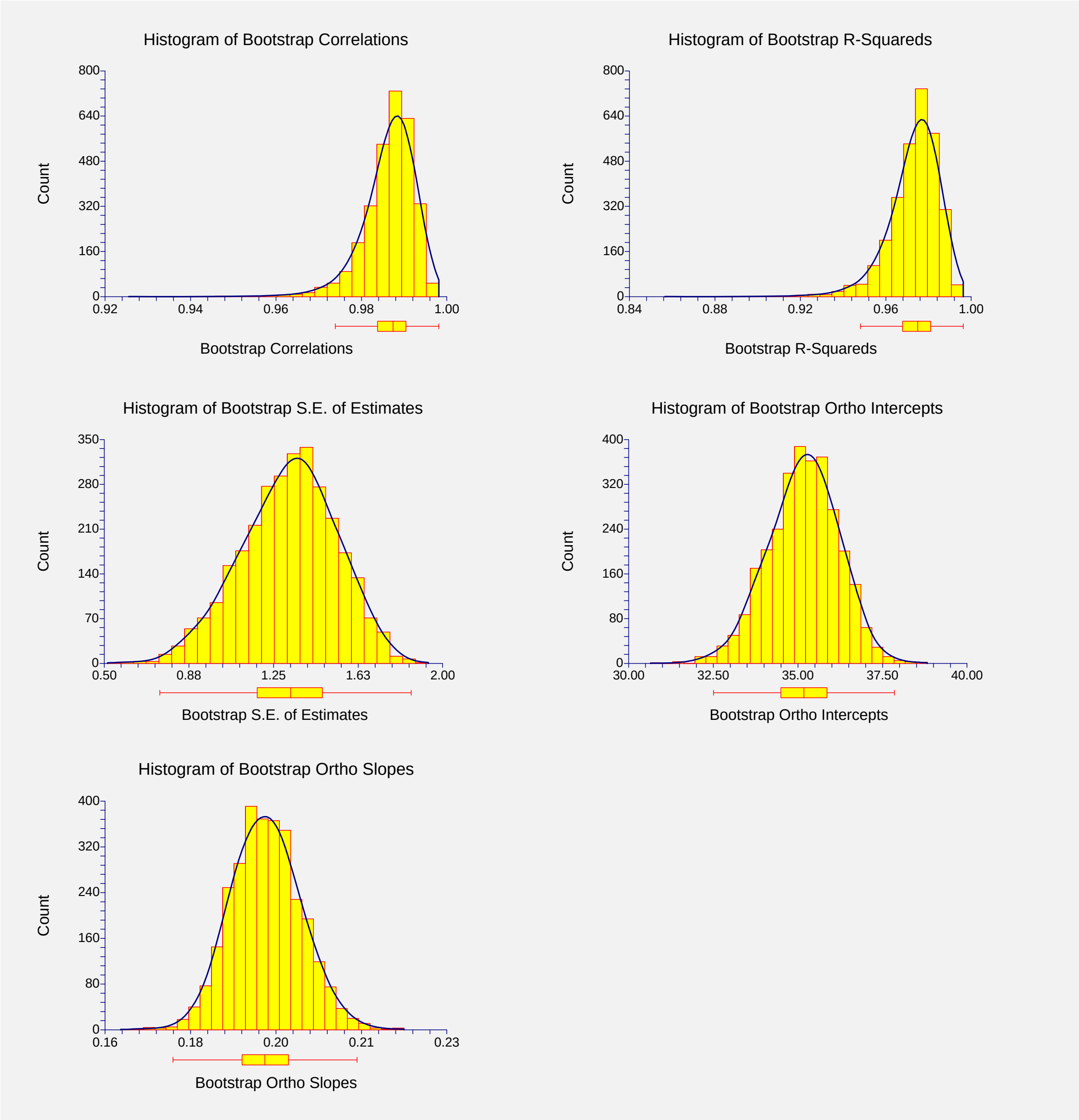
Bootstrap Confidence Limits - Lower and Upper

These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

Bootstrap Histograms Section





Each histogram shows the distribution of the corresponding parameter estimate.

Note that the number of decimal places shown in the horizontal axis is controlled by which histogram style file is selected. In this example, we selected Bootstrap2, which was created to provide two decimal places.

Correlation and R-Squared Section

```
                                                                 Spearman
                                         Pearson                     Rank
                                     Correlation              Correlation
Parameter                            Coefficient    R-Squared  Coefficient
Estimated Value                           0.9868       0.9738       0.9759
Lower 95% Conf. Limit (r dist'n)          0.9646
Upper 95% Conf. Limit (r dist'n)          0.9945
Lower 95% Conf. Limit (Fisher's z)        0.9662                    0.9387
Upper 95% Conf. Limit (Fisher's z)        0.9949                    0.9906
Adjusted (Rbar)                                        0.9723
T-Value for H0: Rho = 0                  25.8679      25.8679      18.9539
Prob Level for H0: Rho = 0                0.0000       0.0000       0.0000
```

This report provides results about Pearson's correlation, R-squared, and Spearman's rank correlation.

Pearson Correlation Coefficient

Details of the calculation of this value were given earlier in the chapter. Remember that this value is an index of the strength of the linear association between *X* and *Y*. The range of values is from -1 to 1. Strong association occurs when the magnitude of the correlation is close to one. Low correlations are those near zero.

Two sets of confidence limits are given. The first is a set of exact limits computed from the distribution of the correlation coefficient. These limits assume that *X* and *Y* follow the bivariate normal distribution. The second set of limits are limits developed by R. A. Fisher as an approximation to the exact limits. The approximation is quite good as you can see by comparing the two sets of limits. The second set is provided because they are often found in statistics books. In most cases, you should use the first set based on the *r* distribution because they are exact. You may want to compare these limits with those found for the correlation in the Bootstrap report.

The two-sided hypothesis test and probability level are for testing whether the correlation is zero.

Prob Level (Randomization Test)

This is the two-sided *p*-value for the randomization test of whether the slope is zero. This probability value may also be used to test whether the Pearson correlation is zero. Since this value is based on a randomization test, it does not require all of the assumptions that the parametric test does. The number of Monte Carlo samples of the permutation distribution of the slope is shown in parentheses.

Spearman Rank Correlation Coefficient

The Spearman's rank correlation is simply the Pearson correlation computed on the ranks of *X* and *Y* rather than on the actual data. By using the ranks, some of the assumptions may be relaxed. However, the interpretation of the correlation is much more difficult.

The confidence interval for this correlation is calculated using the Fisher's z transformation of the rank correlation.

The two-sided hypothesis test and probability level are for testing whether the rank correlation is zero.

R-Squared

R^2, officially known as the coefficient of determination, is defined as

$$R^2 = \frac{SS_{Model}}{SS_{Total}}$$

R^2 is probably the most popular statistical measure of how well the regression model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship between the *Y* and *X*, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *X* and *Y* was a perfect circle. The R^2 value of this relationship would be zero.

2. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.

3. *No-intercept model*. The definition of R^2 assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of R^2 changes dramatically. The fact that your R^2 value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying meaning of R^2.

4. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Adjusted R-Squared

This is an adjusted version of R^2. The adjustment seeks to remove the distortion due to a small sample size.

$$R^2_{adjusted} = 1 - \left(1 - R^2\right)\left(\frac{N - 1}{N - 2}\right)$$

Analysis of Variance Section

```
                       Sum of          Mean                  Prob     Power
Source         DF     Squares        Square     F-Ratio     Level      (5%)
Intercept       1     77128.2       77128.2
Slope           1    1318.337      1318.337    669.1468    0.0000    1.0000
Error          18    35.46317      1.970176
  Lack of Fit  16    34.96317      2.185198      8.7408    0.1074
  Pure Error    2         0.5          0.25
Adj. Total     19      1353.8      71.25263
Total          20       78482
s = Square Root(1.970176) = 1.40363
```

An analysis of variance (ANOVA) table summarizes the information related to the sources of variation in data.

Source

This represents the partitions of the variation in *Y*. There are four sources of variation listed: intercept, slope, error, and total (adjusted for the mean).

DF

The degrees of freedom are the number of dimensions associated with this term. Note that each observation can be interpreted as a dimension in *N*-dimensional space. The degrees of freedom for the intercept, model, error, and adjusted total are 1, *1*, *N* - 2, and *N* - 1, respectively.

Sum of Squares

These are the sums of squares associated with the corresponding sources of variation. Note that these values are in terms of the dependent variable, *Y*. The formulas for each are

$$SS_{intercept} = N\overline{Y}^2$$

$$SS_{slope} = \sum\left(\hat{Y} - \overline{Y}\right)^2$$

$$SS_{error} = \sum\left(Y - \hat{Y}\right)^2$$

$$SS_{total} = \sum\left(Y - \overline{Y}\right)^2$$

Note that the *lack of fit* and *pure error* values are provided if there are observations with identical values of the independent variable.

Mean Square

The mean square is the sum of squares divided by the degrees of freedom. This mean square is an estimated variance. For example, the mean square error is the estimated variance of the residuals (the residuals are sometimes called the *errors*).

F-Ratio

This is the *F* statistic for testing the null hypothesis that the slope equals zero. This *F*-statistic has 1 degree of freedom for the numerator variance and *N* - 2 degrees of freedom for the denominator variance.

Prob Level

This is the *p*-value for the above *F* test. The *p*-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the *p*-value is less than alpha, say 0.05, the null hypothesis is rejected. If the *p*-value is greater than alpha, the null hypothesis is accepted.

Power(5%)

Power is the probability of rejecting the null hypothesis that the slope is zero when it is not.

S = Root Mean Square Error

s is the square root of the mean square error. It is an estimate of the standard deviation of the residuals.

Summary Matrices

```
                        X'X          X'X          X'Y    X'X Inverse    X'X Inverse
Index                     0            1            2              0              1
0                        20         2792         1242      0.6015912  -3.951227E-03
1                      2792       425094       180208  -3.951227E-03   2.830392E-05
2 (Y'Y)                                         78482
Determinant                       706616                               1.415196E-06

Variance - Covariance Matrix of Regression Coefficients
                      VC(b)          VC(b)
Index                     0              1
0                  1.185241  -7.784612E-03
1             -7.784612E-03   5.576369E-05
```

This section provides the matrices from which the least square regression values are calculated. Occasionally, these values may be useful in hand calculations.

Tests of Assumptions Section

```
                                                         Is the Assumption
                                       Test       Prob   Reasonable at the 0.2000
Assumption/Test                       Value      Level   Level of Significance?
Residuals follow Normal Distribution?
Shapiro Wilk                         0.9728   0.812919   Yes
Anderson Darling                     0.2652   0.694075   Yes
D'Agostino Skewness                 -0.9590   0.337543   Yes
D'Agostino Kurtosis                  0.1205   0.904066   Yes
D'Agostino Omnibus                   0.9343   0.626796   Yes

Constant Residual Variance?
Modified Levene Test                 0.0946   0.761964   Yes

Relationship is a Straight Line?
Lack of Linear Fit F(16, 2) Test     8.7408   0.107381   No

No Serial Correlation?
Evaluate the Serial-Correlation report and the Durbin-Watson test if you have
equal-spaced, time series data.
```

This report presents numeric tests of some of the assumptions made when using linear regression. The results of these tests should be compared to an appropriate plot to determine if the assumption is valid or not.

Note that a 'Yes' means that there is not enough evidence to reject the assumption. This lack of assumption test rejection may be because the sample size is too small or the assumptions of the test were no met. It does not necessarily mean that the data met assumption. Likewise, a 'No' may occur because the sample size is very large. It is almost always possible to fail a preliminary test given a large enough sample size. No assumption is every fits perfectly. Bottom line, you should also investigate plots designed to check the assumptions.

Residuals follow Normal Distribution?

This section displays the results of five normality tests of the residuals. The Shapiro-Wilk and Anderson-Darling tests are usually considered as the best.

Unfortunately, these tests have small statistical power (probability of detecting nonnormal data) unless the sample sizes are large, say over 300. Hence, if the decision is to reject normality, you can be reasonably certain that the data are not normal. However, if the decision is not to reject, the situation is not as clear. If you have a sample size of 300 or more, you can reasonably assume that the actual distribution is closely approximated by the normal distribution. If your sample size

is less than 300, all you know for sure is that there was not enough evidence in your data to reject the normality of residuals assumption. In other words, the data might be nonnormal, you just could not prove it. In this case, you must rely on the graphics to justify the normality assumption.

Shapiro-Wilk W Test

This test for normality, developed by Shapiro and Wilk (1965), has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of *N* observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The denominator is the sum of squares of the observations about the sample mean. *W* may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator. Since these weights are asymptotically proportional to the corresponding expected normal order statistics, *W* is roughly a measure of the straightness of the normal quantile-quantile plot. Hence, the closer *W* is to one, the more normal the sample is.

The probability values for *W* are valid for samples in the range of 3 to 5000.

The test is not calculated when a frequency variable is specified.

Anderson-Darling Test

This test, developed by Anderson and Darling (1954), is based on EDF statistics. In some situations, it has been found to be as powerful as the Shapiro-Wilk test.

The test is not calculated when a frequency variable is specified.

D’Agostino Skewness

D’Agostino (1990) proposed a normality test based on the skewness coefficient, $\sqrt{b_1}$. Because the normal distribution is symmetrical, $\sqrt{b_1}$ is equal to zero for normal data. Hence, a test can be developed to determine if the value of $\sqrt{b_1}$ is significantly different from zero. If it is, the data are obviously nonnormal. The test statistic is, under the null hypothesis of normality, approximately normally distributed. The computation of this statistic is restricted to sample sizes greater than 8. The formula and further details are given in the Descriptive Statistics chapter.

D’Agostino Kurtosis

D’Agostino (1990) proposed a normality test based on the kurtosis coefficient, b_2. For the normal distribution, the theoretical value of b_2 is 3. Hence, a test can be developed to determine if the value of b_2 is significantly different from 3. If it is, the residuals are obviously nonnormal. The test statistic is, under the null hypothesis of normality, approximately normally distributed for sample sizes $N > 20$. The formula and further details are given in the Descriptive Statistics chapter.

D’Agostino Omnibus

D’Agostino (1990) proposed a normality test that combines the tests for skewness and kurtosis. The statistic, K^2, is approximately distributed as a chi-square with two degrees of freedom.

Constant Residual Variance?

Linear regression assumes that the residuals have constant variance. The validity of this assumption can be checked by looking at a plot of the absolute values of the residuals versus the *X* variable. The modified Levene test may be used when a numerical answer is needed.

If your data fail this test, you may want to use a logarithm transformation or a weighted regression.

Modified Levene Test

The *modified Levene test* can be used to evaluated the validity of the assumption of constant variance. It has been shown to be reliable even when the residuals do not follow a normal distribution. The mathematical details of the test were presented earlier in this chapter.

Relationship is a Straight Line?

Linear regression assumes that the relationship between *X* and *Y* is a straight line (linear). The validity of this assumption can be checked by looking at the plot *Y* versus *X* and at the plot of the residuals versus *X*. The lack of fit test may be used when a numerical answer is needed.

If your data fail this test, you may want to use a different model which accounts for the curvature. The Growth and Other Models procedure in curve fitting is a good choice when curvature exists in your data.

Lack of Linear Fit Test

The *lack-of-fit* test is used to test for a departure from the linear fit. This test requires that there are multiple observations for at least one *X* value. When such is the case, an estimate of *pure error* and *lack of fit* can be found and an *F* test created. The mathematical details of the test were presented earlier in this chapter.

Serial Correlation and Durbin-Watson Sections

```
Serial Correlation of Residuals Section
        Serial                Serial                 Serial
Lag     Correlation     Lag   Correlation     Lag    Correlation
1       0.1029          9     -0.2353         17
2       -0.4127*        10    -0.0827         18
3       0.0340          11    -0.0316         19
4       0.2171          12    -0.0481         20
5       -0.1968         13    0.0744          21
6       -0.0194         14    0.0073          22
7       0.2531          15                    23
8       -0.0744         16                    24

Durbin-Watson Test For Serial Correlation
                                                    Did the Test Reject
Parameter                                  Value      H0: Rho(1) = 0?
Durbin-Watson Value                        1.6978
Prob. Level: Positive Serial Correlation   0.2366                  No
Prob. Level: Negative Serial Correlation   0.7460                  No
```

This section reports on the autocorrelation structure of the residuals. Of course, if your data were not taken through time, this section should be ignored.

Lag

The lag, *k*, is the number of periods back.

Serial Correlation

The serial correlation reported here is the sample autocorrelation coefficient of lag *k*. It is computed as

$$r_k = \frac{\sum e_{i-k} e_i}{\sum e_i^2} \qquad \mathit{for}\ k = 1,2,\ldots,24$$

The distribution of these autocorrelations may be approximated by the distribution of the regular correlation coefficient. Using this fact, Fisher's *Z* transformation may be used to find large autocorrelations. If the Fisher's *Z* transformation of the autocorrelation is greater than 1.645, the autocorrelation is assumed to be large and the observation is starred.

Durbin-Watson Value

The Durbin-Watson test is often used to test for positive or negative, first-order, serial correlation. It is calculated as follows

$$DW = \frac{\sum_{j=2}^{N} \left(e_j - e_{j-1}\right)^2}{\sum_{j=1}^{N} e_j^2}$$

The distribution of this test is mathematically difficult because it involves the *X* values. Originally, Durbin-Watson (1950, 1951) gave a pair of bounds to be used. However, there is a large range of indecision that can be found when using these bounds. Instead of using these bounds, ***NCSS*** calculates the exact probability using the beta distribution approximation suggested by Durbin-Watson (1951). This approximation has been shown to be accurate to three decimal places in most cases.

PRESS Section

```
                                From          From
                               PRESS       Regular
Parameter                  Residuals     Residuals
Sum of Squared Residuals    43.15799      35.46317
Sum of |Residuals|          24.27421      22.02947
R-Squared                     0.9681        0.9738
```

This section reports on the PRESS statistics. The regular statistics, computed on all of the data, are provided to the side to make comparison between corresponding values easier.

Sum of Squared Residuals

PRESS is an acronym for prediction sum of squares. It was developed for use in variable selection to validate a regression model. To calculate *PRESS*, each observation is individually omitted. The remaining *N* - 1 observations are used to calculate a regression and estimate the value of the omitted observation. This is done *N* times, once for each observation. The difference between the actual *Y* value and the predicted *Y* with the observation deleted is called the prediction error or *PRESS* residual. The sum of the squared prediction errors is the *PRESS* value. The smaller *PRESS* is, the better the predictability of the model.

Sum of |Press residuals|

This is the sum of the absolute value of the *PRESS* residuals or prediction errors. If a large value for the *PRESS* is due to one or a few large *PRESS* residuals, this statistic may be a more accurate way to evaluate predictability.

Press R-Squared

The PRESS value above can be used to compute an R^2 -like statistic, called *R2Predict*, which reflects the prediction ability of the model. This is a good way to validate the prediction of a regression model without selecting another sample or splitting your data. It is very possible to have a high R^2 and a very low *R2Predict*. When this occurs, it implies that the fitted model is data dependent. This *R2Predict* ranges from below zero to above one. When outside the range of zero to one, it is truncated to stay within this range.

Predicted Values and Confidence Limits Section

```
                 Predicted    Standard     Lower 95%    Upper 95%
Weight           Height       Error        Confidence   Confidence
(X)              (Yhat|X)     of Yhat      Limit of Y|X Limit of Y|X
90.0000          52.5188      0.4855       51.4989      53.5388
100.0000         54.4505      0.4312       53.5446      55.3565
150.0000         64.1090      0.3233       63.4297      64.7882
200.0000         73.7674      0.5495       72.6129      74.9218
250.0000         83.4258      0.8821       81.5725      85.2791
```

The predicted values and confidence intervals of the mean response of *Y* given *X* are provided here. The values of *X* used here were specified in the *Predict Y at these X Values* option on the *Variables* panel.

It is important to note that violations of any regression assumptions will invalidate this interval estimate.

X

This is the value of *X* at which the prediction is made.

Predicted Y (Yhat|X)

The predicted value of *Y* for the value of *X* indicated.

Standard Error of Yhat

This is the estimated standard deviation of the predicted value.

Lower 95% Confidence Limit of Y|X

This is the lower limit of a 95% confidence interval estimate of the mean of *Y* at this value of *X*.

Upper 95% Confidence Limit of Y|X

This is the upper limit of a 95% confidence interval estimate of the mean of *Y* at this value of *X*. Note that you set the alpha level on the *Variables* panel.

Predicted Values and Prediction Limits Section

```
               Predicted      Standard       Lower 95%      Upper 95%
Weight         Height         Error          Prediction     Prediction
(X)            (Yhat|X)       of Yhat        Limit of Y|X   Limit of Y|X
90.0000        52.5188        1.4852         49.3985        55.6392
100.0000       54.4505        1.4684         51.3656        57.5355
150.0000       64.1090        1.4404         61.0828        67.1351
200.0000       73.7674        1.5074         70.6005        76.9342
250.0000       83.4258        1.6578         79.9429        86.9087
```

The predicted values and prediction intervals of the response of *Y* given *X* are provided here. The values of *X* used here were specified in the *Predict Y at these X Values* option on the *Variables* panel.

It is important to note that violations of any regression assumptions will invalidate this interval estimate.

X

This is the value of *X* at which the prediction is made.

Predicted Y (Yhat|X)

The predicted value of *Y* for the value of *X* indicated.

Standard Error of Yhat

This is the estimated standard deviation of the predicted value.

Lower 95% Prediction Limit of Y|X

This is the lower limit of a 95% prediction interval estimate of the mean of *Y* at this value of *X*.

Upper 95% Prediction Limit of Y|X

This is the upper limit of a 95% prediction interval estimate of the mean of *Y* at this value of *X*. Note that you set the alpha level on the *Variables* panel.

Residual Plots

The residuals can be graphically analyzed in numerous ways. For certain, the regression analyst should examine all of the basic residual graphs: the histogram, the density trace, the normal probability plot, the serial correlation plots (for time series data), the scatter plot of the residuals versus the sequence of the observations (for time series data), and the scatter plot of the residuals versus the independent variable.

For the scatter plots of residuals versus either the predicted values of *Y* or the independent variables, Hoaglin (1983) explains that there are several patterns to look for. You should note that these patterns are very difficult, if not impossible, to recognize for small data sets.

Point Cloud

A point cloud, basically in the shape of a rectangle or a horizontal band, would indicate no relationship between the residuals and the variable plotted against them. This is the preferred condition.

Wedge

An increasing or decreasing wedge would be evidence that there is increasing or decreasing (nonconstant) variation. A transformation of *Y* may correct the problem, or weighted least squares may be needed.

Bowtie

This is similar to the wedge above in that the residual plot shows a decreasing wedge in one direction while simultaneously having an increasing wedge in the other direction. A transformation of *Y* may correct the problem, or weighted least squares may be needed.

Sloping Band

This kind of residual plot suggests adding a linear version of the independent variable to the model.

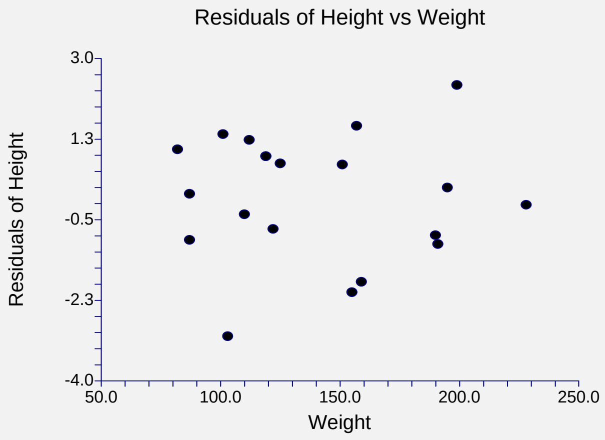
Curved Band

This kind of residual plot may be indicative of a nonlinear relationship between *Y* and the independent variable that was not accounted for. The solution might be to use a transformation on *Y* to create a linear relationship with *X*. Another possibility might be to add quadratic or cubic terms of a particular independent variable.

Curved Band with Increasing or Decreasing Variability

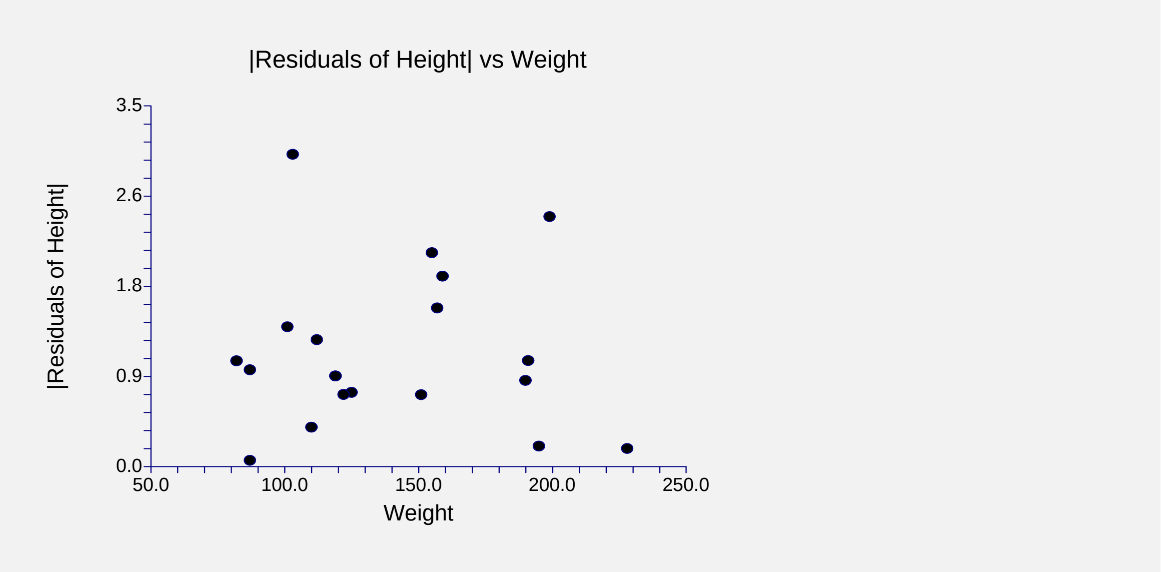
This residual plot is really a combination of the wedge and the curved band. It too must be avoided.

Residual vs X Plot



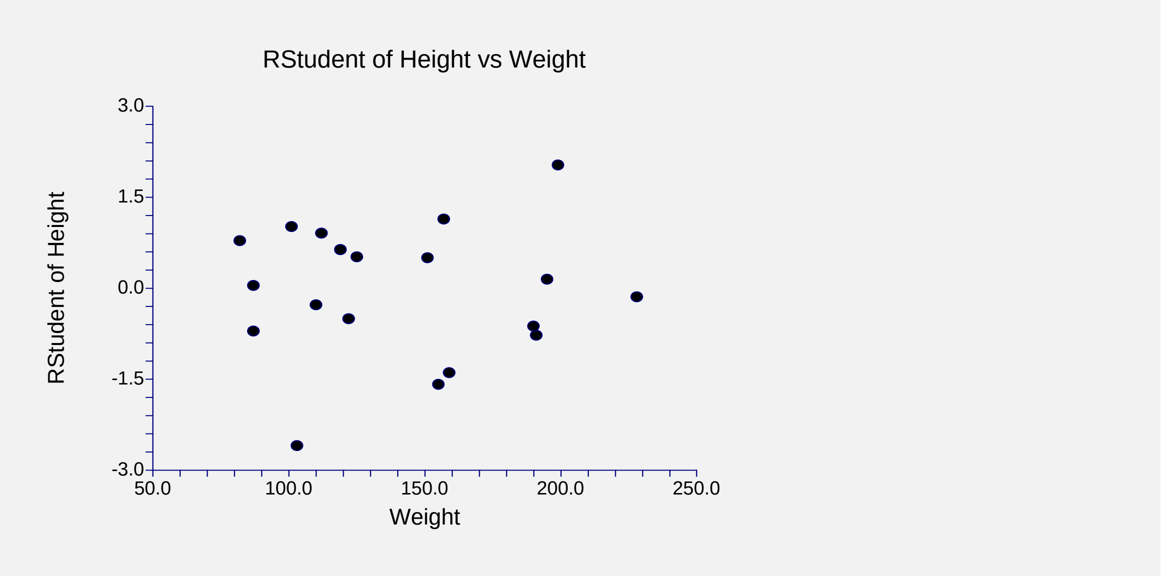
This plot is useful for showing nonlinear patterns and outliers. The preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.

|Residual| vs X Plot



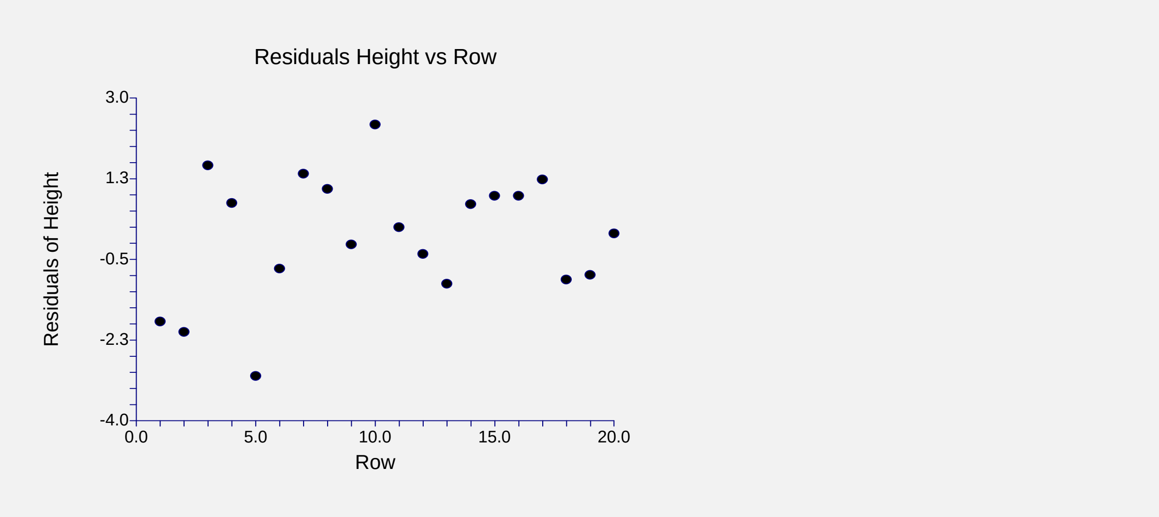
This plot is useful for showing nonconstant variance in the residuals. The preferred pattern is a rectangular shape or point cloud. The most common type of nonconstant variance occurs when the variance is proportion to *X*. This is shown by a funnel shape. Remedies for nonconstant variances were discussed earlier.

RStudent vs X Plot



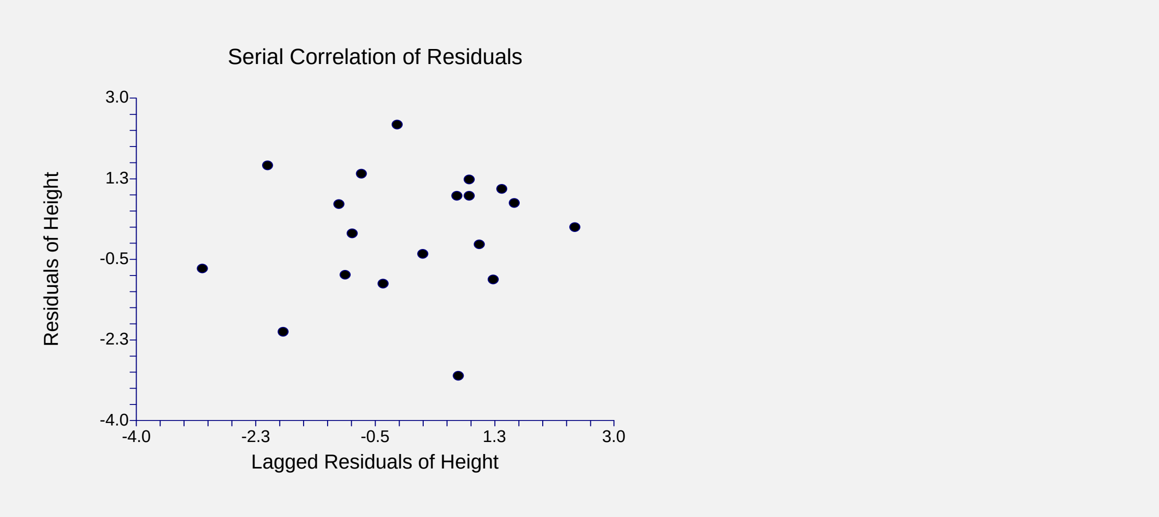
This is a scatter plot of the RStudent residuals versus the independent variable. The preferred pattern is a rectangular shape or point cloud. This plot is helpful in identifying any outliers.

Sequence Plot



Sequence plots may be useful in finding variables that are not accounted for by the regression equation. They are especially useful if the data were taken over time.

Serial Correlation of Residuals Plot



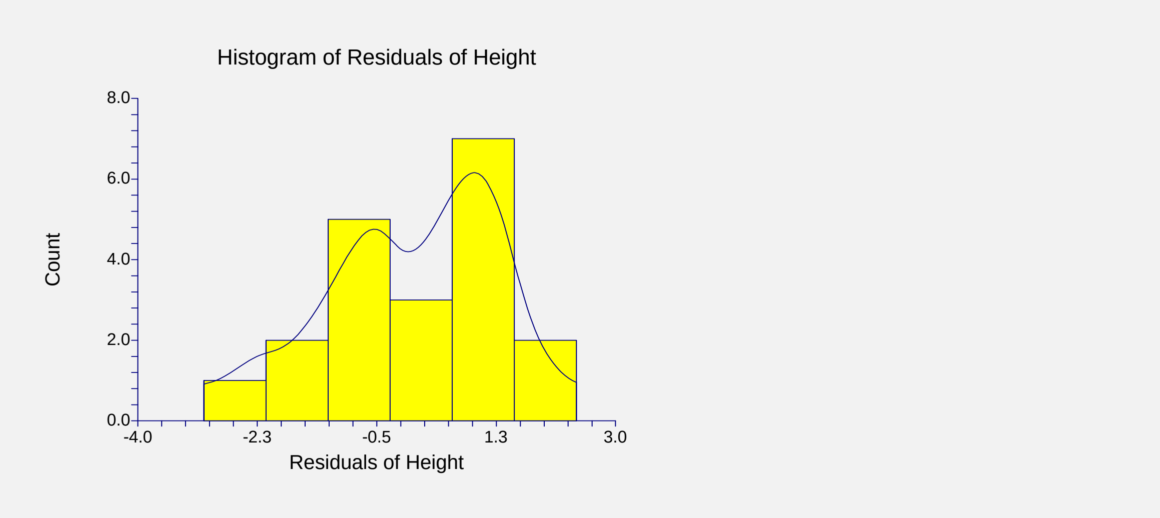
This is a scatter plot of the i^{th} residual versus the i^{th}-1 residual. It is only useful for time series data where the order of the rows on the database is important.

The purpose of this plot is to check for first-order autocorrelation. You would like to see a random pattern, i.e., a rectangular or uniform distribution of the points. A strong positive or negative trend indicates a need to redefine the model with some type of autocorrelation component.

Positive autocorrelation or serial correlation means that the residual in time period *t* tends to have the same sign as the residual in time period (*t* - 1). On the other hand, a strong negative autocorrelation means that the residual in time period *t* tends to have the opposite sign as the residual in time period (*t* - 1).

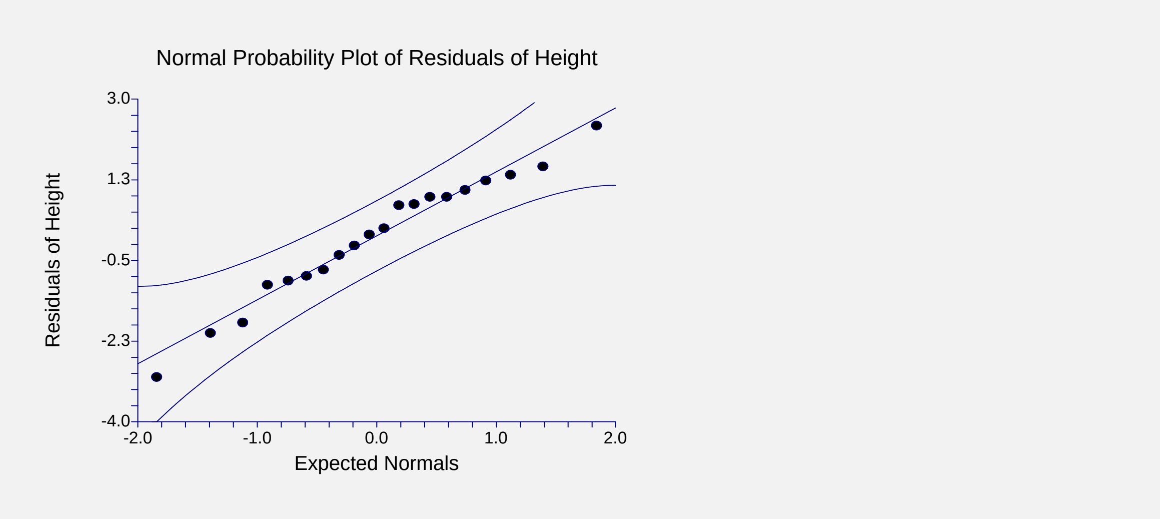
Be sure to check the Durbin-Watson statistic.

Histogram



The purpose of the histogram and density trace of the residuals is to evaluate whether they are normally distributed. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating normality of the residuals. The better choice would be the normal probability plot.

Probability Plot of Residuals



If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers. Curvature at both ends of the plot indicates long or short distributional tails. Convex, or concave, curvature indicates a lack of symmetry. Gaps, plateaus, or segmentation indicate clustering and may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is unwise.

If the residuals are not normally distributed, the *t*-tests on regression coefficients, the *F*-tests, and the interval estimates are not valid. This is a critical assumption to check.

Original Data Section

```
                                             Predicted
           Weight            Height             Height
Row           (X)               (Y)            (Yhat|X)         Residual
1        159.0000           64.0000            65.8475          -1.8475
2        155.0000           63.0000            65.0748          -2.0748
3        157.0000           67.0000            65.4611           1.5389
4        125.0000           60.0000            59.2797           0.7203
5        103.0000           52.0000            55.0300          -3.0300
6        122.0000           58.0000            58.7002          -0.7002
.               .                 .                  .                .
.               .                 .                  .                .
.               .                 .                  .                .
```

This report lists the values of *X*, *Y*, the predicted value of *Y*, and the residual.

Predicted Values of Means Section

```
                                   Predicted      Standard      Lower 95%      Upper 95%
           Weight       Height        Height         Error     Conf. Limit    Conf. Limit
Row           (X)          (Y)       (Yhat|X)       of Yhat    of Y Mean|X    of Y Mean|X
1        159.0000      64.0000       65.8475        0.3457        65.1212        66.5737
2        155.0000      63.0000       65.0748        0.3343        64.3725        65.7771
3        157.0000      67.0000       65.4611        0.3397        64.7475        66.1748
4        125.0000      60.0000       59.2797        0.3323        58.5817        59.9778
5        103.0000      52.0000       55.0300        0.4162        54.1557        55.9044
6        122.0000      58.0000       58.7002        0.3403        57.9854        59.4151
.               .            .             .             .              .              .
.               .            .             .             .              .              .
.               .            .             .             .              .              .
```

The predicted values and confidence intervals of the mean response of *Y* given *X* are given for each observation.

X

This is the value of *X* at which the prediction is made.

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

The predicted value of *Y* for the value of *X* indicated.

Standard Error of Yhat

This is the estimated standard deviation of the predicted mean value.

Lower 95% Confidence Limit of Y|X

This is the lower limit of a 95% confidence interval estimate of the mean of *Y* at this value of *X*.

Upper 95% Confidence Limit of Y|X

This is the upper limit of a 95% confidence interval estimate of the mean of *Y* at this value of *X*. Note that you set the alpha level on the *Variables* panel.

Predicted Values and Prediction Limits Section

```
                                 Predicted    Standard    Lower 95%    Upper 95%
           Weight      Height       Height       Error   Prediction   Prediction
Row           (X)         (Y)    (Yhat|X)     of Yhat   Limit of Y|X Limit of Y|X
1        159.0000     64.0000     65.8475      1.4456      62.8104      68.8845
2        155.0000     63.0000     65.0748      1.4429      62.0434      68.1062
3        157.0000     67.0000     65.4611      1.4441      62.4271      68.4952
4        125.0000     60.0000     59.2797      1.4424      56.2493      62.3101
5        103.0000     52.0000     55.0300      1.4640      51.9542      58.1058
6        122.0000     58.0000     58.7002      1.4443      55.6659      61.7346
.               .           .           .           .            .            .
.               .           .           .           .            .            .
.               .           .           .           .            .            .
```

The predicted values and confidence intervals of the mean response of *Y* given *X* are given for each observation.

X

This is the value of *X* at which the prediction is made.

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

The predicted value of *Y* for the value of *X* indicated.

Standard Error of Yhat

This is the estimated standard deviation of the predicted value suitable for creating a prediction limit for an individual.

Lower 95% Prediction Limit of Y|X

This is the lower limit of a 95% prediction interval estimate of *Y* at this value of *X*.

Upper 95% Prediction Limit of Y|X

This is the upper limit of a 95% prediction interval estimate of *Y* at this value of *X*. Note that you set the alpha level on the *Variables* panel.

Working-Hotelling Simultaneous Confidence Band

Row	Weight (X)	Height (Y)	Predicted Height (Yhat|X)	Standard Error of Yhat	Lower 95% Conf. Band of Y Mean|X	Upper 95% Conf. Band of Y Mean|X
1	159.0000	64.0000	65.8475	0.3457	63.3900	68.3050
2	155.0000	63.0000	65.0748	0.3343	62.6985	67.4511
3	157.0000	67.0000	65.4611	0.3397	63.0462	67.8761
4	125.0000	60.0000	59.2797	0.3323	56.9177	61.6418
5	103.0000	52.0000	55.0300	0.4162	52.0713	57.9887
6	122.0000	58.0000	58.7002	0.3403	56.2812	61.1192
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

The predicted values and confidence band of the mean response function are given for each observation. Note that this is a confidence band for all possible values of *X* along the real number line. The confidence coefficient is the proportion of time that this procedure yields a band that includes the true regression line when a large number of samples are taken using the *X* values as in this sample.

X

This is the value of *X* at which the prediction is made.

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

The predicted value of *Y* for the value of *X* indicated.

Standard Error of Yhat

This is the estimated standard deviation of the predicted mean value.

Lower 95% Confidence Band of Y|X

This is the lower limit of the 95% confidence band for the value of *Y* at this *X*.

Upper 95% Confidence Band of Y|X

This is the upper limit of the 95% confidence band for the value of *Y* at this *X*.

Residual Section

```
                                 Predicted                                Percent
              Weight     Height     Height              Standardized     Absolute
Row              (X)        (Y)    (Yhat|X)   Residual      Residual        Error
1           159.0000    64.0000    65.8475    -1.8475       -1.3580       2.8867
2           155.0000    63.0000    65.0748    -2.0748       -1.5220       3.2933
3           157.0000    67.0000    65.4611     1.5389        1.1299       2.2968
4           125.0000    60.0000    59.2797     0.7203        0.5282       1.2004
5           103.0000    52.0000    55.0300    -3.0300       -2.2604       5.8270
6           122.0000    58.0000    58.7002    -0.7002       -0.5142       1.2073
.                  .          .          .          .             .            .
.                  .          .          .          .             .            .
.                  .          .          .          .             .            .
```

This is a report showing the value of the residual at each observation.

X

This is the value of *X* at which the prediction is made.

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

The predicted value of *Y* for the value of *X* indicated.

Residual

This is the difference between the actual and predicted values of *Y*.

Standardized Residual

The variance of the observed residuals is not constant. This makes comparisons among the residuals difficult. One solution is to standardize the residuals by dividing by their standard deviations. This gives a set of residuals with constant variance.

The formula for this residual is

$$r_j = \frac{e_j}{s\sqrt{1-h_{jj}}}$$

Percent Absolute Error

The percent is the absolute value of the *Residual* divided by the *Actual* value. Scrutinize observations with the large percent errors.

Residual Diagnostics Section

```
              Weight                                   Hat
Row              (X)     Residual     RStudent    Diagonal      Cook's D         MSEi
1           159.0000      -1.8475      -1.3931      0.0607        0.0595       1.8723
2           155.0000      -2.0748      -1.5845      0.0567        0.0696       1.8176
3           157.0000       1.5389       1.1392      0.0586        0.0397       1.9381
4           125.0000       0.7203       0.5173      0.0560        0.0083       2.0537
5           103.0000      -3.0300     *-2.5957      0.0879        0.2462       1.4939
6           122.0000      -0.7002      -0.5034      0.0588        0.0083       2.0554
.                  .            .            .           .             .            .
.                  .            .            .           .             .            .
.                  .            .            .           .             .            .
```

This is a report gives residual diagnostics for each observation. These were discussed earlier in the technical of this chapter and we refer you to that section for the technical details.

X

This is the value of *X* at which the prediction is made.

Residual

This is the difference between the actual and predicted values of *Y*.

RStudent

Sometimes called the externally studentized residual, *RStudent* is a standardized residual that has the impact of a single observation removed from the mean square error. If the regression assumption of normality is valid, a single value of the RStudent has a *t* distribution with *N* - 2 degrees of freedom.

An observation is starred as an outlier if the absolute value of RStudent is greater than 2.

Hat Diagonal

The hat diagonal captures an observation's remoteness in the *X*-space. Some authors refer to the hat diagonal as a measure of *leverage* in the *X*-space.

Hat diagonals greater than 4 / *N* are considered influential. However, an influential observation is not a bad observation. An influential observation should be checked to determine if it is also an outlier.

Cook's D

Cook's D attempts to measure the influence the observation on all *N* fitted values. The formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j \left[\hat{y}_j - \hat{y}_j(i)\right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. A Cook's *D* value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

MSEi

This is the value of the mean squared error calculated without observation *j*.

Leave One Row Out Section

```
Row   RStudent      DFFITS     Cook's D    CovRatio    DFBETAS(0)  DFBETAS(1)
1      -1.3931     -0.3540      0.0595      0.9615        0.0494     -0.1483
2      -1.5845     -0.3885      0.0696      0.9023        0.0228     -0.1337
3       1.1392      0.2842      0.0397      1.0279       -0.0284      0.1087
4       0.5173      0.1260      0.0083      1.1511        0.0739     -0.0414
5    * -2.5957     -0.8059      0.2462      0.6304       -0.6820      0.5292
6      -0.5034     -0.1258      0.0083      1.1564       -0.0800      0.0486
.            .           .           .           .             .           .
.            .           .           .           .             .           .
.            .           .           .           .             .           .
```

Each column gives the impact on some aspect of the linear regression of omitting that row.

RStudent

Sometimes called the externally studentized residual, *RStudent* is a standardized residual that has the impact of a single observation removed from the mean square error. If the regression assumption of normality is valid, a single value of the RStudent has a *t* distribution with *N* - 2 degrees of freedom.

An observation is starred as an outlier if the absolute value of RStudent is greater than 2.

Dffits

Dffits is the standardized difference between the predicted value of *Y* with and without observation *j*. It represents the number of estimated standard errors that the predicted value changes if that observation is omitted. Dffits > 1 would flag observations as being influential in prediction.

Cook's D

Cook's D attempts to measure the influence the observation on all *N* fitted values. The formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j \left[\hat{y}_j - \hat{y}_j(i)\right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. A Cook's *D* value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

CovRatio

This diagnostic flags observations that have a major impact on the generalized variance of the regression coefficients. A value exceeding 1.0 implies that the observation provides an improvement, i.e., a reduction in the generalized variance of the coefficients. A value of CovRatio less than 1.0 flags an observation that increases the estimated generalized variance. This is not a favorable condition.

DFBETAS(0) and DFBETAS(1)

DFBETAS(0) and *DFBETAS(1)* are the standardized change in the intercept and slope when an observation is omitted from the analysis. Belsley, Kuh, and Welsch (1980) recommend using a cutoff of $2/\sqrt{N}$ when *N* is greater than 100. When *N* is less than 100, others have suggested using a cutoff of 1.0 or 2.0 for the absolute value of *DFBETAS*.

Outlier Detection Chart

```
             Weight                          Standardized
Row             (X)     Residual                 Residual                 RStudent
1          159.0000     -1.8475  |.............   -1.3580  |.............    -1.3931  |.............
2          155.0000     -2.0748  |.............   -1.5220  |.............    -1.5845  |.............
3          157.0000      1.5389  |.............    1.1299  |.............     1.1392  |.............
4          125.0000      0.7203  |.............    0.5282  |.............     0.5173  |.............
5          103.0000     -3.0300  |.............   -2.2604  |.............  * -2.5957  |.............
6          122.0000     -0.7002  |.............   -0.5142  |.............    -0.5034  |.............
.                 .           .              .          .              .           .              .
.                 .           .              .          .              .           .              .
.                 .           .              .          .              .           .              .
```

Outliers are rows that are far removed from the rest of the data. Since outliers can have dramatic effects on the results, corrective action, such as elimination, must be carefully considered. Outlying rows should not be removed unless a good reason for their removal can be given.

An outlier may be defined as a row in which |RStudent| > 2. Rows with this characteristic have been starred.

X

This is the value of *X*.

Residual

This is the difference between the actual and predicted values of *Y*.

Standardized Residual

The variance of the observed residuals is not constant. This makes comparisons among the residuals difficult. One solution is to standardize the residuals by dividing by their standard deviations. This gives a set of residuals with constant variance.

RStudent

Sometimes called the externally studentized residual, *RStudent* is a standardized residual that has the impact of a single observation removed from the mean square error. If the regression assumption of normality is valid, a single value of the RStudent has a *t* distribution with *N* - 2 degrees of freedom.

An observation is starred as an outlier if the absolute value of RStudent is greater than 2.

Influence Detection Chart

```
           Weight
Row           (X)     DFFITS                     Cook's D                 DFBETAS(1)
1        159.0000    -0.3540  |.............       0.0595  |.............    -0.1483  |.............
2        155.0000    -0.3885  |.............       0.0696  |.............    -0.1337  |.............
3        157.0000     0.2842  |.............       0.0397  |.............     0.1087  |.............
4        125.0000     0.1260  |.............       0.0083  |.............    -0.0414  |.............
5        103.0000    -0.8059  ||............       0.2462  ||||||........     0.5292  |||...........
6        122.0000    -0.1258  |.............       0.0083  |.............     0.0486  |.............
.               .          .              .             .              .          .              .
.               .          .              .             .              .          .              .
.               .          .              .             .              .          .              .
```

Influential rows are those whose omission results in a relatively large change in the results. They are not necessarily harmful. However, they will distort the results if they are also outliers. The impact of influential rows should be studied very carefully. The accuracy of the data values should be double-checked.

X

This is the value of *X*.

Dffits

Dffits is the standardized difference between the predicted value of *Y* with and without observation *j*. It represents the number of estimated standard errors that the predicted value changes if that observation is omitted. Dffits > 1 would flag observations as being influential in prediction.

Cook's D

Cook's D attempts to measure the influence the observation on all *N* fitted values. The formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j \left[\hat{y}_j - \hat{y}_j(i) \right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. A Cook's *D* value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

DFBETAS(1)

DFBETAS(1) is the standardized change in the slope when an observation is omitted from the analysis. Belsley, Kuh, and Welsch (1980) recommend using a cutoff of $2/\sqrt{N}$ when *N* is greater than 100. When *N* is less than 100, others have suggested using a cutoff of 1.0 or 2.0 for the absolute value of *DFBETAS*.

Outlier & Influence Detection Chart

```
                                                                          Hat
          Weight   RStudent                        Cooks D           Diagonal
Row          (X)   (Outlier)                    (Influence)         (Leverage)
1       159.0000    -1.3931  |.............        0.0595 |.............   0.0607 |.............
2       155.0000    -1.5845  |.............        0.0696 |.............   0.0567 |.............
3       157.0000     1.1392  |.............        0.0397 |.............   0.0586 |.............
4       125.0000     0.5173  |.............        0.0083 |.............   0.0560 |.............
5       103.0000  * -2.5957  |.............        0.2462 ||||||........   0.0879 |.............
6       122.0000    -0.5034  |.............        0.0083 |.............   0.0588 |.............
.              .          .               .             .             .         .             .
.              .          .               .             .             .         .             .
.              .          .               .             .             .         .             .
```

This report provides diagnostics about whether a row is an outlier, influential, and has high leverage. Outliers are rows that are removed from the rest of the data. Influential rows are those whose omission results in a relatively large change in the results. This report lets you see both.

X

This is the value of *X*.

RStudent (Outlier)

RStudent is a standardized residual that has the impact of a single observation removed from the mean square error. If the regression assumption of normality is valid, a single value of the RStudent has a *t* distribution with *N* - 2 degrees of freedom.

An observation is starred as an outlier if the absolute value of RStudent is greater than 2.

Cook's D (Influence)

Cook's D attempts to measure the influence the observation on all *N* fitted values. The formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j \left[\hat{y}_j - \hat{y}_j(i) \right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. A Cook's *D* value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

Hat Diagonal (Leverage)

The hat diagonal captures an observation's remoteness in the *X*-space. Some authors refer to the hat diagonal as a measure of *leverage* in the *X*-space.

Hat diagonals greater than 4 / *N* are considered influential. However, an influential observation is not a bad observation. An influential observation should be checked to determine if it is also an outlier.

Inverse Prediction of X Means

```
                                   Predicted                     Lower 95%    Upper 95%
            Height       Weight       Weight                     Conf. Limit  Conf. Limit
Row            (Y)          (X)      (Xhat|Y)      X-Xhat|Y    of X Mean|Y  of X Mean|Y
1          64.0000     159.0000     149.4360        9.5640      145.9832     153.0193
2          63.0000     155.0000     144.2591       10.7409      140.8441     147.7361
3          67.0000     157.0000     164.9664       -7.9664      161.1310     169.1387
4          60.0000     125.0000     128.7287       -3.7287      125.1181     132.1948
5          52.0000     103.0000      87.3141       15.6859       81.4894      92.4444
6          58.0000     122.0000     118.3750        3.6250      114.3947     122.0735
.                .            .            .             .             .            .
.                .            .            .             .             .            .
.                .            .            .             .             .            .
```

This report provides inverse prediction or calibration results. Although a regression of *Y* on *X* has been fit, our interest here is predicting the value of *X* from the value of *Y*. This report provides both a point estimate and an interval estimate of the predicted mean of *X* given *Y*.

Y

This is the actual value of *Y*.

X

This is the value of *X* at which the prediction is made.

Predicted X (Xhat|Y)

The predicted value of *X* for the value of *Y* indicated.

Lower 95% Confidence Limit of X Mean|Y

This is the lower limit of a 95% confidence interval estimate of the mean of *X* at this value of *Y*.

Upper 95% Confidence Limit of X Mean|Y

This is the upper limit of a 95% confidence interval estimate of the mean of *X* at this value of *Y*.

Inverse Prediction of X Individuals

```
                                      Predicted                     Lower 95%     Upper 95%
              Height       Weight        Weight                    Prediction    Prediction
Row              (Y)          (X)      (Xhat|Y)      X-Xhat|Y    Limit of X|Y  Limit of X|Y
1            64.0000     159.0000      149.4360        9.5640        133.7858      165.2167
2            63.0000     155.0000      144.2591       10.7409        128.5906      159.9896
3            67.0000     157.0000      164.9664       -7.9664        149.3036      180.9662
4            60.0000     125.0000      128.7287       -3.7287        112.9365      144.3765
5            52.0000     103.0000       87.3141       15.6859         70.7003      103.2335
6            58.0000     122.0000      118.3750        3.6250        102.4436      134.0246
.                  .            .             .             .               .             .
.                  .            .             .             .               .             .
.                  .            .             .             .               .             .
```

This report provides inverse prediction or calibration results. Although a regression of *Y* on *X* has been fit, our interest here is predicting the value of *X* from the value of *Y*. This report provides both a point estimate and an interval estimate of the predicted value of *X* given *Y*.

Y

This is the actual value of *Y*.

X

This is the value of *X* at which the prediction is made.

Predicted X (Xhat|Y)

The predicted value of *X* for the value of *Y* indicated.

Lower 95% Prediction Limit of X|Y

This is the lower limit of a 95% prediction interval estimate of *X* at this value of *Y*.

Upper 95% Prediction Limit of X|Y

This is the upper limit of a 95% prediction interval estimate of *X* at this value of *Y*.

Chapter 305

Multiple Regression

Introduction

Multiple Regression Analysis refers to a set of techniques for studying the straight-line relationships among two or more variables. Multiple regression estimates the $\beta' s$ in the equation

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + \varepsilon_j$$

The *X's* are the *independent variables* (IV's). *Y* is the *dependent variable*. The subscript *j* represents the observation (row) number. The $\beta' s$ are the unknown *regression coefficients*. Their estimates are represented by *b's*. Each β represents the original unknown (population) parameter, while *b* is an estimate of this β. The ε_j is the error (residual) of observation *j*.

Although the regression problem may be solved by a number of techniques, the most-used method is least squares. In least squares regression analysis, the *b's* are selected so as to minimize the sum of the squared residuals. This set of *b's* is not necessarily the set you want, since they may be distorted by *outliers*--points that are not representative of the data. Robust regression, an alternative to least squares, seeks to reduce the influence of outliers.

Multiple regression analysis studies the relationship between a *dependent* (response) *variable* and *p independent variables* (*predictors, regressors, IV's*). The sample multiple regression equation is

$$\hat{y}_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \ldots + b_p x_{pj}$$

If p = 1, the model is called *linear regression.*

The intercept, b_0, is the point at which the regression plane intersects the *Y* axis. The b_i are the slopes of the regression plane in the direction of x_i. These coefficients are called the partial-regression coefficients. Each partial regression coefficient represents the net effect the i^{th} variable has on the dependent variable, holding the remaining *X's* in the equation constant.

A large part of a regression analysis consists of analyzing the sample *residuals*, e_j, defined as

$$e_j = y_j - \hat{y}_j$$

Once the $\beta' s$ have been estimated, various indices are studied to determine the reliability of these estimates. One of the most popular of these reliability indices is the correlation coefficient. The correlation coefficient, or simply the correlation, is an index that ranges from -1 to 1. When the value is near zero, there is no linear relationship. As the correlation gets closer to plus or minus one, the relationship is stronger. A value of one (or negative one) indicates a perfect linear relationship between two variables.

The regression equation is only capable of measuring linear, or straight-line, relationships. If the data form a circle, for example, regression analysis would not detect a relationship. For this reason, it is always advisable to plot each independent variable with the dependent variable, watching for curves, outlying points, changes in the amount of variability, and various other anomalies that may occur.

If the data are a random sample from a larger population and the $\varepsilon' s$ are independent and normally distributed, a set of statistical tests may be applied to the *b's* and the correlation coefficient. These *t*-tests and *F*-tests are valid only if the above assumptions are met.\

Regression Models

In order to make good use of multiple regression, you must have a basic understanding of the regression model. The basic regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

This expression represents the relationship between the dependent variable (DV) and the independent variables (IV's) as a weighted average in which the regression coefficients ($\beta' s$) are the weights. Unlike the usual weights in a weighted average, it is possible for the regression coefficients to be negative.

A fundamental assumption in this model is that the effect of each IV is additive. Now, no one really believes that the true relationship is actually additive. Rather, they believe that this model is a reasonable first-approximation to the true model. To add validity to this approximation, you might consider this additive model to be a Taylor-series expansion of the true model. However, this appeal to the Taylor-series expansion usually ignores the 'local-neighborhood' assumption.

Another assumption is that the relationship of the DV with each IV is linear (straight-line). Here again, no one really believes that the relationship is a straight-line. However, this is a reasonable first approximation.

In order obtain better approximations, methods have been developed to allow regression models to approximate curvilinear relationships as well as nonadditivity. Although nonlinear regression models can be used in these situations, they add a higher level of complexity to the modeling process. An experienced user of multiple regression knows how to include curvilinear components in a regression model when it is needed.

Another issue is how to add categorical variables into the model. Unlike regular numeric variables, categorical variables may be alphabetic. Examples of categorical variables are gender, producer, and location. In order to effectively use multiple regression, you must know how to include categorical IV's in your regression model.

This section shows how ***NCSS*** may be used to specify and estimate advanced regression models that include curvilinearity, interaction, and categorical variables.

Representing a Curvilinear Relationship

A curvilinear relationship between a DV and one or more IV's is often modeled by adding new IV's which are created from the original IV by squaring, and occasionally cubing, them. For example, the regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

might be expanded to

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 \\ &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 \end{aligned}$$

Note that this model is still additive in terms of the new IV's.

One way to adopt such a new model is to create the new IV's using the transformations of existing variables. However, the same effect can be achieved using the Custom Model statement. The details of writing a Custom Model will be presented later, but we note in passing that the above model would be written as

$$\mathrm{X}_1 \quad \mathrm{X}_2 \quad \mathrm{X}_1 * \mathrm{X}_1 \quad \mathrm{X}_1 * \mathrm{X}_2 \quad \mathrm{X}_2 * \mathrm{X}_2$$

Representing Categorical Variables

Categorical variables take on only a few unique values. For example, suppose a therapy variable has three possible values: A, B, and C. One question is how to include this variable in the regression model. At first glance, we can convert the letters to numbers by recoding A to 1, B to 2, and C to 3. Now we have numbers. Unfortunately, we will obtain completely different results if we recode A to 2, B to 3, and C to 1. Thus, a direct recode of letters to numbers will not work.

To convert a categorical variable to a form usable in regression analysis, we have to create a new set of numeric variables. If a categorical variable has *k* values, *k* - 1 new variables must be generated.

There are many ways in which these new variables may be generated. We will present a few examples here.

Indicator Variables

Indicator (dummy or binary) variables are a popular type of generated variables. They are created as follows. A *reference value* is selected. Usually, the most common value is selected as the reference value. Next, a variable is generated for each of the values other than the reference value. For example, suppose that C is selected as the reference value. An indicator variable is generated for each of the remaining values: A and B. The value of the indicator variable is one if the value of the original variable is equal to the value of interest, or zero otherwise. Here is how the original variable T and the two new indicator variables TA and TB look in a short example.

T	TA	TB
A	1	0
A	1	0
B	0	1
B	0	1
C	0	0
C	0	0

The generated IV's, TA and TB, would be used in the regression model.

Contrast Variables

Contrast variables are another popular type of generated variables. Several types of contrast variables can be generated. We will present a few here. One method is to contrast each value with the reference value. The value of interest receives a one. The reference value receives a negative one. All other values receive a zero.

Continuing with our example, one set of contrast variables is

T	CA	CB
A	1	0
A	1	0
B	0	1
B	0	1
C	-1	-1
C	-1	-1

The generated IV's, CA and CB, would be used in the regression model.

Another set of contrast variables that is commonly used is to compare each value with those remaining. For this example, we will suppose that T takes on four values: A, B, C, and D. The generate variables are

T	C1	C2	C3
A	-3	0	0
A	-3	0	0
B	1	-2	0
B	1	-2	0
C	1	1	-1
C	1	1	-1
D	1	1	1
D	1	1	1

Many other methods have been developed to provide meaningful numeric variables that represent categorical variable. We have presented these because they may be generated automatically by ***NCSS***.

Representing Interactions of Numeric Variables

The interaction between two variables is represented in the regression model by creating a new variable that is the product of the variables that are interacting. Suppose you have two variables *X1* and *X2* for which an interaction term is necessary. A new variable is generated by multiplying the values of *X1* and *X2* together.

X1	X2	Int
1	1	1
2	1	2
3	2	6
2	2	4
0	4	0
5	-2	-10

The new variable, *Int*, is added to the regression equation and treated like any other variable during the analysis. With *Int* in the regression model, the interaction between *X1* and *X2* may be investigated.

Representing Interactions of Numeric and Categorical Variables

When the interaction between a numeric IV and a categorical IV is to be included in the model, all proceeds as above, except that an interaction variable must be generated for each categorical variable. This can be accomplished automatically in ***NCSS*** using an appropriate Model statement.

In the following example, the interaction between the categorical variable *T* and the numeric variable *X* is created.

T	CA	CB	X	XCA	XCB
A	1	0	1.2	1.2	0
A	1	0	1.4	1.4	0
B	0	1	2.3	0	2.3
B	0	1	4.7	0	4.7
C	-1	-1	3.5	-3.5	-3.5
C	-1	-1	1.8	-1.8	-1.8

When the variables *XCA* and *XCB* are added to the regression model, they will account for the interaction between *T* and *X*.

Representing Interactions Two or More Categorical Variables

When the interaction between two categorical variables is included in the model, an interaction variable must be generated for each combination of the variables generated for each categorical variable. This can be accomplished automatically in ***NCSS*** using an appropriate Model statement.

In the following example, the interaction between the categorical variables *T* and *S* are generated. Try to determine the reference value used for variable *S*.

T	CA	CB	S	S1	S2	CAS1	CAS2	CBS1	CBS2
A	1	0	D	1	0	1	0	0	0
A	1	0	E	0	1	0	1	0	0
B	0	1	F	0	0	0	0	0	0
B	0	1	D	1	0	0	0	1	0
C	-1	-1	E	0	1	0	-1	0	-1
C	-1	-1	F	0	0	0	0	0	0

When the variables, *CAS1*, *CAS2*, *CBS1*, and *CBS2* are added to the regression model, they will account for the interaction between *T* and *S*.

Possible Uses of Regression Analysis

Montgomery (1982) outlines the following five purposes for running a regression analysis.

Description

The analyst is seeking to find an equation that describes or summarizes the relationships in a set of data. This purpose makes the fewest assumptions.

Coefficient Estimation

This is a popular reason for doing regression analysis. The analyst may have a theoretical relationship in mind, and the regression analysis will confirm this theory. Most likely, there is specific interest in the magnitudes and signs of the coefficients. Frequently, this purpose for regression overlaps with others.

Prediction

The prime concern here is to predict some response variable, such as sales, delivery time, efficiency, occupancy rate in a hospital, reaction yield in some chemical process, or strength of some metal. These predictions may be very crucial in planning, monitoring, or evaluating some process or system. There are many assumptions and qualifications that must be made in this case. For instance, you must not extrapolate beyond the range of the data. Also, interval estimates require special, so-called normality, assumptions to hold.

Control

Regression models may be used for monitoring and controlling a system. For example, you might want to calibrate a measurement system or keep a response variable within certain guidelines. When a regression model is used for control purposes, the independent variables must be related to the dependent in a causal way. Furthermore, this functional relationship must continue over time. If it does not, continual modification of the model must occur.

Variable Selection or Screening

In this case, a search is conducted for those independent variables that explain a significant amount of the variation in the dependent variable. In most applications, this is not a one-time process but a continual model-building process. This purpose is manifested in other ways, such as using historical data to identify factors for future experimentation.

Assumptions

The following assumptions must be considered when using multiple regression analysis.

Linearity

Multiple regression models the linear (straight-line) relationship between *Y* and the *X's*. Any curvilinear relationship is ignored. This is most easily evaluated by scatter plots early on in your analysis. Nonlinear patterns can show up in residual plots.

Constant Variance

The variance of the $\varepsilon' s$ is constant for all values of the *X's*. This can be detected by residual plots of e_j versus $\hat{y}_j$ or the *X's*. If these residual plots show a rectangular shape, we can assume constant variance. On the other hand, if a residual plot shows an increasing or decreasing wedge or bowtie shape, nonconstant variance exists and must be corrected.

Special Causes

We assume that all special causes, outliers due to one-time situations, have been removed from the data. If not, they may cause nonconstant variance, nonnormality, or other problems with the regression model.

Normality

We assume the $\varepsilon's$ are normally distributed when hypothesis tests and confidence limits are to be used.

Independence

The $\varepsilon's$ are assumed to be uncorrelated with one another, which implies that the *Y's* are also uncorrelated. This assumption can be violated in two ways: model misspecification or time-sequenced data.

1. *Model misspecification.* If an important independent variable is omitted or if an incorrect functional form is used, the residuals may not be independent. The solution to this dilemma is to find the proper functional form or to include the proper independent variables.
2. *Time-sequenced data*. Whenever regression analysis is performed on data taken over time (frequently called time series data), the residuals are often correlated. This correlation among residuals is called serial correlation or autocorrelation. Positive autocorrelation means that the residual in time period *j* tends to have the same sign as the residual in time period (*j*-*k*), where *k* is the lag in time periods. On the other hand, negative autocorrelation means that the residual in time period *j* tends to have the opposite sign as the residual in time period (*j*-*k*).

The presence of autocorrelation among the residuals has several negative impacts:

1. The regression coefficients are unbiased but no longer efficient, i.e., minimum variance estimates.
2. With positive serial correlation, the mean square error may be seriously underestimated. The impact of this is that the standard errors are underestimated, the partial t-tests are inflated (show significance when there is none), and the confidence intervals are shorter than they should be.
3. Any hypothesis tests or confidence limits that required the use of the t or F distribution would be invalid.

You could try to identify these serial correlation patterns informally, with the residual plots versus time. A better analytical way would be to compute the serial or autocorrelation coefficient for different time lags and compare it to a critical value.

Multicollinearity

Collinearity, or multicollinearity, is the existence of near-linear relationships among the set of independent variables. The presence of multicollinearity causes all kinds of problems with regression analysis, so you could say that we assume the data do not exhibit it.

Effects of Multicollinearity

Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial *t*-tests for the regression coefficients, give false nonsignificant p-values, and degrade the predictability of the model.

Sources of Multicollinearity

To deal with collinearity, you must be able to identify its source. The source of the collinearity impacts the analysis, the corrections, and the interpretation of the linear model. There are five sources (see Montgomery [1982] for details):

1. *Data collection*. In this case, the data has been collected from a narrow subspace of the independent variables. The collinearity has been created by the sampling methodology. Obtaining more data on an expanded range would cure this collinearity problem.
2. *Physical constraints* of the linear model or population. This source of collinearity will exist no matter what sampling technique is used. Many manufacturing or service processes have constraints on independent variables (as to their range), either physically, politically, or legally, which will create collinearity.
3. *Over-defined model*. Here, there are more variables than observations. This situation should be avoided.
4. *Model choice or specification*. This source of collinearity comes from using independent variables that are higher powers or interactions of an original set of variables. It should be noted that if sampling subspace of X_j is narrow, then any combination of variables with x_j will increase the collinearity problem even further.
5. *Outliers*. Extreme values or outliers in the *X*-space can cause collinearity as well as hide it.

Detection of Collinearity

The following steps for detecting collinearity proceed from simple to complex.

1. Begin by studying pairwise scatter plots of pairs of independent variables, looking for near-perfect relationships. Also glance at the correlation matrix for high correlations. Unfortunately, multicollinearity does not always show up when considering the variables two at a time.
2. Next, consider the variance inflation factors (*VIF*). Large *VIF*'s flag collinear variables.
3. Finally, focus on small eigenvalues of the correlation matrix of the independent variables. An eigenvalue of zero or close to zero indicates that an exact linear dependence exists. Instead of looking at the numerical size of the eigenvalue, use the condition number. Large condition numbers indicate collinearity.

Correction of Collinearity

Depending on what the source of collinearity is, the solutions will vary. If the collinearity has been created by the data collection, then collect additional data over a wider *X*-subspace. If the choice of the linear model has accented the collinearity, simplify the model by variable selection techniques. If an observation or two has induced the collinearity, remove those observations and proceed accordingly. Above all, use care in selecting the variables at the outset.

Centering and Scaling Issues in Collinearity

When the variables in regression are centered (by subtracting their mean) and scaled (by dividing by their standard deviation), the resulting $X'X$ matrix is in correlation form. The centering of each independent variable has removed the constant term from the collinearity diagnostics. Scaling and centering permit the computation of the collinearity diagnostics on standardized variables. On the other hand, there are many regression applications where the intercept is a vital part of the linear model. The collinearity diagnostics on the uncentered data may provide a more realistic picture of the collinearity structure in these cases.

Multiple Regression Checklist

This checklist, prepared by a professional statistician, is a flowchart of the steps you should complete to conduct a valid multiple regression analysis. Several of these steps should be performed prior to this phase of the regression analysis, but they are briefly listed here again as a reminder. You should complete these tasks in order.

Step 1 – Data Preparation

Scan your data for anomalies, keypunch errors, typos, and so on. You should have a minimum of five observations for each variable in the analysis, including the dependent variable. This discussion assumes that the pattern of missing values is random. All data preparation should be done prior to the use of one of the variable selection strategies.

Special attention must be paid to categorical IV's to make certain that you have chosen a reasonable method of converting them to numeric values.

Also, you must decide how complicated of a model to use. Do you want to include powers of variables and interactions between terms?

One the best ways to accomplish this data preparation is to run your data through the Data Screening procedure, since it provides reports about missing value patterns, discrete and continuous variables, and so on.

Step 2 – Variable Selection

Variable selection seeks to reduce the number of IV's to a manageable few. There are several variable selection methods in regression: Stepwise Regression, All Possible Regressions, or Multivariate Variable Selection. Each of these variable selection methods has advantages and disadvantages. We suggest that you begin with the Hierarchical Stepwise procedure included in this procedure since it allows you to look at interactions, powers, and categorical variables. Use this to narrow your search down to fifteen or fewer IV's. Next, apply All Possible Regressions to those fifteen variables to find the best four or five variables.

It is extremely important that you complete Step 1 before beginning this step, since variable selection can be greatly distorted by outliers. Every effort should be taken to find outliers before beginning this step.

Step 3 – Setup and Run the Regression

Introduction

Now comes the fun part: running the program. **NCSS** is designed to be simple to operate, but it can still seem complicated. When you go to run a procedure such as this for the first time, take a few minutes to read through the chapter again and familiarize yourself with the issues involved.

Enter Variables

The ***NCSS*** panels are set with ready-to-run defaults, but you have to select the appropriate variables (columns of data). There should be only one dependent variable and one or more independent variables enumerated. In addition, if a weight variable is available from a previous analysis, it needs to be specified.

Choose Report Options

In multiple linear regression, there is a wide assortment of report options available. As a minimum, you are interested in the coefficients for the regression equation, the analysis of variance report, normality testing, serial correlation (for time-sequenced data), regression diagnostics (looking for outliers), and multicollinearity insights.

Specify Alpha

Most beginners at statistics forget this important step and let the alpha value default to the standard 0.05. You should make a conscious decision as to what value of alpha is appropriate for your study. The 0.05 default came about during the dark ages when people had to rely on printed probability tables and there were only two values available: 0.05 or 0.01. Now you can set the value to whatever is appropriate.

Select All Plots

As a rule, select all residual plots. They add a great deal to your analysis of the data.

Step 4 – Check Model Adequacy

Introduction

Once the regression output is displayed, you will be tempted to go directly to the probability of the *F*-test from the regression analysis of variance table to see if you have a significant result. However, it is very important that you proceed through the output in an orderly fashion. The main conditions to check for relate to linearity, normality, constant variance, independence, outliers, multicollinearity, and predictability. Return to the statistical sections and plot descriptions for more detailed discussions.

Check 1. Linearity

- Look at the Residual vs. Predicted plot. A curving pattern here indicates nonlinearity.
- Look at the Residual vs. Predictor plots. A curving pattern here indicates nonlinearity.
- Look at the *Y* versus X plots. For simple linear regression, a linear relationship between Y and X in a scatter plot indicates that the linearity assumption is appropriate. The same holds if the dependent variable is plotted against each independent variable in a scatter plot.
- If linearity does not exist, take the appropriate action and return to Step 2. Appropriate action might be to add power terms (such as Log(X), X squared, or X cubed) or to use an appropriate nonlinear model.

Check 2. Normality

- Look at the *Normal Probability Plot*. If all of the residuals fall within the confidence bands for the *Normal Probability Plot*, the normality assumption is likely met. One or two residuals outside the confidence bands may be an indicator of outliers, not nonnormality.
- Look at the *Normal Assumptions Section*. The formal normal goodness of fit tests are given in the *Normal Assumptions Section*. If the decision is accepted for the *Normality (Omnibus)* test, there is no evidence that the residuals are not normal.

- If normality does not exist, take the appropriate action and return to Step 2. Appropriate action includes removing outliers and/or using the logarithm of the dependent variable.

Check 3. Nonconstant Variance

- Look at the Residual vs. Predicted plot. If the Residual vs. Predicted plot shows a rectangular shape instead of an increasing or decreasing wedge or a bowtie, the variance is constant.
- Look at the Residual vs. Predictor plots. If the Residual vs. Predictor plots show a rectangular shape, instead of an increasing or decreasing wedge or a bowtie, the variance is constant.
- If nonconstant variance does not exist, take the appropriate action and return to Step 2. Appropriate action includes taking the logarithm of the dependent variable or using weighted regression.

Check 4. Independence or Serial Correlation

- If you have time series data, look at the Serial-Correlations Section. If none of the serial correlations in the Serial-Correlations Section are greater than the critical value that is provided, independence may be assumed.
- Look at the Residual vs. Row plot. A visualization of what the Serial-Correlations Section shows will be exhibited by adjacent residuals being similar (a roller coaster trend) or dissimilar (a quick oscillation).
- If independence does not exist, use a first difference model and return to Step 2. More complicated choices require time series models.

Check 5. Outliers

- Look at the Regression Diagnostics Section. Any observations with an asterisk by the diagnostics RStudent, Hat Diagonal, DFFITS, or the CovRatio, are potential outliers. Observations with a Cook's *D* greater than 1.00 are also potentially influential.
- Look at the Dfbetas Section. Any Dfbetas beyond the cutoff of $\pm 2/\sqrt{N}$ indicate influential observations.
- Look at the Rstudent vs. Hat Diagonal plot. This plot will flag an observation that may be jointly influential by both diagnostics.
- If outliers do exist in the model, go to robust regression and run one of the options there to confirm these outliers. If the outliers are to be deleted or down weighted, return to Step 2.

Check 6. Multicollinearity

- Look at the Multicollinearity Section. If any variable has a variance inflation factor greater than 10, collinearity could be a problem.
- Look at the Eigenvalues of Centered Correlations Section. Condition numbers greater than 1000 indicate severe collinearity. Condition numbers between 100 and 1000 imply moderate to strong collinearity.
- Look at the Correlation Matrix Section. Strong pairwise correlation here may give some insight as to the variables causing the collinearity.

- If multicollinearity does exist in the model, it could be due to an outlier (return to Check 5 and then Step 2) or due to strong interdependencies between independent variables. In the latter case, return to Step 2 and try a different variable selection procedure.

Check 7. Predictability

- Look at the PRESS Section. If the Press R2 is almost as large as the R2, you have done as well as could be expected. It is not unusual in practice for the Press R2 to be half of the R2. If R2 is 0.50, a Press R2 of 0.25 would be unacceptable.
- Look at the Predicted Values with Confidence Limits for Means and Individuals. If the confidence limits are too wide to be practical, you may need to add new variables or reassess the outlier and collinearity possibilities.
- Look at the Residual Report. Any observation that has percent error grossly deviant from the values of most observations is an indication that this observation may be impacting predictability.
- Any changes in the model due to poor predictability require a return to Step 2.

Step 5 – Record Your Results

Since multiple regression can be quite involved, it is best make notes of why you did what you did at different steps of the analysis. Jot down what decisions you made and what you have found. Explain what you did, why you did it, what conclusions you reached, which outliers you deleted, areas for further investigation, and so on. Be sure to examine the following sections closely and in the indicated order:

1. Analysis of Variance Section. Check for the overall significance of the model.
2. Regression Equation and Coefficient Sections. Significant individual variables are noted here.

Regression analysis is a complicated statistical tool that frequently demands revisions of the model. Your notes of the analysis process as well as of the interpretation will be worth their weight in gold when you come back to an analysis a few days later!

Multiple Regression Technical Details

This section presents the technical details of least squares regression analysis using a mixture of summation and matrix notation. Because this module also calculates weighted multiple regression, the formulas will include the weights, w_j. When weights are not used, the w_j are set to one.

Define the following vectors and matrices:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_N \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} \cdots x_{1p} \\ \vdots & \\ 1 & x_{1j} \cdots x_{pj} \\ \vdots & \\ 1 & x_{1N} \cdots x_{pN} \end{bmatrix}, \ \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_N \end{bmatrix}, \ \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & \vdots \\ 0 & 0 & w_j & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & w_N \end{bmatrix}$$

Least Squares

Using this notation, the least squares estimates are found using the equation.

$$\mathbf{b} = (\mathbf{X'WX})^{-1}\mathbf{X'\,WY}$$

Note that when the weights are not used, this reduces to

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'\,Y}$$

The predicted values of the dependent variable are given by

$$\hat{\mathbf{Y}} = \mathbf{b'\,X}$$

The residuals are calculated using

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Estimated Variances

An estimate of the variance of the residuals is computed using

$$s^2 = \frac{\mathbf{e'\,We}}{N - p - 1}$$

An estimate of the variance of the regression coefficients is calculated using

$$\mathrm{V}\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = s^2(\mathbf{X'\,WX})^{-1}$$

An estimate of the variance of the predicted mean of *Y* at a specific value of *X*, say X_0, is given by

$$s^2_{Y_m|X_0} = s^2(1, X_0)(\mathbf{X'\,WX})^{-1}\begin{pmatrix} 1 \\ X_0 \end{pmatrix}$$

An estimate of the variance of the predicted value of *Y* for an individual for a specific value of *X*, say X_0, is given by

$$s^2_{Y_I|X_0} = s^2 + s^2_{Y_m|X_0}$$

Hypothesis Tests of the Intercept and Slopes

Using these variance estimates and assuming the residuals are normally distributed, hypothesis tests may be constructed using the Student's *t* distribution with *N* - *p* - 1 degrees of freedom using

$$t_{b_i} = \frac{b_i - \mathrm{B}_i}{s_{b_i}}$$

Usually, the hypothesized value of B_i is zero, but this does not have to be the case.

Confidence Intervals of the Intercept and Slope

A $100(1-\alpha)\%$ confidence interval for the true regression coefficient, β_i, is given by

$$b_{\mathrm{i}} \pm \left(t_{1-\alpha/2,N-p-1}\right)s_{b_i}$$

Confidence Interval of Y for Given X

A $100(1-\alpha)\%$ confidence interval for the mean of *Y* at a specific value of *X*, say X_0, is given by

$$b'X_0 \pm \left(t_{1-\alpha/2,N-p-1}\right)s_{Y_m|X_0}$$

A $100(1-\alpha)\%$ prediction interval for the value of *Y* for an individual at a specific value of *X*, say X_0, is given by

$$b'X_0 \pm \left(t_{1-\alpha/2,N-p-1}\right)s_{Y_I|X_0}$$

R-Squared (Percent of Variation Explained)

Several measures of the goodness-of-fit of the regression model to the data have been proposed, but by far the most popular is R^2. R^2 is the square of the correlation coefficient between *Y* and $\hat{Y}$. It is the proportion of the variation in *Y* that is accounted by the variation in the independent variables. R^2 varies between zero (no linear relationship) and one (perfect linear relationship).

R^2, officially known as the *coefficient of determination*, is defined as the sum of squares due to the regression divided by the adjusted total sum of squares of *Y*. The formula for R^2 is

$$R^2 = 1 - \left(\frac{\mathbf{e'We}}{\mathbf{Y'WY} - \frac{(\mathbf{1'WY})^2}{\mathbf{1'W1}}} \right)$$

$$= \frac{SS_{Model}}{SS_{Total}}$$

R^2 is probably the most popular measure of how well a regression model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Additional independent variables*. It is possible to increase R^2 by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This usually happens when the sample size is small.

2. *Range of the independent variables*. R^2 is influenced by the range of the independent variables. R^2 increases as the range of the *X's* increases and decreases as the range of the *X's* decreases.
3. *Slope magnitudes*. R^2 does not measure the magnitude of the slopes.
4. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *X* and *Y* was a perfect sphere. Although there is a perfect relationship between the variables, the R^2 value would be zero.
5. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.
6. *No-intercept model*. The definition of R^2 assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of R^2 changes dramatically. The fact that your R^2 value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying definition of R^2.
7. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Rbar-Squared (Adjusted R-Squared)

R^2 varies directly with *N*, the sample size. In fact, when $N = p$, $R^2 = 1$. Because R^2 is so closely tied to the sample size, an adjusted R^2 value, called $\overline{R}^2$, has been developed. $\overline{R}^2$ was developed to minimize the impact of sample size. The formula for $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \frac{(N-1)(1-R^2)}{N-p-1}$$

Testing Assumptions Using Residual Diagnostics

Evaluating the amount of departure in your data from each assumption is necessary to see if remedial action is necessary before the fitted results can be used. First, the types of plots and statistical analyses the are used to evaluate each assumption will be given. Second, each of the diagnostic values will be defined.

Notation – Use of (j) and p

Several of these residual diagnostic statistics are based on the concept of studying what happens to various aspects of the regression analysis when each row is removed from the analysis. In what follows, we use the notation (*j*) to mean that observation *j* has been omitted from the analysis. Thus, *b*(*j*) means the value of *b* calculated without using observation *j*.

Some of the formulas depend on whether the intercept is fitted or not. We use *p* to indicate the number of regression parameters. When the intercept is fit, *p* will include the intercept.

1 – No Outliers

Outliers are observations that are poorly fit by the regression model. If outliers are influential, they will cause serious distortions in the regression calculations. Once an observation has been determined to be an outlier, it must be checked to see if it resulted from a mistake. If so, it must be corrected or omitted. However, if no mistake can be found, the outlier should not be discarded just because it is an outlier. Many scientific discoveries have been made because outliers, data points that were different from the norm, were studied more closely. Besides being caused by simple data-entry mistakes, outliers often suggest the presence of an important independent variable that has been ignored.

Outliers are easy to spot on scatter plots of the residuals and RStudent. RStudent is the preferred statistic for finding outliers because each observation is omitted from the calculation making it less likely that the outlier can mask its presence. Scatter plots of the residuals and RStudent against the X variables are also helpful because they may show other problems as well.

2 – Linear Regression Function - No Curvature

The relationship between Y and each X is assumed to be linear (straight-line). No mechanism for curvature is included in the model. Although scatter plots of Y versus each X can show curvature in the relationship, the best diagnostic tool is the scatter plot of the residual versus each X. If curvature is detected, the model must be modified to account for the curvature. This may mean adding a quadratic term, taking logarithms of Y or X, or some other appropriate transformation.

3 – Constant Variance

The errors are assumed to have constant variance across all values of X. If there are a lot of data ($N > 100$), nonconstant variance can be detected on the scatter plots of the residuals versus each X. However, the most direct diagnostic tool to evaluate this assumption is a scatter plot of the absolute values of the residuals versus each X. Often, the assumption is violated because the variance increases with X. This will show up as a ‘megaphone’ pattern on the scatter plot.

When nonconstant variance is detected, a variance-stabilizing transformation such as the square-root or logarithm may be used. However, the best solution is probably to use weighted regression, with weights inversely proportional to the magnitude of the residuals.

4 – Independent Errors

The Y’s, and thus the errors, are assumed to be independent. This assumption is usually ignored unless there is a reason to think that it has been violated, such as when the observations were taken across time. An easy way to evaluate this assumption is a scatter plot of the residuals versus their sequence number (assuming that the data are arranged in time sequence order). This plot should show a relative random pattern.

The Durbin-Watson statistic is used as a formal test for the presence of first-order serial correlation. A more comprehensive method of evaluation is to look at the autocorrelations of the residuals at various lags. Large autocorrelations are found by testing each using Fisher’s z transformation. Although Fisher’s z transformation is only approximate in the case of autocorrelations, it does provide a reasonable measuring stick with which to judge the size of the autocorrelations.

If independence is violated, confidence intervals and hypothesis tests are erroneous. Some remedial method that accounts for the lack of independence must be adopted, such as using first differences or the Cochrane-Orcutt procedure.

Durbin-Watson Test

The Durbin-Watson test is often used to test for positive or negative, first-order, serial correlation. It is calculated as follows

$$DW = \frac{\sum_{j=2}^{N}\left(e_j - e_{j-1}\right)^2}{\sum_{j=1}^{N} e_j^2}$$

The distribution of this test is difficult because it involves the X values. Originally, Durbin-Watson (1950, 1951) gave a pair of bounds to be used. However, there is a large range of ‘inclusion’ found when using these bounds. Instead of using these bounds, we calculate the exact probability using the beta distribution approximation suggested by Durbin-Watson (1951). This approximation has been shown to be accurate to three decimal places in most cases which is all that are needed for practical work.

5 – Normality of Residuals

The residuals are assumed to follow the normal probability distribution with zero mean and constant variance. This can be evaluated using a normal probability plot of the residuals. Also, normality tests are used to evaluate this assumption. The most popular of the five normality tests provided is the Shapiro-Wilk test.

Unfortunately, a breakdown in any of the other assumptions results in a departure from this assumption as well. Hence, you should investigate the other assumptions first, leaving this assumption until last.

Influential Observations

Part of the evaluation of the assumptions includes an analysis to determine if any of the observations have an extra large influence on the estimated regression coefficients, on the fit of the model, or on the value of Cook’s distance. By looking at how much removing an observation changes the results, an observation’s influence can be determined.

Five statistics are used to investigate influence. These are Hat diagonal, DFFITS, DFBETAS, Cook’s D, and COVARATIO.

Definitions Used in Residual Diagnostics

Residual

The residual is the difference between the actual *Y* value and the *Y* value predicted by the estimated regression model. It is also called the *error*, the *deviate*, or the *discrepancy*.

$$e_j = y_j - \hat{y}_j$$

Although the true errors, ε_j , are assumed to be independent, the computed residuals, e_j , are not. Although the lack of independence among the residuals is a concern in developing theoretical tests, it is not a concern on the plots and graphs.

By assumption, the variance of the ε_j is σ^2 . However, the variance of the e_j is not σ^2 . In vector notation, the covariance matrix of **e** is given by

$$\begin{aligned} \mathrm{V}(\mathbf{e}) &= \sigma^2\left(\mathbf{I} - \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'W}^{\frac{1}{2}}\right) \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

The matrix **H** is called the *hat matrix* since it puts the 'hat' on *y* as is shown in the unweighted case.

$$\begin{aligned} \hat{Y} &= \mathbf{Xb} \\ &= \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y} \\ &= \mathbf{HY} \end{aligned}$$

Hence, the variance of e_j is given by

$$\mathrm{V}(e_j) = \sigma^2(1 - h_{jj})$$

where h_{jj} is the jth diagonal element of **H**. This variance is estimated using

$$\hat{\mathrm{V}}(e_j) = s^2(1 - h_{jj})$$

Hat Diagonal

The hat diagonal, h_{jj}, is the jth diagonal element of the hat matrix, H where

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'W}^{\frac{1}{2}}$$

H captures an observation's remoteness in the *X*-space. Some authors refer to the hat diagonal as a measure of *leverage* in the *X*-space. As a rule of thumb, hat diagonals greater than 4/*N* are considered influential and are called high-leverage observations.

Note that a high-leverage observation is not a bad observation. Rather, high-leverage observations exert extra influence on the final results, so care should be taken to insure that they are correct. You should not delete an observation just because it has a high-influence. However, when you interpret the regression equation, you should bear in mind that the results may be due to a few, high-leverage observations.

Standardized Residual

As shown above, the variance of the observed residuals is not constant. This makes comparisons among the residuals difficult. One solution is to standardize the residuals by dividing by their standard deviations. This will give a set of residuals with constant variance.

The formula for this residual is

$$r_j = \frac{e_j}{s\sqrt{1-h_{jj}}}$$

s(j) or MSEi

This is the value of the mean squared error calculated without observation j. The formula for $s(j)$ is given by

$$s(j)^2 = \frac{1}{N-p-1}\sum_{i=1,i\neq j}^{N} w_i\left(y_i - \mathbf{x}_i'\mathbf{b}(j)\right)$$
$$= \frac{(N-p)s^2 - \frac{w_j e_j^2}{1-h_{jj}}}{N-p-1}$$

RStudent

Rstudent is similar to the studentized residual. The difference is the $s(j)$ is used rather than s in the denominator. The quantity $s(j)$ is calculated using the same formula as s, except that observation j is omitted. The hope is that be excluding this observation, a better estimate of σ^2 will be obtained. Some statisticians refer to these as the *studentized deleted residuals*.

$$t_j = \frac{e_j}{s(j)\sqrt{1-h_{jj}}}$$

If the regression assumptions of normality are valid, a single value of the RStudent has a t distribution with N - 2 degrees of freedom. It is reasonable to consider |RStudent| > 2 as outliers.

DFFITS

DFFITS is the standardized difference between the predicted value with and without that observation. The formula for *DFFITS* is

$$DFFITS_j = \frac{\hat{y}_j - \hat{y}_j(j)}{s(j)\sqrt{h_{jj}}}$$
$$= t_j\sqrt{\frac{h_{jj}}{1-h_{jj}}}$$

The values of $\hat{y}_j(j)$ and $s^2(j)$ are found by removing observation j before the doing the calculations. It represents the number of estimated standard errors that the fitted value changes if the j^{th} observation is omitted from the data set. If |*DFFITS*| > 1, the observation should be considered to be influential with regards to prediction.

Cook's D

The DFFITS statistic attempts to measure the influence of a single observation on its fitted value. Cook's distance (Cook's *D*) attempts to measure the influence each observation on all *N* fitted values. The formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j\left[\hat{y}_j - \hat{y}_j(i)\right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. Rather than go to all the time of recalculating the regression coefficients *N* times, we use the following approximation

$$D_j = \frac{w_j e_j^2 h_{jj}}{ps^2\left(1 - h_{jj}\right)^2}$$

This approximation is exact when no weight variable is used.

A Cook's *D* value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

CovRatio

This diagnostic flags observations that have a major impact on the generalized variance of the regression coefficients. A value exceeding 1.0 implies that the i^{th} observation provides an improvement, i.e., a reduction in the generalized variance of the coefficients. A value of CovRatio less than 1.0 flags an observation that increases the estimated generalized variance. This is not a favorable condition.

The general formula for the CovRatio is

$$\text{CovRatio}_j = \frac{\det\left[s(j)^2\left(\mathbf{X}(j)'\mathbf{W}\mathbf{X}(j)\right)^{-1}\right]}{\det\left[s^2\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\right]}$$

$$= \frac{1}{1 - h_{jj}}\left[\frac{s(j)^2}{s^2}\right]^p$$

Belsley, Kuh, and Welsch (1980) give the following guidelines for the CovRatio.

If CovRatio > 1 + 3*p* / *N* then omitting this observation significantly damages the precision of at least some of the regression estimates.

If CovRatio < 1 - 3*p* / *N* then omitting this observation significantly improves the precision of at least some of the regression estimates.

DFBETAS

The *DFBETAS* criterion measures the standardized change in a regression coefficient when an observation is omitted. The formula for this criterion is

$$DFBETAS_{kj} = \frac{b_k - b_k(j)}{s(j)\sqrt{c_{kk}}}$$

where c_{kk} is a diagonal element of the inverse matrix $(\mathbf{X'WX})^{-1}$.

Belsley, Kuh, and Welsch (1980) recommend using a cutoff of $2/\sqrt{N}$ when *N* is greater than 100. When *N* is less than 100, others have suggested using a cutoff of 1.0 or 2.0 for the absolute value of *DFBETAS*.

Press Value

PRESS is an acronym for prediction sum of squares. It was developed for use in variable selection to validate a regression model. To calculate *PRESS*, each observation is individually omitted. The remaining *N* - 1 observations are used to calculate a regression and estimate the value of the omitted observation. This is done *N* times, once for each observation. The difference between the actual *Y* value and the predicted *Y* with the observation deleted is called the prediction error or *PRESS* residual. The sum of the squared prediction errors is the *PRESS* value. The smaller *PRESS* is, the better the predictability of the model.

The formula for PRESS is

$$PRESS = \sum_{j=1}^{N} w_j \left[y_j - \hat{y}_j(j) \right]^2$$

Press R-Squared

The PRESS value above can be used to compute an R^2 -like statistic, called *R2Predict*, which reflects the prediction ability of the model. This is a good way to validate the prediction of a regression model without selecting another sample or splitting your data. It is very possible to have a high R^2 and a very low *R2Predict*. When this occurs, it implies that the fitted model is data dependent. This *R2Predict* ranges from below zero to above one. When outside the range of zero to one, it is truncated to stay within this range.

$$R^2_{predict} = 1 - \frac{PRESS}{SS_{tot}}$$

Sum |Press residuals|

This is the sum of the absolute value of the *PRESS* residuals or prediction errors. If a large value for the *PRESS* is due to one or a few large *PRESS* residuals, this statistic may be a more accurate way to evaluate predictability. This quantity is computed as

$$\sum |PRESS| = \sum_{j=1}^{N} w_j \left| y_j - \hat{y}_j(j) \right|$$

Bootstrapping

Bootstrapping was developed to provide standard errors and confidence intervals for regression coefficients and predicted values in situations in which the standard assumptions are not valid. In these nonstandard situations, bootstrapping is a viable alternative to the corrective action suggested earlier. The method is simple in concept, but it requires extensive computation time.

The bootstrap is simple to describe. You assume that your sample is actually the population and you draw *B* samples (*B* is over 1000) of size *N* from your original sample with replacement. With replacement means that each observation may be selected more than once. For each bootstrap sample, the regression results are computed and stored.

Suppose that you want the standard error and a confidence interval of the slope. The bootstrap sampling process has provided *B* estimates of the slope. The standard deviation of these *B* estimates of the slope is the bootstrap estimate of the standard error of the slope. The bootstrap confidence interval is found the arranging the *B* values in sorted order and selecting the appropriate percentiles from the list. For example, a 90% bootstrap confidence interval for the slope is given by fifth and ninety-fifth percentiles of the bootstrap slope values. The bootstrap method can be applied to many of the statistics that are computed in regression analysis.

The main assumption made when using the bootstrap method is that your sample approximates the population fairly well. Because of this assumption, bootstrapping does not work well for small samples in which there is little likelihood that the sample is representative of the population. Bootstrapping should only be used in medium to large samples.

When applied to linear regression, there are two types of bootstrapping that can be used.

Modified Residuals

Davison and Hinkley (1999) page 279 recommend the use of a special rescaling of the residuals when bootstrapping to keep results unbiased. These modified residuals are calculated using

$$e_j^* = \frac{e_j}{\sqrt{\dfrac{1-h_{jj}}{w_j}}} - \overline{e}^*$$

where

$$\overline{e}^* = \frac{\sum_{j=1}^{N} w_j e_j^*}{\sum_{j=1}^{N} w_j}$$

Bootstrap the Observations

The bootstrap samples are selected from the original sample. This method is appropriate for data in which both *X* and *Y* have been selected at random. That is, the *X* values were not predetermined, but came in as measurements just as the *Y* values.

An example of this situation would be if a population of individuals is sampled and both *Y* and *X* are measured on those individuals only after the sample is selected. That is, the value of *X* was not used in the selection of the sample.

Bootstrap Prediction Intervals

Bootstrap confidence intervals for the mean of *Y* given *X* are generated from the bootstrap sample in the usual way. To calculate prediction intervals for the predicted value (not the mean) of *Y* given *X* requires a modification to the predicted value of *Y* to be made to account for the variation of *Y* about its mean. This modification of the predicted *Y* values in the bootstrap sample, suggested by Davison and Hinkley, is as follows.

$$\hat{y}_{+} = \hat{y} - \sum x_i \left(b_i^* - b_i \right) + e_{+}^*$$

where e_{+}^* is a randomly selected modified residual. By adding the randomly sample residual we have added an appropriate amount of variation to represent the variance of individual *Y*’s about their mean value.

Subset Selection

Subset selection refers to the task of finding a small subset of the available independent variables that does a good job of predicting the dependent variable. Exhaustive searches are possible for regressions with up to 15 IV’s. However, when more than 15 IV’s are available, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

An issue that comes up because of categorical IV’s is what to do with the individual-degree of freedom variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms search on model terms rather than on the individual binary variables. Thus, the whole set of generated variables associated with a given term are considered together for inclusion in, or deletion from, the model. Its all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can save the generated set of variables in the first run and designate them as Numeric Variables.

Hierarchical Models

Another issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to consider only hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of *R*-Squared. Enter this term into the model.
3. Continue adding terms until a target value for *R*-Squared is achieved or until a preset limit on the maximum number of terms in the model is reached. Note that these terms can be limited to those keeping the model hierarchical.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations and terms so that other, more time consuming, methods are not feasible.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of *R*-Squared. If a switch can be found, it is made and the pool of terms is again searched to determine if another switch can be made. Note that this switching can be limited to those keeping the model hierarchical.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some *F*-to-enter and *F*-to-remove tests whose properties are not well understood to begin with.

Robust Regression

Regular multiple regression is optimum when all of its assumptions are valid. When some of these assumptions are invalid, least squares regression can perform poorly. Thorough residual analysis can point to these assumption breakdowns and allow you to work around these limitations. However, this residual analysis is time consuming and requires a great deal of training.

Robust regression provides an alternative to least squares regression that works with less restrictive assumptions. Specifically, it provides much better regression coefficient estimates when outliers are present in the data. Outliers violate the assumption of normally distributed residuals in least squares regression. They tend to pull the least squares fit too much in their direction by receiving much more "weight" than they deserve. Typically, you would expect that the weight attached to each observation would be about 1/*N* in a dataset with *N* observations. However, these outlying observations may receive a weight of 10, 20, or even 50 %. This leads to serious distortions in the estimated regression coefficients.

Because of this distortion, these outliers are difficult to identify since their residuals are much smaller than they should be. When only one or two independent variables are used, these outlying points may be visually detected in various scatter plots. However, the complexity added by additional independent variables hides the outliers from view in these scatter plots. Robust regression down-weights the influence of outliers. This makes their residuals larger and easier to spot. Robust regression techniques are iterative procedures that seek to identify these outliers and minimize their impact on the coefficient estimates.

The amount of weighting assigned to each observation in robust regression is controlled by a special curve called an *influence function*. There are three influence functions available in **NCSS**.

Although robust regression can particularly benefit untrained users, careful consideration should be given to the results. Essentially, robust regression conducts its own residual analysis and down-weights or completely removes various observations. You should study the weights that are assigned to each observation, determine which have been largely eliminated, and decide if you want these observations in your analysis.

M-Estimators

Several families of robust estimators have been developed. The robust methods found in ***NCSS*** fall into the family of *M-estimators*. This estimator minimizes the sum of a function $\rho(\cdot)$ of the residuals. That is, these estimators are defined as the $\beta's$ that minimize

$$\min_{\beta}\sum_{j=1}^{N}\rho\left(y_j - x_j'\beta\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(e_j\right)$$

M in *M*-estimators stands for maximum likelihood since the function $\rho(\cdot)$ is related to the likelihood function for a suitable choice of the distribution of the residuals. In fact, when the residuals follow the normal distribution, setting $\rho(u) = \frac{1}{2}u^2$ results in the usual method of least squares.

Unfortunately, *M*-estimators are not necessarily *scale invariant*. That is, these estimators may be influenced by the scale of the residuals. A scale-invariant estimator is found by solving

$$\min_{\beta}\sum_{j=1}^{N}\rho\left(\frac{y_j - x_j'\beta}{s}\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(\frac{e_j}{s}\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(u_j\right)$$

where *s* is a robust estimate of scale. The value of *s* is given by

$$s = \frac{median\left|e_j - median\left(e_j\right)\right|}{0.6745}$$

This estimate of *s* yields an approximately unbiased estimator of the standard deviation of the residuals when *N* is large and the error distribution is normal.

The function

$$\sum_{j=1}^{N}\rho\left(\frac{y_j - x_j'\beta}{s}\right)$$

is minimized by setting the first partial derivatives of $\rho(\cdot)$ with respect to each β_i to zero which forms a set of *p* + 1 nonlinear equations

$$\sum_{j=1}^{N}x_{ij}\psi\left(\frac{y_j - x_j'\beta}{s}\right) = 0, \quad i = 0,1,\cdots,p$$

where $\psi(u) = \rho'(u)$ is the *influence function*.

These equations are solved iteratively using an approximate technique called iteratively reweighted least squares (IRLS). At each step, new estimates of the regression coefficients are found using the matrix equation

$$\beta_{t+1} = \left(\mathbf{X}'\mathbf{W}_t\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}_t\mathbf{Y}$$

where W_t is an *N-by-N* diagonal matrix of weights $w_{1t}, w_{2t}, \cdots, w_{Nt}$ defined as

$$w_{jt} = \begin{cases} \dfrac{\psi\left[\left(y_j - x'\beta_{jt}\right)/s_t\right]}{\left(y_j - x'\beta_{jt}\right)/s_t} & \text{if } y_j \neq x'\beta_{jt} \\ 1 & \text{if } y_j = x'\beta_{jt} \end{cases}$$

The ordinary least squares regression coefficients are used at the first iteration to begin the iteration process. Iterations are continued until there is little or no change in the regression coefficients from one iteration to the next. Because of the masking nature of outliers, it is a good idea to run through at least five iterations to allow the outliers to be found.

Three functions are available in ***NCSS.*** These are Andrew's Sine, Huber's method, and Tukey's biweight. Huber's method is currently the most frequently recommended in the regression texts that we have seen. The specifics for each of these functions are as follows.

Andrew's Sine

$$\rho(u) = \begin{cases} c\left[1 - \cos(u/c)\right] & \text{if } |u| < \pi c \\ 2c & \text{if } |u| \geq \pi c \end{cases}$$

$$\psi(u) = \begin{cases} \sin(u/c) & \text{if } |u| < \pi c \\ 0 & \text{if } |u| \geq \pi c \end{cases}$$

$$w(u) = \begin{cases} \dfrac{\sin(u/c)}{u/c} & \text{if } |u| < \pi c \\ 0 & \text{if } |u| \geq \pi c \end{cases}$$

$$c = 1.339$$

Huber's Method

$$\rho(u) = \begin{cases} u^2 & \text{if } |u| < c \\ |2u|c - c^2 & \text{if } |u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u & \text{if } |u| < c \\ c\,\text{sign}(u) & \text{if } |u| \geq c \end{cases}$$

$$w(u) = \begin{cases} 1 & \text{if } |u| < c \\ c/|u| & \text{if } |u| \geq c \end{cases}$$

$$c = 1.345$$

Tukey's Biweight

$$\rho(u) = \begin{cases} \dfrac{c^2}{3}\left\{1 - \left[1 - \left(\dfrac{u}{c}\right)^2\right]^3\right\} & \text{if } |u| < c \\ 2c & \text{if } |u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u\left[1-\left(\frac{u}{c}\right)^2\right]^2 & \text{if } |u| < c \\ 0 & \text{if } |u| \geq c \end{cases}$$

$$w(u) = \begin{cases} \left[1-\left(\frac{u}{c}\right)^2\right]^2 & \text{if } |u| < c \\ 0 & \text{if } |u| \geq c \end{cases}$$

$$c = 4.685$$

This gives you a sketch of what robust regression is about. If you find yourself using the technique often, we suggest that you study one of the modern texts on regression analysis. All of these texts have chapters on robust regression. A good introductory discussion of robust regression is found in Hamilton (1991). A more thorough discussion is found in Montgomery and Peck (1992).

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown below. These data are from a study of the relationship of several variables with a person's I.Q. Fifteen people were studied. Each person's IQ was recorded along with scores on five different personality tests. The data are contained in the IQ database. We suggest that you open this database now so that you can follow along with the example.

IQ dataset

Test1	Test2	Test3	Test4	Test5	IQ
83	34	65	63	64	106
73	19	73	48	82	92
54	81	82	65	73	102
96	72	91	88	94	121
84	53	72	68	82	102
86	72	63	79	57	105
76	62	64	69	64	97
54	49	43	52	84	92
37	43	92	39	72	94
42	54	96	48	83	112
71	63	52	69	42	130
63	74	74	71	91	115
69	81	82	75	54	98
81	89	64	85	62	96
50	75	72	64	45	103

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present on a row for all but the dependent variable, a predicted value and confidence limits are generated for that row.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable(s)

This option specifies one or more dependent (*Y*) variables. If more than one variable is specified, a separate analysis is run for each.

Weight Variable

Weight Variable

When used, this is the name of a variable containing observation weights for generating a weighted-regression analysis. These weight values should be non-negative.

Numeric Independent Variables

X's: Numeric Independent Variable(s)

Specify any numeric independent variables in this box. Numeric variables are those whose values are numeric and are at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are more appropriately analyzed when you specify them as Categorical Independent Variables.

If you want to create powers and cross-products of these variables, specify an appropriate model in the 'Custom Model' field under the Model tab.

If you want to create predicted values of *Y* for values of *X* not in your database, add the *X* values to the bottom of the database. These rows will not be used during estimation phase, but predicted values will be generated for them on the reports.

Categorical Independent Variables

X's: Categorical Independent Variable(s)

Specify categorical (nominal) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories.

The values in a categorical variable are not used directly in the regression analysis. Instead, a set of numeric variables is substituted for them. Suppose a categorical variable has *G* categories. ***NCSS*** automatically generates the *G*-1 indicator variables that are needed for the analysis. The type of indicator variable created is determined by the selection for the *Default Reference Value* and the *Default Contrast Type*. The type of indicator created can also be controlled by entering the reference value and contrast type directly according to the syntax below. See the Default

Reference Value and Default Contrast Type sections below for a discussion of the reference value and contrast type options.

You can create the interactions among these variables automatically using the *Custom Model* field under the Model tab.

Syntax

The syntax for specifying a categorical variable is *VarName*(*RefValue*;*CType*) where *VarName* is the name of the variable, *RefValue* is the reference value, and *CType* is the type of numeric variables generated: B for binary, P for polynomial, R for contrast with the reference value, and S for a standard set of contrasts.

For example, suppose a categorical variable, STATE, has four values: Texas, California, Florida, and New York. To process this variable, the values are arranged in sorted order: California, Florida, New York, and Texas. Next, the reference value is selected. If a reference value is not specified, the default value specified in the *Default Reference Value* window is used. Finally, the method of generating numeric variables is selected. If such a method is not specified, the contrast type selected in the *Default Contrast Type* window is used. Possible ways of specifying this variable are

STATE	**RefValue = Default, CType = Default**
STATE(New York)	**RefValue = New York, CType = Default**
STATE(California;R)	**RefValue = California, CType = Contrast with Reference**
STATE(Texas;S)	**RefValue = Texas, CType = Standard Set**

More than one category variable may be designated using a list. Examples of specifying three variables with various options are shown next.

STATE BLOODTYPE GENDER

STATE(California;R) BLOODTYPE(O) GENDER(F)

STATE(Texas;S) BLOODTYPE(O;R) GENDER(F;B)

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting**

  Use the first value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**

  Use the last value in alpha-numeric sorted order as the reference value.

The reference value may also be designated within parentheses after the name of the categorical independent variable, in which case the default reference value is ignored. For example, suppose that the categorical independent variable, STATE, has four values: 1, 3, 4, and 5.

1. If this option is set to 'First Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 1.
2. If this option is set to 'Last Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 5.

3. If the categorical independent variable is entered as 'STATE(4)', the choice for this setting would be ignored, and the reference value would be 4.

Default Contrast Type

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to 'Contrast with Reference' or 'Standard Set' in order to match GLM results for factor effects.

- **Binary (This is the default)**

  Categories are converted to numbers using a set of binary indicator variables by assigning a '1' to the active category and a '0' to all other values. For example, suppose a categorical variable has G categories. ***NCSS*** automatically generates the G-1 binary (indicator) variables that are used in the regression. These indicator variables are set to 1 for those rows in which the value of this variable is equal to a certain value. They are set to 0 otherwise. The G-1 occurs because the Gth indicator variable is redundant (when all G-1 indicators are 0, wIfe know that the Gth indicator variable would be a 1). The value that is skipped is called the Reference Value.

  If your model includes interactions, using the binary indicator type may cause strange results.

  For the STATE variable, three binary variables would be generated. Suppose that the *Default Contrast Type* was 'Binary' and the statement used was 'STATE(Florida)'. The categories would be converted to numbers as follows:

STATE	B1	B2	B3
California	1	0	0
Florida	0	0	0
New York	0	1	0
Texas	0	0	1

- **Contrast with Reference**

  Categories are converted to numbers using a set of contrast variables by assigning a '1' to the active category, a '-1' to the reference value, and a '0' to all other values. A separate contrast is generated for each value other than the reference value.

  For the STATE variable, three numeric variables would be generated. Suppose the *Default Contrast Type* was 'Contrast with Reference', the *Default Reference Type* was 'Last Value after Sorting', and the variable was entered as 'STATE'. The categories would be converted to numbers as follows:

STATE	R1	R2	R3
California	1	0	0
Florida	0	1	0
New York	0	0	1
Texas	-1	-1	-1

- **Polynomial**

  If a variable has five or fewer categories, it can be converted to a set of polynomial contrast variables that account for the linear, quadratic, cubic, quartic, and quintic relationships. Note that these assignments are made after the values are sorted. Usually, the polynomial method is used on a variable for which the categories represent the actual values. That is, the values themselves are ordinal, not just category identifiers. Also, it is assumed that these values are equally spaced. Note that with this method, the reference value is ignored.

  For the STATE variable, linear, quadratic, and cubic variables are generated. Suppose that the *Default Contrast Type* was 'Polynomial' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	Linear	Quadratic	Cubic
California	-3	1	-1
Florida	-1	-1	3
New York	1	-1	-3
Texas	3	1	1

- **Standard Set**

  A variable can be converted to a set of contrast variables using a standard set of contrasts. This set is formed by comparing each value with those below it. Those above it are ignored. Note that these assignments are made after the values are sorted. The reference value is ignored.

  For the STATE variable, three numeric variables are generated. Suppose that the *Default Contrast Type* was 'Standard Set' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	S1	S2	S3
California	-3	0	0
Florida	1	-2	0
New York	1	1	-1
Texas	1	1	1

Regression Options

Perform Robust Regression

When checked, the program performs a robust regression analysis using the options specified under the ‘Robust’ tab.

Resampling

Calculate Bootstrap C.I.’s

Specify whether to calculate the bootstrap confidence intervals of the regression coefficients and predicted values. Note that this option uses Monte Carlo simulation and may require a long time to complete, especially for robust regression.

Alpha Levels

Alpha for C.I.'s and Tests

The value of alpha for the statistical tests and confidence intervals is specified here. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your particular study.

Alpha for Assumptions

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests use a larger alpha such as 0.10, 0.15, or 0.20.

We recommend 0.20.

Model Tab

These options control the regression model. To run a standard multiple regression analysis, set Subset Selection to *None* and Which Model Terms to *Up to 1-Way*.

Subset Selection

Subset Selection

This option specifies the subset selection algorithm used to reduce the number of independent variables used in the regression model. The Forward algorithm is much quicker than the Forward with Switching algorithm, but the Forward algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the individual generated variables. That is, either all numeric variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **None**

  No subset selection is attempted. All specified independent variables are used in the regression equation.

- **(Hierarchical) Forward**

  With this algorithm, the term that adds the most to *R*-Squared is entered into the model. Next, the term that increases the *R*-Squared the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the *R*-Squared does not change significantly.

- **(Hierarchical) Forward with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest *R*-Squared is entered into the regression model. The term which increases largest *R*-Squared the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, *R*-Squared is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in *R*-Squared. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Max Terms in Subset

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of *R*-Squared.

Note that the intercept is counted in this number.

Model Specification

Which Model Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a straight-forward regression model, select Up to 1-Way.

The options are

- **Full Model**

  The complete, saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables).

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

  Note that the discussion of the Custom Model option discusses the interpretation of this model.

- **Up to 1-Way**

  This option generates a model in which each variable is represented by a single model term. No cross-products or interaction terms are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

  This is the option to select when you want to analyze the independent variables specified without adding any other terms.

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C

- **Up to 2-Way**

  This option specifies that all main effects and two-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C

- **Up to 3-Way**

  All main effects, two-way interactions, and three-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

- **Up to 4-Way**

  All main effects, two-way interactions, three-way interactions, and four-way interactions are included in the model. For example, if you have four independent variables A, B, C, and D, this would generate the model:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D

- **Custom Model**

  The model specified in the *Custom Model* box is used.

Remove Intercept

Unchecked indicates that the intercept term, β_0, is to be included in the regression. Checked indicates that the intercept should be omitted from the regression model. Note that deleting the intercept distorts most of the diagnostic statistics (R-Squared, etc.). In most situations, you should include the intercept in the model.

Write Model in Custom Model Field

When this option is checked, no data analysis is performed when the procedure is run. Instead, a copy of the full model is stored in the Custom Model box. You can then edit the model as desired. This option is useful when you have several variables and you want to be selective about which terms are used.

Note that the program will not do any calculations while this option is checked.

Model Specification – Custom Model

Max Term Order

This option specifies that maximum number of variables that can occur in an interaction term in a custom model. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would be excluded from the model.

This option is particularly useful when used with the bar notation of a custom model to allow a simple way to remove unwanted high-order interactions.

Custom Model

This options specifies a custom model. It is only used when the Which Model Terms option is set to *Custom Model*. A custom model specifies the terms (single variables and interactions) that are to be kept in the model.

Interactions

An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of generated variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the generated variables.

Syntax

A model is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A + B + C + A*B + A*C + B*C + A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B + A*C.

Some examples will help to indicate how the model syntax works:

A|B = A + B + A*B

A|B A*A B*B = A + B + A*B + A*A + B*B

Note that you should only repeat numeric variables. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A + B + A*A + A*B + B*B

A|B|C = A + B + C + A*B + A*C + B*C + A*B*C

(A + B)*(C + D) = A*C + A*D + B*C + B*D

(A + B)|C = (A + B) + C + (A + B)*C = A + B + C + A*C + B*C

Reports Tab

The following options control which reports and plots are displayed. Since over 30 reports are available, you may want to spend some time deciding which reports to display on a routine basis and create a template that saves your favorite choices.

Select Report / Plot Group

Select a Group of Reports and Plots

This option allows you to specify a group of reports and plots without checking them individually. The checking of individual reports and plots is only useful when this option is set to *Display only those items that are CHECKED BELOW*. Otherwise, the checking of individual reports and plots is ignored.

Report Options

Show All Rows

This option makes it possible to display predicted values for only a few designated rows.

When checked predicted values, residuals, and other row-by-row statistics, will be displayed for all rows used in the analysis.

When not checked, predicted values and other row-by-row statistics will be displayed for only those rows in which the dependent variable's value is missing.

Select Reports – Summaries

Run Summary ... Correlations

Each of these options specifies whether the indicated report is calculated and displayed.

Select Reports – Subset Selection

Subset Summary and Subset Detail

Indicate whether to display these subset selection reports.

Select Reports – Estimation

Equation ... Robust Percentiles

Indicate whether to display these estimation reports.

Select Reports – ANOVA

ANOVA Summary and ANOVA Detail

Indicate whether to display these ANOVA reports.

Select Reports – Assumptions

PRESS Statistics ... Durbin-Watson

Indicate whether to display these assumptions reports.

Select Reports – IV Diagnostics

R-Squared ... Multicollinearity

Indicate whether to display these independent variable diagnostic reports.

Select Reports – Study Design

Eigenvalues - Centered Corr ... Eigenvectors - Uncentered Corr

Indicate whether to display these design reports.

Select Reports – Row-by-Row Lists

Robust Residuals ... DFBetas

Indicate whether to display these list reports. The number of rows in the output is controlled by the 'Show All Rows' option. Note that since these reports provide results for each row, they may be too long for normal use when requested on large databases.

Select Plots

Histogram ... Partial Resid vs X Plot

Indicate whether to display these plots.

Format Tab

These options specify the number of decimal places shown when the indicated value is displayed in a report. The number of decimal places shown in plots is controlled by the Tick Label Settings buttons on the Axes tabs.

Report Options

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Skip Line After

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable’s results printed on two lines.

Enter *100* when you want each variable’s results printed on a single line.

Report Options – Decimal Places

Probability ... Mean Square Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'All' to display all digits available. The number of digits displayed by this option is controlled by whether the PRECISION option is SINGLE or DOUBLE.

Plot Options Tab

These options control the titles and style files used on each of the plots.

Plot Titles and Style Files

Plot Titles

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Plot Style Files

Designate various plot style files. These files set all plot options that are not set directly by this procedure. Unless you choose otherwise, the default style file (Default) is used. These files are created in the various graphics procedures, depending on the plot type.

Plotting Symbol

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Histogram Options

Number of Bars

Specify the number of intervals, bins, or bars used in the histogram. Select ‘0 - Automatic’ to have the program select an appropriate number based on the number of residuals.

Axes Tabs

The options on these panels control the appearance of the *X* variables, *Y* variable, residuals, RStudent, Hat Diagonal, Rows Numbers, Counts, and Expected axes whenever they are included on a plot. This makes it easy to give a consistent look to all of your plots without modifying them individually.

Y-Variable ... Expected Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by the names of the corresponding variables. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the axis associated with this variable. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the associated axis.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on the associated axes.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Robust Tab

The options on this panel control the robust regression analysis, if designated.

Robust Regression Options

Robust Method

This option specifies which of the three robust influence functions is used: Andrews' sine, Tukey's biweight, or Huber's. We recommend that use Huber's method.

Robust Regression Options – Robust Truncation Constants

Andrew's Sine Constant

This option specifies the robust truncation constant for Andrew's Sine method. This is a cutoff point on the influence function designating when an observation's weight should be set to zero.

The recommended value is 1.339.

Huber's Constant

This option specifies the robust truncation constant for Huber's method. This is a cutoff point on the influence function designating when an observation's weight should be reduced.

The recommended value is 1.345.

Tukey’s Biweight Constant

This option specifies the robust truncation constant for Tukey’s Biweight method. This is a cutoff point on the influence function designating when an observation’s weight should be set to zero.

The recommended value is 4.685.

Robust Regression Options – Robust Truncation Constants

Minimum % Beta Change

This option specifies an early stopping value for the iteration procedure. Normally, the number of iterations is specified in the next option. However, if the percentage change in each of the estimated regression coefficients is less than this amount, the iteration procedure is terminated. If you want this option to be ignored, set it to zero.

Maximum Iterations

This is the maximum number of iterations that the robust procedure will cycle through. Normally, you should have at least five iterations.

MAD Constant

Specify the constant used to scale MAD. The default value of 0.6745 has been suggested in several regression texts.

Note that in the older Robust Regression procedure, this value was set to 1.0.

Robust Regression Options – Reporting Options

Cutoff for Weight Report

On the Robust Residuals and Weights report, only rows with weights less than this amount are displayed. Since this report may be several pages long when the number of rows is large, this cutoff value allows you to see only those rows that have been severely down-weighted.

The permissible range is from 0.01 to 1.00. The recommended value is 0.20.

Resampling Tab

This panel controls the bootstrapping. Note that bootstrapping is only used when the Calculate Bootstrap C.I.’s is checked on the Variables panel.

Bootstrap Calculation Options – Sampling

Samples (N)

This is the number of bootstrap samples used. A general rule of thumb is that you use at least 100 when standard errors are your focus or at least 1000 when confidence intervals are your focus. If computing time is available, it does not hurt to do 4000 or 5000.

We recommend setting this value to at least 3000.

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. We recommend setting the parameter to at least 50.

Bootstrap Calculation Options – Estimation

Percentile Type

The method used to create the percentiles when forming bootstrap confidence limits. You can read more about the various types of percentiles in the Descriptive Statistics chapter. We suggest you use the Ave *X*(p[n+1]) option.

C.I. Method

This option specifies the method used to calculate the bootstrap confidence intervals. The reflection method is recommended.

- **Percentile**

  The confidence limits are the corresponding percentiles of the bootstrap values.

- **Reflection**

  The confidence limits are formed by reflecting the percentile limits. If *X*0 is the original value of the parameter estimate and *XL* and *XU* are the percentile confidence limits, the Reflection interval is (2 *X*0 - *XU*, 2 *X*0 - *XL*).

Bootstrap Confidence Coefficients

These are the confidence coefficients of the bootstrap confidence intervals. Since bootstrapping calculations may take several minutes, it may be useful to obtain confidence intervals using several different confidence coefficients.

All values must be between 0.50 and 1.00. You may enter several values, separated by blanks or commas. A separate confidence interval is given for each value entered.

Examples:

0.90 0.95 0.99

0.90:.99(0.01)

0.90.

Bootstrap Histogram Options

Vertical Axis Label

This is the label of the vertical axis of a bootstrap histogram.

Horizontal Axis Label

This is the label of the horizontal axis of a bootstrap histogram.

Plot Style File

This is the histogram style file. We have provided several different style files to choose from, or you can create your own in the Histogram procedure.

Histogram Title

This is the title used on the bootstrap histograms.

Number of Bars

The number of bars shown in a bootstrap histogram. We recommend setting this value to at least 25 when the number of bootstrap samples is over 1000.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful.

Data Storage Options – Select Items to Store

Predicted Y ... VC(Betas) Matrix

Indicate whether to store these row-by-row values, beginning at the variable indicated by the *Store First Variable In* option.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Multiple Regression (All Reports)

This section presents an example of how to run a multiple regression analysis of the data presented earlier in this chapter. The data are in the IQ database. This example will run a regression of *IQ* on *Test1* through *Test5*. This regression program outputs over thirty different reports and plots, many of which contain duplicate information. For the purposes of annotating the output, all output is displayed. Normally, you would only select a few these reports.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Multiple Regression window.

1 **Open the IQ dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **IQ.s0**.
- Click **Open**.

2 **Open the Multiple Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **IQ**.
- Set the **X's: Numeric Independent Variables** box to **Test1-Test5**.

4 Specify the reports.

- Select the **Reports tab**.
- Set the **Select a Group of Reports and Plots** to **Display ALL reports & plots**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Section

```
Parameter                 Value          Parameter                  Value
Dependent Variable        IQ             Rows Processed             17
Number Ind. Variables     5              Rows Filtered Out          0
Weight Variable           None           Rows with X's Missing      0
R2                        0.3991         Rows with Weight Missing   0
Adj R2                    0.0652         Rows with Y Missing        0
Coefficient of Variation  0.1021         Rows Used in Estimation    15
Mean Square Error         113.4648       Sum of Weights             15.000
Square Root of MSE        10.65198       Completion Status          Normal Completion
Ave Abs Pct Error         6.218
```

This report summarizes the multiple regression results. It presents the variables used, the number of rows used, and the basic results.

R-Squared

R^2, officially known as the coefficient of determination, is defined as

$$R^2 = \frac{SS_{Model}}{SS_{Total(Adjusted)}}$$

R^2 is probably the most popular statistical measure of how well the regression model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship between the Y and the X's, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Additional independent variables*. It is possible to increase R^2 by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This case happens when your sample size is small.

2. *Range of the independent variables*. R^2 is influenced by the range of each independent variable. R^2 increases as the range of the X's increases and decreases as the range of the X's decreases.

3. *Slope magnitudes*. R^2 does not measure the magnitude of the slopes.

4. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between x and Y was a perfect circle. The R^2 value of this relationship would be zero.

5. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.

6. *No-intercept model*. The definition of R^2 assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of R^2 changes dramatically. The fact that your R^2 value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying meaning of R^2.

7. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Adjusted R-Squared

This is an adjusted version of R^2. The adjustment seeks to remove the distortion due to a small sample size.

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, computed by dividing root mean square error by the mean of the dependent variable. By itself, it has little value, but it can be useful in comparative studies.

$$CV = \frac{\sqrt{MSE}}{\bar{y}}$$

Ave Abs Pct Error

This is the average of the absolute percent errors. It is another measure of the goodness of fit of the regression model to the data. It is calculated using the formula

$$AAPE = \frac{100\sum_{j=1}^{N}\left|\frac{y_j - \hat{y}_j}{y_j}\right|}{N}$$

Note that when the dependent variable is zero, its predicted value is used in the denominator.

Descriptive Statistics Section

```
                                      Standard
Variable      Count    Mean           Deviation      Minimum    Maximum
Test1         15       67.93333       17.39239       37         96
Test2         15       61.4           19.39735       19         89
Test3         15       72.33334       14.73415       43         96
Test4         15       65.53333       13.95332       39         88
Test5         15       69.93333       16.15314       42         94
IQ            15       104.3333       11.0173        92         130
```

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

```
             Test1      Test2      Test3      Test4
Test1       1.0000     0.1000    -0.2608     0.7539
Test2       0.1000     1.0000     0.0572     0.7196
Test3      -0.2608     0.0572     1.0000    -0.1409
Test4       0.7539     0.7196    -0.1409     1.0000
Test5       0.0140    -0.2814     0.3473    -0.1729
IQ          0.2256     0.2407     0.0741     0.3714

             Test5         IQ
Test1       0.0140     0.2256
Test2      -0.2814     0.2407
Test3       0.3473     0.0741
Test4      -0.1729     0.3714
Test5       1.0000    -0.0581
IQ         -0.0581     1.0000
```

Pearson correlations are given for all variables. Outliers, nonnormality, nonconstant variance, and nonlinearities can all impact these correlations. Note that these correlations may differ from pair-wise correlations generated by the correlation matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with one another may cause collinearity problems.

Regression Equation Section

```
            Regression   Standard     T-Value               Reject     Power
Independent Coefficient  Error        to test     Prob      H0 at      of Test
Variable    b(i)         Sb(i)        H0:B(i)=0   Level     5%?        at 5%
Intercept   85.2404      23.6951      3.597       0.0058    Yes        0.8915
Test1       -1.9336      1.0291       -1.879      0.0930    No         0.3896
Test2       -1.6599      0.8729       -1.902      0.0897    No         0.3974
Test3       0.1050       0.2199       0.477       0.6445    No         0.0713
Test4       3.7784       1.8345       2.060       0.0695    No         0.4522
Test5       -0.0406      0.2012       -0.202      0.8447    No         0.0538

Estimated Model
 85.2403846967438-1.9335712381893*Test1-1.6598811696115*Test2+ .104954325385776*Test3
+3.7783766794138*Test4-4.05775409260278E-02*Test5
```

This section reports the values and significance tests of the regression coefficients. Before using this report, check that the assumptions are reasonable. For instance, collinearity can cause the t-tests to give false results and the regression coefficients to be of the wrong magnitude or sign.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

Regression Coefficient

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in that particular X when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests or confidence limits.

T-Value to test Ho: B(i)=0

This is the t-test value for testing the hypothesis that $\beta_j = 0$ versus the alternative that $\beta_j \neq 0$ after removing the influence of all other *X's*. This *t*-value has *n*-*p*-1 degrees of freedom.

To test for a value other than zero, use the formula below. There is an easier way to test hypothesized values using confidence limits. See the discussion below under Confidence Limits. The formula for the *t*-test is

$$t_j = \frac{b_j - \beta_j^*}{s_{b_j}}$$

Prob Level

This is the *p*-value for the significance test of the regression coefficient. The *p*-value is the probability that this *t*-statistic will take on a value at least as extreme as the actually observed value, assuming that the null hypothesis is true (i.e., the regression estimate is equal to zero). If the *p*-value is less than alpha, say 0.05, the null hypothesis of equality is rejected. This *p*-value is for a two-tail test.

Reject H0 at 5%?

This is the conclusion reached about the null hypothesis. It will be either reject *H0* at the 5% level of significance or not.

Note that the level of significance is specified in the Alpha of C.I.'s and Tests box on the Format tab panel.

Power (5%)

Power is the probability of rejecting the null hypothesis that $\beta_j = 0$ when $\beta_j = \beta_j^* \neq 0$. The power is calculated for the case when $\beta_j^* = b_j$, $\sigma^2 = s^2$, and alpha is as specified in the Alpha of C.I.'s and Tests option.

High power is desirable. High power means that there is a high probability of rejecting the null hypothesis that the regression coefficient is zero when this is false. This is a critical measure of sensitivity in hypothesis testing. This estimate of power is based upon the assumption that the residuals are normally distributed.

Estimated Model

This is the least squares regression line presented in double precision. Besides showing the regression model in long form, it may be used as a transformation by copying and pasting it into the Transformation portion of the spreadsheet.

Note that a transformation must be less than 255 characters. Since these formulas are often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file, including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

Regression Coefficient Section

```
Independent  Regression        Standard       Lower       Upper   Standardized
Variable     Coefficient          Error    95% C.L.    95% C.L.    Coefficient
Intercept        85.2404        23.6951     31.6383    138.8425         0.0000
Test1            -1.9336         1.0291     -4.2615      0.3944        -3.0524
Test2            -1.6599         0.8729     -3.6345      0.3147        -2.9224
Test3             0.1050         0.2199     -0.3925      0.6024         0.1404
Test4             3.7784         1.8345     -0.3715      7.9283         4.7853
Test5            -0.0406         0.2012     -0.4958      0.4146        -0.0595
Note: The T-Value used to calculate these confidence limits was 2.262.
```

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

Regression Coefficient

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in *x* when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

Lower - Upper 95% C.L.

These are the lower and upper values of a $100(1-\alpha)\%$ interval estimate for β_j based on a *t*-distribution with *n*-*p*-1 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed.

These confidence limits may be used for significance testing values of β_j other than zero. If a specific value is not within this interval, it is significantly different from that value. Note that these confidence limits are set up as if you are interested in each regression coefficient separately.

The formulas for the lower and upper confidence limits are:

$$b_j \pm t_{1-\alpha/2,n-p-1} \, s_{b_j}$$

Standardized Coefficient

Standardized regression coefficients are the coefficients that would be obtained if you standardized the independent variables and the dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When the independent variables have vastly different scales of measurement, this value provides a way of making comparisons among variables. The formula for the standardized regression coefficient is:

$$b_{j,\,std} = b_j \left(\frac{s_{X_j}}{s_Y} \right)$$

where s_Y and s_{X_j} are the standard deviations for the dependent variable and the j^{th} independent variable.

Note: The T-Value …

This is the value of $t_{1-\alpha/2,n-p-1}$ used to construct the confidence limits.

Analysis of Variance Section

```
                                     Sum of        Mean                Prob     Power
Source            DF       R2       Squares      Square    F-Ratio     Level     (5%)
Intercept          1              163281.7    163281.7
Model              5   0.3991     678.1504    135.6301      1.195    0.3835   0.2565
Error              9   0.6009     1021.183    113.4648
Total(Adjusted)   14   1.0000     1699.333     121.381
```

An analysis of variance (ANOVA) table summarizes the information related to the variation in data.

Source

This represents a partition of the variation in *Y*.

R2

This is the overall R^2 of this the regression model.

DF

The degrees of freedom are the number of dimensions associated with this term. Note that each observation can be interpreted as a dimension in *n*-dimensional space. The degrees of freedom for the intercept, model, error, and adjusted total are 1, *p*, *n*-*p*-1, and *n*-1, respectively.

Sum of Squares

These are the sums of squares associated with the corresponding sources of variation. Note that these values are in terms of the dependent variable. The formulas for each are

$$SS_{Intercept} = n\bar{y}^2$$

$$SS_{Model} = \sum \left(\hat{y}_j - \bar{y} \right)^2$$

$$SS_{Error} = \Sigma\left(y_j - \hat{y}_j\right)^2$$

$$SS_{Total} = \Sigma\left(y_j - \bar{y}\right)^2$$

Mean Squares

The mean square is the sum of squares divided by the degrees of freedom. This mean square is an estimated variance. For example, the mean square error is the estimated variance of the residuals.

F-Ratio

This is the *F*-statistic for testing the null hypothesis that all $\beta_j = 0$. This *F*-statistic has *p* degrees of freedom for the numerator variance and *n*-*p*-1 degrees of freedom for the denominator variance.

Prob Level

This is the *p*-value for the above *F-test*. The *p*-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the *p*-value is less than α, say 0.05, the null hypothesis is rejected. If the *p*-value is greater than α, then the null hypothesis is accepted.

Power(5%)

Power is the probability of rejecting the null hypothesis that all the regression coefficients are zero when at least one is not.

Analysis of Variance Detail Section

Model Term	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		163281.7	163281.7			
Model	5	0.3991	678.1504	135.6301	1.195	0.3835	0.2565
Test1	1	0.2357	400.562	400.562	3.530	0.0930	0.3896
Test2	1	0.2414	410.2892	410.2892	3.616	0.0897	0.3974
Test3	1	0.0152	25.8466	25.8466	0.228	0.6445	0.0713
Test4	1	0.2832	481.3241	481.3241	4.242	0.0695	0.4522
Test5	1	0.0027	4.614109	4.614109	0.041	0.8447	0.0538
Error	9	0.6009	1021.183	113.4648			
Total(Adjusted)	14	1.0000	1699.333	121.381			

This analysis of variance table provides a line for each term in the model. It is especially useful when you have categorical independent variables.

Model Term

This is the term from the design model.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the number of degrees of freedom that the model is degrees of freedom is reduced when this term is removed from the model. This is the numerator degrees of freedom of the *F-test*.

R2

This is the amount that R^2 is reduced when this term is removed from the regression model.

Sum of Squares

This is the amount that the model sum of squares that are reduced when this term is removed from the model.

Mean Square

The mean square is the sum of squares divided by the degrees of freedom.

F-Ratio

This is the *F*-statistic for testing the null hypothesis that all β_j associated with this term are zero. This *F-statistic* has *DF* and *n*-*p*-1 degrees of freedom.

Prob Level

This is the *p*-value for the above *F*-test. The *p*-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the *p*-value is less than α , say 0.05, the null hypothesis is rejected. If the *p*-value is greater than α , then the null hypothesis is accepted.

Power(5%)

Power is the probability of rejecting the null hypothesis that all the regression coefficients associated with this term are zero, assuming that the estimated values of these coefficients are their true values.

PRESS Section

```
                                      From          From
                                     PRESS       Regular
Parameter                        Residuals     Residuals
Sum of Squared Residuals          2839.941      1021.183
Sum of |Residuals|                169.6438      99.12155
R2                                 -0.6712        0.3991
```

This section reports on the PRESS statistics. The regular statistics, computed on all of the data, are provided to the side to make comparison between corresponding values easier.

Sum of Squared Residuals

PRESS is an acronym for prediction sum of squares. It was developed for use in variable selection to validate a regression model. To calculate *PRESS*, each observation is individually omitted. The remaining *N* - 1 observations are used to calculate a regression and estimate the value of the omitted observation. This is done *N* times, once for each observation. The difference between the actual *Y* value and the predicted *Y* with the observation deleted is called the prediction error or *PRESS* residual. The sum of the squared prediction errors is the *PRESS* value. The smaller *PRESS* is, the better the predictability of the model.

$$\Sigma \left(y_j - \hat{y}_{j,\text{-}j} \right)^2$$

Sum of |Press residuals|

This is the sum of the absolute value of the *PRESS* residuals or prediction errors. If a large value for the *PRESS* is due to one or a few large *PRESS* residuals, this statistic may be a more accurate way to evaluate predictability.

$$\Sigma \left| y_j - \hat{y}_{j,-j} \right|$$

Press R2

The PRESS value above can be used to compute an R^2-like statistic, called *R2Predict*, which reflects the prediction ability of the model. This is a good way to validate the prediction of a regression model without selecting another sample or splitting your data. It is very possible to have a high R^2 and a very low *R2Predict*. When this occurs, it implies that the fitted model is data dependent. This *R2Predict* ranges from below zero to above one. When outside the range of zero to one, it is truncated to stay within this range.

$$R^2_{PRESS} = 1 - \frac{PRESS}{SS_{Total}}$$

Normality Tests Section

```
Test                            Test        Prob       Reject H0
Name                           Value       Level   At Alpha = 20%?
Shapiro Wilk                  0.9076    0.124280               Yes
Anderson Darling              0.4365    0.297324                No
D'Agostino Skewness           2.0329    0.042064               Yes
D'Agostino Kurtosis           1.5798    0.114144               Yes
D'Agostino Omnibus            6.6285    0.036361               Yes
```

This report gives the results of applying several normality tests to the residuals. The Shapiro-Wilk test is probably the most popular, so it is given first. These tests are discussed in detail in the Normality Test section of the Descriptive Statistics procedure.

Serial-Correlation and Durbin-Watson Test

```
          Serial                Serial                Serial
Lag  Correlation      Lag  Correlation      Lag  Correlation
1         0.4529        9      -0.2769       17       0.0000
2        -0.2507       10      -0.2287       18       0.0000
3        -0.5518       11       0.0000       19       0.0000
4        -0.3999       12       0.0000       20       0.0000
5         0.0780       13       0.0000       21       0.0000
6         0.2956       14       0.0000       22       0.0000
7         0.1985       15       0.0000       23       0.0000
8        -0.0016       16       0.0000       24       0.0000
Above serial correlations significant if their absolute values are greater than 0.516398

Durbin-Watson Test For Serial Correlation
                                                 Did the Test Reject
Parameter                              Value      H0: Rho(1) = 0?
Durbin-Watson Value                   1.0010
Prob. Level: Positive Serial Correlation   0.0072                Yes
Prob. Level: Negative Serial Correlation   0.9549                 No
```

This section reports the autocorrelation structure of the residuals. Of course, this report is only useful if the data represent a time series.

Lag and Correlation

The lag, *k*, is the number of periods (rows) back. The correlation here is the sample autocorrelation coefficient of lag *k*. It is computed as:

$$r_k = \frac{\sum e_{i-k} e_i}{\sum e_i^2} \qquad \textit{for } k = 1,2,\ldots,24$$

To test the null hypothesis that $\rho_k = 0$ at a 5% level of significance with a large-sample normal approximation, reject when the absolute value of the autocorrelation coefficient, $|r_k|$, is greater than two over the square root of *N*.

Durbin-Watson Value

The Durbin-Watson test is often used to test for positive or negative, first-order, serial correlation. It is calculated as follows

$$r_k = \frac{\sum e_{i-k} e_i}{\sum e_i^2} \qquad \textit{for } k = 1,2,\ldots,24$$

The distribution of this test is mathematically difficult because it involves the *X* values. Originally, Durbin-Watson (1950, 1951) gave a pair of bounds to be used. However, there is a large range of indecision that can be found when using these bounds. Instead of using these bounds, ***NCSS*** calculates the exact probability using the beta distribution approximation suggested by Durbin-Watson (1951). This approximation has been shown to be accurate to three decimal places in most cases.

R-Squared Section

```
                                R2 Increase   R2 Decrease     R2 When    Partial R2
                  Total R2 for    When This     When This    This I.V.     Adjusted
Independent      This I.V. And  I.V. Added To     I.V. Is       Is Fit      For All
Variable          Those Above   Those Above      Removed        Alone  Other I.V.'s
Test1                  0.0509        0.0509       0.2357       0.0509        0.2817
Test2                  0.0990        0.0480       0.2414       0.0579        0.2866
Test3                  0.1131        0.0142       0.0152       0.0055        0.0247
Test4                  0.3964        0.2832       0.2832       0.1379        0.3203
Test5                  0.3991        0.0027       0.0027       0.0034        0.0045
```

R^2 reflects the percent of variation in *Y* explained by the independent variables in the model. A value of R^2 near zero indicates a complete lack of fit between *Y* and the *X*s, while a value near one indicates a perfect fit.

In this section, various types of R^2 values are given to provide insight into the variation in the dependent variable explained either by the independent variables added in order (i.e., sequential) or by the independent variables added last. This information is valuable in an analysis of which variables are most important.

Independent Variable

This is the name of the independent variable reported on in this row.

Total R2 for This I.V. and Those Above

This is the R^2 value that would result from fitting a regression with this independent variable and those listed above it. The IV’s below it are ignored.

R2 Increase When This IV Added to Those Above

This is the amount that this IV adds to R^2 when it is added to a regression model that includes those IV’s listed above it in the report.

R2 Decrease When This IV is Removed

This is the amount that R^2 would be reduced if this IV were removed from the model. Large values here indicate important independent variables, while small values indicate insignificant variables.

One of the main problems in interpreting these values is that each assumes all other variables are already in the equation. This means that if two variables both represent the same underlying information, they will each seem to be insignificant after considering the other. If you remove both, you will lose the information that either one could have brought to the model.

R2 When This IV Is Fit Alone

This is the R^2 that would be obtained if the dependent variable were only regressed against this one independent variable. Of course, a large R^2 value here indicates an important independent variable that can stand alone.

Partial R2 Adjusted For All Other IV’s

The is the square of the partial correlation coefficient. The partial R^2 reflects the percent of variation in the dependent variable explained by one independent variable controlling for the effects of the rest of the independent variables. Large values for this partial R^2 indicate important independent variables.

Variable Omission Section

Independent Variable	R2 When I.V. Omitted	MSE When I.V. Omitted	Mallow's Cp When I.V. Omitted	H0: B=0 Prob Level	Regress. Of This I.V. On Other I.V.'s
Full Model	0.3991	113.4648			
Test1	0.1634	142.1745	7.5303	0.0930	0.9747
Test2	0.1576	143.1472	7.6160	0.0897	0.9717
Test3	0.3839	104.703	4.2278	0.6445	0.2280
Test4	0.1158	150.2507	8.2421	0.0695	0.9876
Test5	0.3964	102.5797	4.0407	0.8447	0.2329

One way of assessing the importance of an independent variable is to examine the impact on various goodness-of-fit statistics of removing it from the model. This section provides this.

Independent Variable

This is the name of the predictor variable reported on in this row. Note that the *Full Model* row gives the statistics when no variables are omitted.

R2 When IV Omitted

This is the R^2 for the multiple regression model when this independent variable is omitted and the remaining independent variables are retained. If this R^2 is close to the R^2 for the full model, this variable is not very important. On the other hand, if this R^2 is much smaller than that of the full model, this independent variable is important.

MSE When IV Omitted

This is the mean square error for the multiple regression model when this IV is omitted and the remaining IV’s are retained. If this MSE is close to the MSE for the full model, this variable may not be very important. On the other hand, if this MSE is much larger than that of the full model, this IV is important.

Mallow's Cp When IV Omitted

Another criterion for variable selection and importance is Mallow’s *Cp* statistic. The optimum model will have a *Cp* value close to *p*+1, where *p* is the number of independent variables. A *Cp* greater than (*p*+1) indicates that the regression model is overspecified (contains too many variables and stands a chance of having collinearity problems). On the other hand, a model with a *Cp* less than (*p*+1) indicates that the regression model is underspecified (at least one important independent variable has been omitted). The formula for the *Cp* statistic is as follows, where *k* is the maximum number of independent variables available

$$C_p = (n-p-1)\left[\frac{MSE_p}{MSE_k}\right] - \left[n-2(p+1)\right]$$

H0: B=0 Prob Level

This is the two-tail *p*-value for testing the significance of the regression coefficient. Most likely, you would deem IV’s with small *p*-values as important. However, you must be careful here. Collinearity can cause extra large *p*-values, so you must check for its presence.

R2 Of Regress. Of This IV Other X’s

This is the R^2 value that would result if this independent variable were regressed on the remaining independent variables. A high value indicates a redundancy between this IV and the other IV’s. IV’s with a high value here (above 0.90) are candidates for omission from the model.

Sum of Squares and Correlation Section

```
               Sequential   Incremental        Last
Independent        Sum of        Sum of      Sum of      Simple        Partial
Variable          Squares       Squares     Squares  Correlation   Correlation
Test1             86.5252       86.5252     400.562      0.2256       -0.5308
Test2            168.1614       81.6362    410.2892      0.2407       -0.5354
Test3            192.2748      24.11342     25.8466      0.0741        0.1571
Test4            673.5363      481.2615    481.3241      0.3714        0.5660
Test5            678.1504      4.614109    4.614109     -0.0581       -0.0671
```

This section provides the sum of squares and correlations equivalent to the *R-Squared Section*.

Independent Variable

This is the name of the IV reported on in this row.

Sequential Sum Squares

The is the sum of squares value that would result from fitting a regression with this independent variable and those above it. The IV’s below it are ignored.

Incremental Sum Squares

This is the amount that this predictor adds to the sum of squares value when it is added to a regression model that includes those predictors listed above it.

Last Sum Squares

This is the amount that the model sum of squares would be reduced if this variable were removed from the model.

Simple Correlation

This is the Pearson correlation coefficient between the dependent variable and the specified independent variable.

Partial Correlation

The partial correlation coefficient is a measure of the strength of the linear relationship between Y and X_j after adjusting for the remaining (p-1) variables.

Sequential Models Section

```
Independent     Included     Omitted    Included    Included     Omitted     Omitted
Variable              R2          R2     F-Ratio      Prob>F     F-Ratio      Prob>F
Test1             0.0509      0.3482       0.697      0.4187       1.304      0.3390
Test2             0.0990      0.3001       0.659      0.5351       1.498      0.2801
Test3             0.1131      0.2859       0.468      0.7107       2.141      0.1735
Test4             0.3964      0.0027       1.641      0.2390       0.041      0.8447
Test5             0.3991      0.0000       1.195      0.3835
Notes
1. INCLUDED variables are those listed from current row up (includes current row).
2. OMITTED variables are those listed below (but not including) this row.
```

This section examines the step-by-step effect of adding variables to the regression model.

Independent Variable

This is the name of the predictor variable reported on in this row.

Included R2

This is the R^2 that would be obtained if only those IV’s on this line and above were in the regression model.

Omitted R2

This is the R^2 for the full model minus the *Included R2*. This is the amount of R^2 explained by the independent variables listed below the current row. Large values indicate that there is much more to come with later independent variables. On the other hand, small values indicate that remaining independent variables contribute little to the regression model.

Included F-ratio

This is an F-ratio for testing the hypothesis that the regression coefficients ($\beta's$) for the IV’s listed on this row and above are zero.

Included Prob>F

This is the p-value for the above F-ratio.

Omitted F-Ratio

This is an F-ratio for testing the hypothesis that the regression coefficients ($\beta's$) for the variables listed below this row are all zero. The alternative is that at least one coefficient is nonzero.

Omitted Prob>F

This is the *p*-value for the above *F*-ratio.

Multicollinearity Section

```
                     Variance           R2                          Diagonal
Independent          Inflation      Versus                            of X'X
Variable             Factor     Other I.V.'s    Tolerance             Inverse
Test1                39.5273        0.9747        0.0253        9.333631E-03
Test2                35.3734        0.9717        0.0283        6.715277E-03
Test3                 1.2953        0.2280        0.7720        4.261841E-04
Test4                80.8456        0.9876        0.0124        2.966012E-02
Test5                 1.3035        0.2329        0.7671        3.568483E-04
```

This report provides information useful in assessing the amount of multicollinearity in your data.

Variance Inflation

The variance inflation factor (*VIF*) is a measure of multicollinearity. It is the reciprocal of $1- R_X^2$, where R_X^2 is the R^2 obtained when this variable is regressed on the remaining IV's. A *VIF* of 10 or more for large data sets indicates a collinearity problem since the R_X^2 with the remaining *IV*'s is 90 percent. For small data sets, even *VIF's* of 5 or more can signify collinearity. Variables with a high *VIF* are candidates for exclusion from the model.

$$VIF_j = \frac{1}{1-R_j^2}$$

R2 Versus Other IV's

R_X^2 is the R^2 obtained when this variable is regressed on the remaining independent variables. A high R_X^2 indicates a lot of overlap in explaining the variation among the remaining independent variables.

Tolerance

Tolerance is just $1- R_X^2$, the denominator of the variance inflation factor.

Diagonal of X'X Inverse

The *X'X* inverse is an important matrix in regression. This is the j^{th} row and j^{th} column element of this matrix.

Eigenvalues of Centered Correlations Section

```
                         Incremental      Cumulative      Condition
No.   Eigenvalue             Percent         Percent         Number
1         2.2150              44.299          44.299          1.000
2         1.2277              24.554          68.853          1.804
3         1.1062              22.124          90.978          2.002
4         0.4446               8.892          99.870          4.982
5         0.0065               0.130         100.000        340.939

Some Condition Numbers greater than 100. Multicollinearity is a MILD problem.
```

This section gives an eigenvalue analysis of the independent variables when they have been centered and scaled.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of IV's. Eigenvalues near zero indicate a high degree of is collinearity in the data.

Incremental Percent

Incremental percent is the percent this eigenvalue is of the total. In an ideal situation, these percentages would be equal. Percents near zero indicate collinearity in the data.

Cumulative Percent

This is the running total of the Incremental Percent.

Condition Number

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variances. Condition numbers greater than 1000 indicate a severe collinearity problem while condition numbers between 100 and 1000 indicate a mild collinearity problem.

Eigenvector Percent of Regression-Coefficent-Variance using Centered Correlations Section

```
No.  Eigenvalue          Test1        Test2        Test3
1        2.2150         0.2705       0.2850       1.8773
2        1.2277         0.0330       0.1208      31.1222
3        1.1062         0.8089       0.8397       7.6430
4        0.4446         0.8059       1.0889      59.3291
5        0.0065        98.0817      97.6657       0.0284

No.  Eigenvalue          Test4        Test5
1        2.2150         0.2331       2.3798
2        1.2277         0.0579      23.6898
3        1.1062         0.0015      14.3442
4        0.4446         0.0002      59.5804
5        0.0065        99.7072       0.0058
```

This report displays how the eigenvectors associated with each eigenvalue are related to the independent variables.

No.

The number of the eigenvalue.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is collinearity in your data.

Values

The rest of this report gives a breakdown of what percentage each eigenvector is of the total variation for the regression coefficient. Hence, the percentages sum to 100 down a column.

A small eigenvalue (large condition number) along with a subset of two or more independent variables having high variance percentages indicates a dependency involving the independent variables in that subset. This dependency has damaged or contaminated the precision of the regression coefficients estimated in the subset. Two or more percentages of at least 50% for an eigenvector or eigenvalue suggest a problem. For certain, when there are two or more variance percentages greater than 90%, there is definitely a collinearity problem.

Again, take the following steps when using this table.

1. Find rows with condition numbers greater than 100 (find these in the *Eigenvalues of Centered Correlations* report).
2. Scan across each row found in step 1 for two or more percentages greater than 50. If two such percentages are found, the corresponding variables are being influenced by collinearity problems. You should remove one and re-run your analysis.

Eigenvalues of Uncentered Correlations Section

```
                          Incremental      Cumulative       Condition
No.   Eigenvalue              Percent         Percent          Number
1         5.7963               96.606          96.606           1.000
2         0.1041                1.735          98.340          55.686
3         0.0670                1.116          99.457          86.532
4         0.0214                0.357          99.814         270.830
5         0.0109                0.181          99.995         533.756
6         0.0003                0.005         100.000       17767.041
Some Condition Numbers greater than 1000. Multicollinearity is a SEVERE problem.
```

This report gives an eigenvalue analysis of the independent variables when they have been scaled but not centered (the intercept is included in the collinearity analysis). The eigenvalues for this situation are generally not the same as those in the previous eigenvalue analysis. Also, the condition numbers are much higher.

Eigenvalue

The eigenvalues of the scaled, but not centered, matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is collinearity in your data.

Incremental Percent

Incremental percent is the percent this eigenvalue is of the total. In an ideal situation, these percentages would be equal. Percents near zero mean that there is collinearity in your data.

Cumulative Percent

This is the running total of the *Incremental Percent.*

Condition Number

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variances. There has not been any formalization of rules on condition numbers for uncentered matrices. You might use the criteria mentioned earlier for mild collinearity and severe collinearity. Since the collinearity will always be worse with the intercept in the model, it is advisable to have more relaxed criteria for mild and severe collinearity, say 500 and 5000, respectively.

Eigenvector Percent of Regression-Coefficent-Variance using Uncentered Correlations

```
No.  Eigenvalue         Test1         Test2         Test3
1        5.7963        0.0042        0.0068        0.0826
2        0.1041        0.0308        0.8177        3.8156
3        0.0670        1.1375        0.9627        7.4272
4        0.0214        0.2675        0.9263       51.4298
5        0.0109        0.4157        0.0499       37.2046
6        0.0003       98.1444       97.2367        0.0402

No.  Eigenvalue         Test4         Test5     Intercept
1        5.7963        0.0015        0.1033        0.0397
2        0.1041        0.0610       11.8930        0.2599
3        0.0670        0.0261        0.0897        0.0106
4        0.0214        0.0006       79.7835        1.6692
5        0.0109        0.0931        8.1292       97.0221
6        0.0003       99.8177        0.0013        0.9986
```

This report displays how the eigenvectors associated with each eigenvalue are related to the independent variables.

No.

The number of the eigenvalue.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is collinearity in your data.

Values

The rest of this report gives a breakdown of what percentage each eigenvector is of the total variation for the regression coefficient. Hence, the percentages sum to 100 down a column.

A small eigenvalue (large condition number) along with a subset of two or more independent variables having high variance percentages indicates a dependency involving the independent variables in that subset. This dependency has damaged or contaminated the precision of the regression coefficients estimated in the subset. Two or more percentages of at least 50% for an eigenvector or eigenvalue suggest a problem. For certain, when there are two or more variance percentages greater than 90%, there is definitely a collinearity problem.

Predicted Values with Confidence Limits of Means

```
                                      Standard     95% Lower     95% Upper
            Actual      Predicted     Error Of     Conf. Limit   Conf. Limit
Row         IQ          IQ            Predicted    Of Mean       Of Mean
1           106.000     110.581       7.157        94.391        126.770
2            92.000      98.248       7.076        82.242        114.255
3           102.000      97.616       6.223        83.539        111.693
4           121.000     118.340       8.687        98.689        137.990
5           102.000      96.006       6.369        81.597        110.414
6           105.000     102.233       5.433        89.942        114.523
.                 .           .           .             .             .
.                 .           .           .             .             .
.                 .           .           .             .             .
```

Confidence intervals for the mean response of *Y* given specific levels for the IV's are provided here. It is important to note that violations of any regression assumptions will invalidate these interval estimates.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y*. It is predicted using the values of the IV's for this row. If the input data had all IV values but no value for *Y*, the predicted value is still provided.

Standard Error of Predicted

This is the standard error of the mean response for the specified values of the IV's. Note that this value is not constant for all IV's values. In fact, it is a minimum at the average value of each IV.

Lower 95% C.L. of Mean

This is the lower limit of a 95% confidence interval estimate of the mean of *Y* for this observation.

Upper 95% C.L. of Mean

This is the upper limit of a 95% confidence interval estimate of the mean of *Y* for this observation. Note that you set the alpha level.

Predicted Values with Prediction Limits of Individuals

```
                                      Standard     95% Lower      95% Upper
            Actual      Predicted     Error Of     Pred. Limit    Pred. Limit
Row         IQ          IQ            Predicted    Of Individual  Of Individual
1           106.000     110.581       12.833       81.551         139.611
2            92.000      98.248       12.788       69.320         127.177
3           102.000      97.616       12.336       69.709         125.523
4           121.000     118.340       13.745       87.247         149.433
5           102.000      96.006       12.411       67.930         124.081
6           105.000     102.233       11.958       75.183         129.283
.                 .           .           .              .              .
.                 .           .           .              .              .
.                 .           .           .              .              .
```

A prediction interval for the individual response of *Y* given specific values of the IV's is provided here for each row.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y*. It is predicted using the levels of the IV’s for this row. If the input data had all values of the IV’s but no value for *Y*, a predicted value is provided.

Standard Error of Predicted

This is the standard deviation of the mean response for the specified levels of the IV’s. Note that this value is not constant for all IV’s. In fact, it is a minimum at the average value of each IV.

Lower 95% Pred. Limit of Individual

This is the lower limit of a 95% prediction interval of the individual value of *Y* for the values of the IV’s for this observation.

Upper 95% Pred. Limit of Individual

This is the upper limit of a 95% prediction interval of the individual value of *Y* for the values of the IV’s for this observation. Note that you set the alpha level.

Residual Report

```
                                                    Absolute     Sqrt(MSE)
          Actual     Predicted                      Percent        Without
Row           IQ            IQ         Residual       Error       This Row
1        106.000       110.581           -4.581       4.322         11.085
2         92.000        98.248           -6.248       6.792         10.905
3        102.000        97.616            4.384       4.298         11.136
4        121.000       118.340            2.660       2.199         11.181
5        102.000        96.006            5.994       5.877         10.984
6        105.000       102.233            2.767       2.635         11.241
.              .             .                .           .              .
.              .             .                .           .              .
.              .             .                .           .              .
```

This section reports on the sample residuals, or e_i’s.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y* using the values of the IV’s given on this row.

Residual

This is the error in the predicted value. It is equal to the *Actual* minus the *Predicted.*

Absolute Percent Error

This is percentage that the absolute value of the *Residual* is of the *Actual* value. Scrutinize rows with the large percent errors.

Sqrt(MSE) Without This Row

This is the value of the square root of the mean square error that is obtained if this row is deleted. A perusal of this statistic for all observations will highlight observations that have an inflationary impact on mean square error and could be outliers.

Regression Diagnostics Section

```
        Standardized                      Hat
Row         Residual    RStudent     Diagonal      Cook's D        Dffits      CovRatio
1            -0.5806     -0.5579       0.4514        0.0462       -0.5061        2.9388
2            -0.7847     -0.7665       0.4413        0.0811       -0.6812        2.3714
3             0.5071      0.4851       0.3413        0.0222        0.3492        2.5863
4             0.4315      0.4111       0.6650        0.0616        0.5792        5.3387
5             0.7021      0.6808       0.3575        0.0457        0.5079        2.2506
6             0.3020      0.2862       0.2601        0.0053        0.1697        2.5777
.                  .           .            .             .             .             .
.                  .           .            .             .             .             .
.                  .           .            .             .             .             .
```

This report presents various statistics known as *regression diagnostics*. They let you conduct an influence analysis of the observations. The interpretation of these values is explained in modern regression books. Belsley, Kuh, and Welsch (1980) devote an entire book to the study of regression diagnostics.

These statistics flag observations that exert three types of influence on the regression.

1. *Outliers in the residual space*. The *Studentized Residual*, the *RStudent*, and the *CovRatio* will flag observations that are influential because of large residuals.

2. *Outliers in the X-space*. The *Hat Diagonal* flags observations that are influential because they are outliers in the *X*-space.

3. *Parameter estimates and fit*. The *Dffits* shows the influence on fitted values. It also measures the impact on the regression coefficients. *Cook's D* measures the overall impact that a single observation has on the regression coefficient estimates.

Standardized Residual

The variances of the observed residuals are not equal, making comparisons among the residuals difficult. One solution is to standardize the residuals by dividing by their standard deviations. This will give a set of standardized residuals with constant variance. The formula for this residual is

$$r_j = \frac{e_j}{\sqrt{\mathrm{MSE}\left(1 - h_{jj}\right)}}$$

RStudent

Rstudent is similar to the standardized residual. The difference is the *MSE*(*j*) is used rather than *MSE* in the denominator. The quantity *MSE*(*j*) is calculated using the same formula as *MSE*, except that observation *j* is omitted. The hope is that be excluding this observation, a better estimate of σ^2 will be obtained. Some statisticians refer to these as the *studentized deleted residuals*.

If the regression assumptions of normality are valid, a single value of the RStudent has a *t* distribution with *n*-*p*-1 degrees of freedom.

$$t_j = \frac{e_j}{\sqrt{MSE\left(j\right)\left(1 - h_{jj}\right)}}$$

Hat Diagonal

The hat diagonal, h_{jj}, captures an observation's remoteness in the *X*-space. Some authors refer to the hat diagonal as a measure of *leverage* in the *X*-space. Hat diagonals greater than two times the number of coefficients in the model divided by the number of observations are said to have *high leverage* (i.e., $h_{ii} > 2p/n$).

Cook's D

Cook's distance (Cook's *D*) attempts to measure the influence each observation on all *N* fitted values. The approximate formula for Cook's *D* is

$$D_j = \frac{\sum_{i=1}^{N} w_j \left[\hat{y}_j - \hat{y}_j(i) \right]^2}{ps^2}$$

The $\hat{y}_j(i)$ are found by removing observation *i* before the calculations. Rather than go to all the time of recalculating the regression coefficients *N* times, we use the following approximation

$$D_j = \frac{w_j e_j^2 h_{jj}}{ps^2 \left(1 - h_{jj}\right)^2}$$

This approximation is exact when no weight variable is used.

A Cook's D value greater than one indicates an observation that has large influence. Some statisticians have suggested that a better cutoff value is 4 / (*N* - 2).

DFFITS

DFFITS is the standardized difference between the predicted value with and without that observation. The formula for *DFFITS* is

$$D_j = \left(\frac{r_j^2}{p} \right) \left(\frac{h_{jj}}{1 - h_{jj}} \right)$$

The values of $\hat{y}_j(j)$ and $s^2(j)$ are found by removing observation *j* before the doing the calculations. It represents the number of estimated standard errors that the fitted value changes if the j^{th} observation is omitted from the data set. If |*DFFITS*| > 1, the observation should be considered to be influential with regards to prediction.

CovRatio

This diagnostic flags observations that have a major impact on the generalized variance of the regression coefficients. A value exceeding 1.0 implies that the i^{th} observation provides an improvement, i.e., a reduction in the generalized variance of the coefficients. A value of CovRatio less than 1.0 flags an observation that increases the estimated generalized variance. This is not a favorable condition.

The general formula for the CovRatio is

$$\mathrm{CovRatio}_j = \frac{\det\left[s(j)^2\left(\mathbf{X}(j)'\mathbf{W}\mathbf{X}(j)\right)^{-1}\right]}{\det\left[s^2\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\right]}$$

$$= \frac{1}{1-h_{jj}}\left[\frac{s(j)^2}{s^2}\right]^p$$

Belsley, Kuh, and Welsch (1980) give the following guidelines for the CovRatio.

If CovRatio > 1 + 3p / N then omitting this observation significantly damages the precision of at least some of the regression estimates.

If CovRatio < 1 - 3p / N then omitting this observation significantly improves the precision of at least some of the regression estimates.

DFBETAS Section

```
Row        Test1      Test2      Test3      Test4      Test5
1         0.2160     0.3128    -0.0390    -0.2556     0.1723
2        -0.1123     0.0190    -0.0830     0.0871     0.0045
3         0.1822     0.2370     0.0291    -0.2075     0.0674
4        -0.1792    -0.2157     0.2157     0.2393     0.1963
5         0.3932     0.3443     0.0108    -0.3638     0.1240
6         0.0969     0.0868    -0.0110    -0.0842    -0.0534
.              .          .          .          .          .
.              .          .          .          .          .
.              .          .          .          .          .
```

DFBETAS

The DFBETAS is an influence diagnostic which gives the number of standard errors that an estimated regression coefficient changes if the j^{th} observation is deleted. If one has N observations and p independent variables, there are Np of these diagnostics. Sometimes, Cook’s D may not show any overall influence on the regression coefficients, but this diagnostic gives the analyst more insight into individual coefficients. The criteria of influence for this diagnostic are varied, but Belsley, Kuh, and Welsch (1980) recommend a cutoff of $2/\sqrt{N}$. Other guidelines are ±1 or ±2. The formula for DFBETAS is

$$dfbetas_k = \frac{b_k - b_{k,-j}}{\sqrt{MSE_j\, c_{kk}}}$$

where c_{kk} is the k^{th} row and k^{th} column element of the inverse matrix $(X'X)^{-1}$.

Graphic Residual Analysis

The residuals can be graphically analyzed in numerous ways. Three types of residuals are graphically analyzed here: residuals, rstudent residuals, and partial residuals. For certain, the regression analyst should examine all of the basic residual graphs: the histogram, the density trace, the normal probability plot, the serial correlation plots, the scatter plot of the residuals versus the sequence of the observations, the scatter plot of the residuals versus the predicted value of the dependent variable, and the scatter plot of the residuals versus each of the independent variables.

For the basic scatter plots of residuals versus either the predicted values of *Y* or the independent variables, Hoaglin (1983) explains that there are several patterns to look for. You should note that these patterns are very difficult, if not impossible, to recognize for small data sets.

Point Cloud

A point cloud, basically in the shape of a rectangle or a horizontal band, would indicate no relationship between the residuals and the variable plotted against them. This is the preferred condition.

Wedge

An increasing or decreasing wedge would be evidence that there is increasing or decreasing (nonconstant) variation. A transformation of *Y* may correct the problem, or weighted least squares may be needed.

Bowtie

This is similar to the wedge above in that the residual plot shows a decreasing wedge in one direction while simultaneously having an increasing wedge in the other direction. A transformation of *Y* may correct the problem, or weighted least squares may be needed.

Sloping Band

This kind of residual plot suggests adding a linear version of the independent variable to the model.

Curved Band

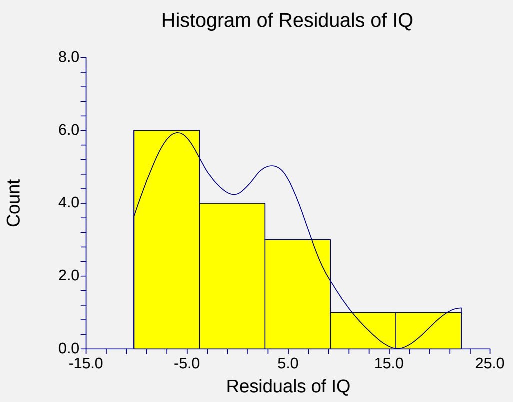
This kind of residual plot may be indicative of a nonlinear relationship between *Y* and the independent variables that was not accounted for. The solution might be to use a transformation on *Y* to create a linear relationship with the *X's*. Another possibility might be to add quadratic or cubic terms of a particular independent variable.

Curved Band with Increasing or Decreasing Variability

This residual plot is really a combination of the wedge and the curved band. It too must be avoided.

Histogram

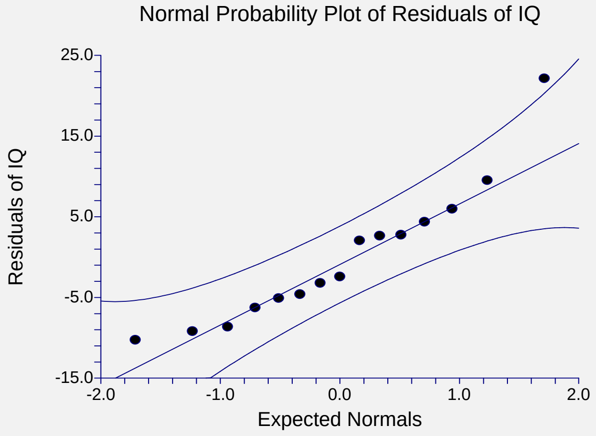
The purpose of the histogram and density trace of the residuals is to evaluate whether they are normally distributed. A dot plot is also given that highlights the distribution of points in each bin of the histogram. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating normality of the residuals. The better choice would be the normal probability plot.



Probability Plot of Residuals

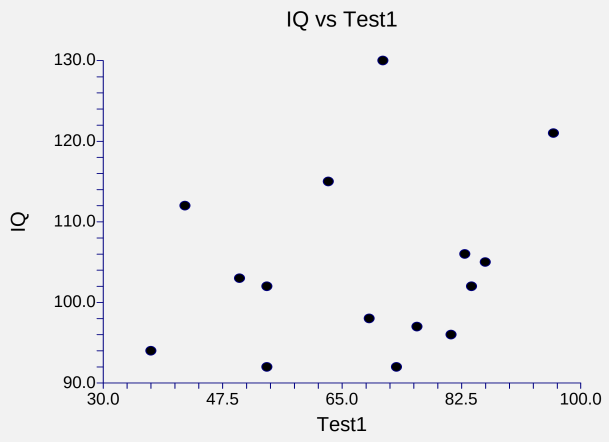
If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.

If the residuals are not normally distributed, then the t-tests on regression coefficients, the F-tests, and any interval estimates are not valid. This is a critical assumption to check.



Plots of Y versus each IV

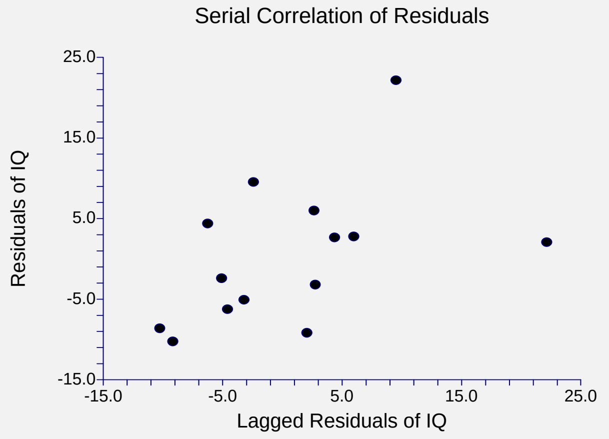
Actually, a regression analysis should always begin with a plot of *Y* versus each IV. These plots often show outliers, curvilinear relationships, and other anomalies.



Serial Correlation of Residuals Plot

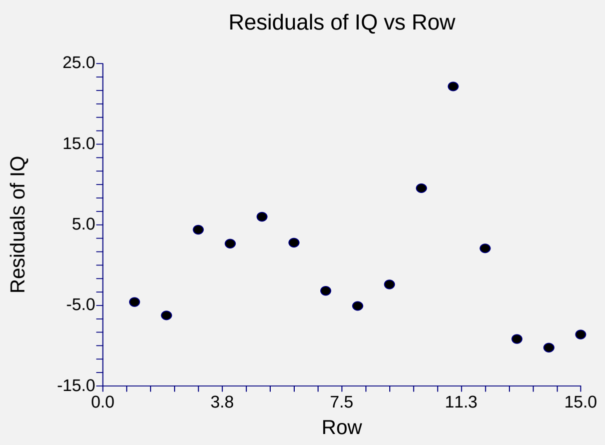
This plot is only useful if your data represent a time series. This is a scatter plot of the j^{th} residual versus the j^{th}-1 residual. The purpose of this plot is to check for first-order autocorrelation.

You would like to see a random pattern of these plotted residuals, i.e., a rectangular or uniform distribution. A strong positive or negative trend would indicate a need to redefine the model with some type of autocorrelation component. Positive autocorrelation or serial correlation means that the residual in time period *j* tends to have the same sign as the residual in time period (*j*-1). On the other hand, a strong negative autocorrelation means that the residual in time period *j* tends to have the opposite sign as the residual in time period (*j*-1). Be sure to check the Durbin-Watson statistic.



Sequence Plot

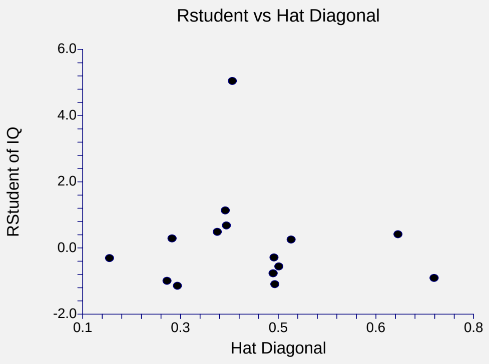
Sequence plots may be useful in finding variables that are not accounted for by the regression equation. They are especially useful if the data were taken over time.



RStudent vs Hat Diagonal Plot

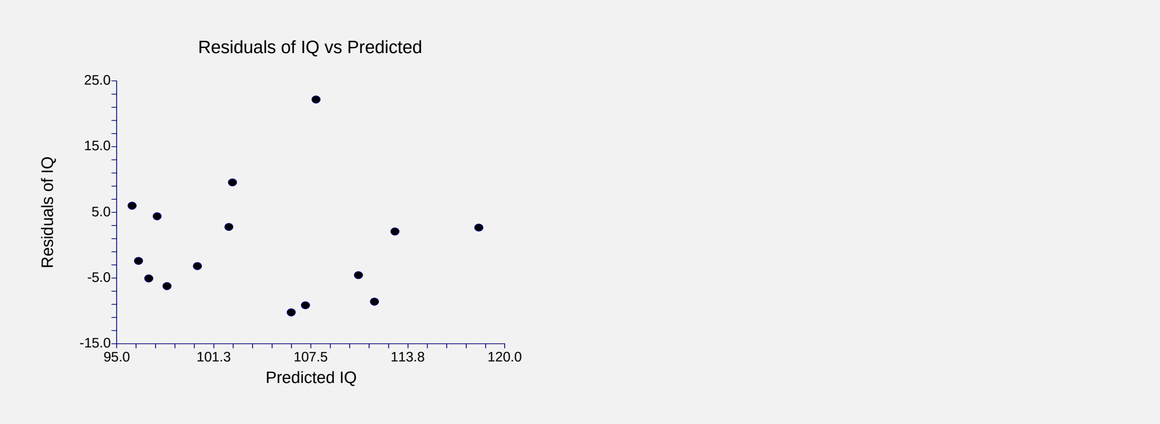
In light of the earlier discussion in the Regression Diagnostics Section, Rstudent is one of the best single-case diagnostics for capturing large residuals, while the hat diagonal flags observations that are remote in the *X*-space. The purpose of this plot is to give a quick visual spotting of observations that are very different from the norm. It is best to rely on the actual regression diagnostics for any formal conclusions on influence. There are three influential realms you might be concerned with

1. Observations that are extreme along the rstudent (vertical) axis are outliers that need closer attention. They may have a major impact on the predictability of the model.
2. Observations that were extreme to the right (i.e., $h_{ii}>2p/n$) are outliers in the *X*-space. These kinds of observations could be data entry errors, so be sure the data is correct before proceeding.
3. Observations that are extreme on both axes are the most influential of all. Double-check these values.



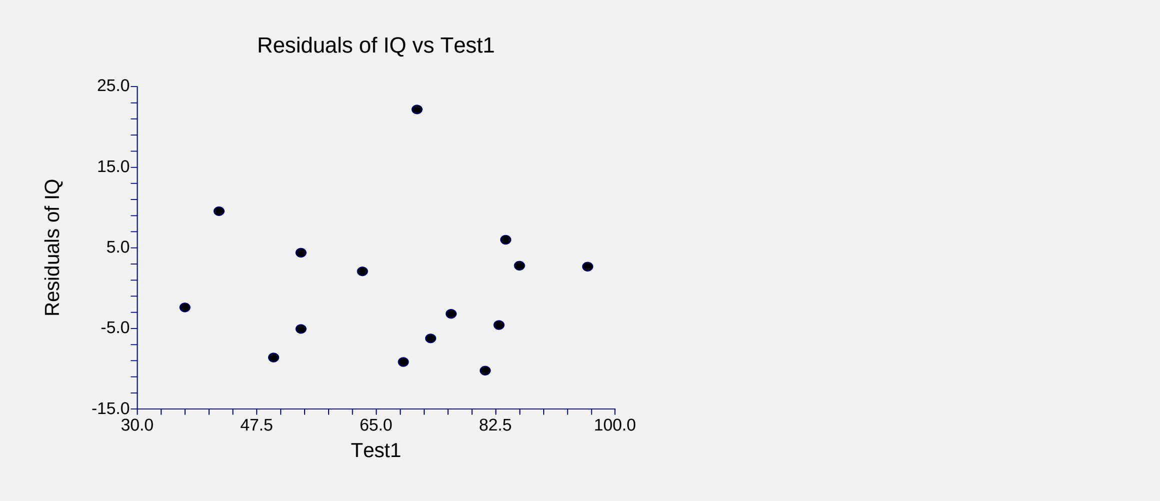
Residual vs Predicted Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance, a violation of a critical regression assumption. The sloping or curved band signifies inadequate specification of the model. The sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.



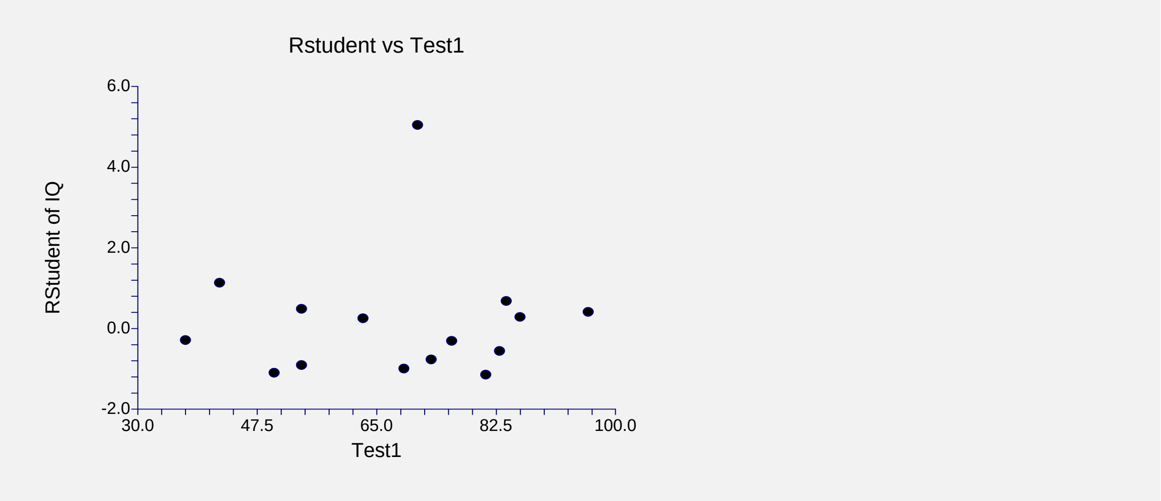
Residual vs Predictor(s) Plot

This is a scatter plot of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.



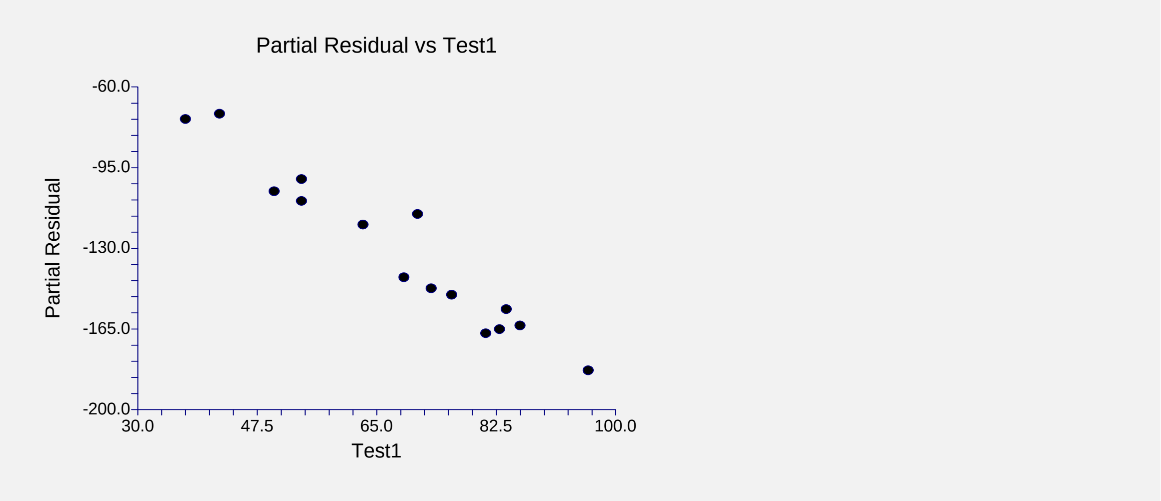
RStudent vs Predictor(s)

This is a scatter plot of the RStudent residuals versus each independent variable. The preferred pattern is a rectangular shape or point cloud. These plots are very helpful in visually identifying any outliers and nonlinear patterns.



Partial Residual Plots

The scatter plot of the partial residuals against each independent variable allows you to examine the relationship between Y and each IV after the effects of the other IV's have been removed. These plots can be used to assess the extent and direction of linearity for each independent variable. In addition, they provide insight as to the correct transformation to apply and information on influential observations. One would like to see a linear pattern between the partial residuals and the independent variable.



Example 2 – Bootstrapping

This section presents an example of how to generate bootstrap confidence intervals with a multiple regression analysis. The tutorial will use the data are in the IQ database. This example will run a regression of IQ on Test1, Test2, and Test4.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Multiple Regression window.

1 Open the IQ dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **IQ.s0**.
- Click **Open**.

2 Open the Multiple Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **IQ**.
- Set the **X's: Numeric Independent Variables** box to **Test1, Test2, Test4**.
- Check the **Calculate Bootstrap C.I.'s** box.

4 Specify the reports.

- Select the **Reports tab**.
- Set the **Select a Group of Reports and Plots** to **Display only those items that are CHECKED BELOW**.
- Check the **Regression Coefficients** box under the **Estimation** heading**.**

5 Specify the bootstrap parameters.

- Select the **Resampling tab**.
- Set the **Samples (N)** to **3000**.
- You may change any of the other parameters as you see fit**.**

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Regression Coefficient Section

Independent Variable	Regression Coefficient	Standard Error	Lower 95% C.L.	Upper 95% C.L.	Standardized Coefficient
Intercept	90.7327	12.8272	62.5003	118.9651	0.0000
Test1	-1.9650	0.9406	-4.0353	0.1053	-3.1020
Test2	-1.6485	0.7980	-3.4048	0.1078	-2.9024
Test4	3.7890	1.6801	0.0912	7.4869	4.7988

This report gives the confidence limits calculated under the assumption of normality. We have displayed it so that we can compare these to the bootstrap confidence intervals.

Bootstrap Section

```
---  Estimation Results    ------    | --- Bootstrap Confidence Limits----
Intercept
Original Value         90.7327   | 0.9000     68.2414          109.0262
Bootstrap Mean         92.1790   | 0.9500     61.6979          113.2208
Bias (BM - OV)          1.4463   | 0.9900     44.6644          123.0653
Bias Corrected         89.2863
Standard Error         13.1402
B(Test1)
Original Value         -1.9650   | 0.9000     -3.0486           -0.1051
Bootstrap Mean         -2.1094   | 0.9500     -3.3245            0.4579
Bias (BM - OV)         -0.1444   | 0.9900     -4.1334            1.5529
Bias Corrected         -1.8206
Standard Error          0.9482
B(Test2)
Original Value         -1.6485   | 0.9000     -2.5591            0.0014
Bootstrap Mean         -1.8020   | 0.9500     -2.8343            0.4772
Bias (BM - OV)         -0.1535   | 0.9900     -3.6419            1.7330
Bias Corrected         -1.4951
Standard Error          0.8336
B(Test4)
Original Value          3.7890   | 0.9000      0.4503            5.7739
Bootstrap Mean          4.0637   | 0.9500     -0.6646            6.3992
Bias (BM - OV)          0.2747   | 0.9900     -2.3355            7.9171
Bias Corrected          3.5144
Standard Error          1.7159
Predicted Mean and Confidence Limits of IQ When Row = 16
Original Value          99.509   | 0.9000      93.047           105.295
Bootstrap Mean          99.796   | 0.9500      90.629           106.717
Bias (BM - OV)           0.287   | 0.9900      85.355           109.743
Bias Corrected          99.222
Standard Error           3.982
Predicted Mean and Confidence Limits of IQ When Row = 17
Original Value         101.264   | 0.9000      96.723           105.457
Bootstrap Mean         101.319   | 0.9500      95.576           106.753
Bias (BM - OV)           0.055   | 0.9900      92.511           108.933
Bias Corrected         101.209
Standard Error           2.810
Predicted Value and Prediction Limits of IQ When Row = 16
Original Value          99.509   | 0.9000      69.984           122.901
Bootstrap Mean         101.009   | 0.9500      63.420           126.959
Bias (BM - OV)           1.500   | 0.9900      44.415           137.078
Bias Corrected          98.009
Standard Error          16.323
Predicted Value and Prediction Limits of IQ When Row = 17
Original Value         101.264   | 0.9000      72.296           123.954
Bootstrap Mean         103.159   | 0.9500      66.145           128.934
Bias (BM - OV)           1.895   | 0.9900      50.664           137.938
Bias Corrected          99.370
Standard Error          15.971
Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.
```

This report provides bootstrap intervals of the regression coefficients and predicted values for rows 16 and 17 which did not have an IQ (*Y*) value. Details of the bootstrap method were presented earlier in this chapter.

It is interesting to compare these confidence intervals with those provided in the Regression Coefficient report. The most striking difference is that the lower limit of the 95% bootstrap confidence interval for B(Test4) is now negative. When the lower limit is negative and the upper limit is positive, we know that a hypothesis test would not find the parameter significantly different from zero. Thus, while the regular confidence interval of B(Test4) indicates statistical significance (since both limits are positive), the bootstrap confidence interval does not.

Note that since these results are based on 3000 random bootstrap samples, they will differ slightly from the results you obtain when you run this report.

Original Value

This is the parameter estimate obtained from the complete sample without bootstrapping.

Bootstrap Mean

This is the average of the parameter estimates of the bootstrap samples.

Bias (BM - OV)

This is an estimate of the bias in the original estimate. It is computed by subtracting the original value from the bootstrap mean.

Bias Corrected

This is an estimated of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate.

Standard Error

This is the bootstrap method’s estimate of the standard error of the parameter estimate. It is simply the standard deviation of the parameter estimate computed from the bootstrap estimates.

Conf. Level

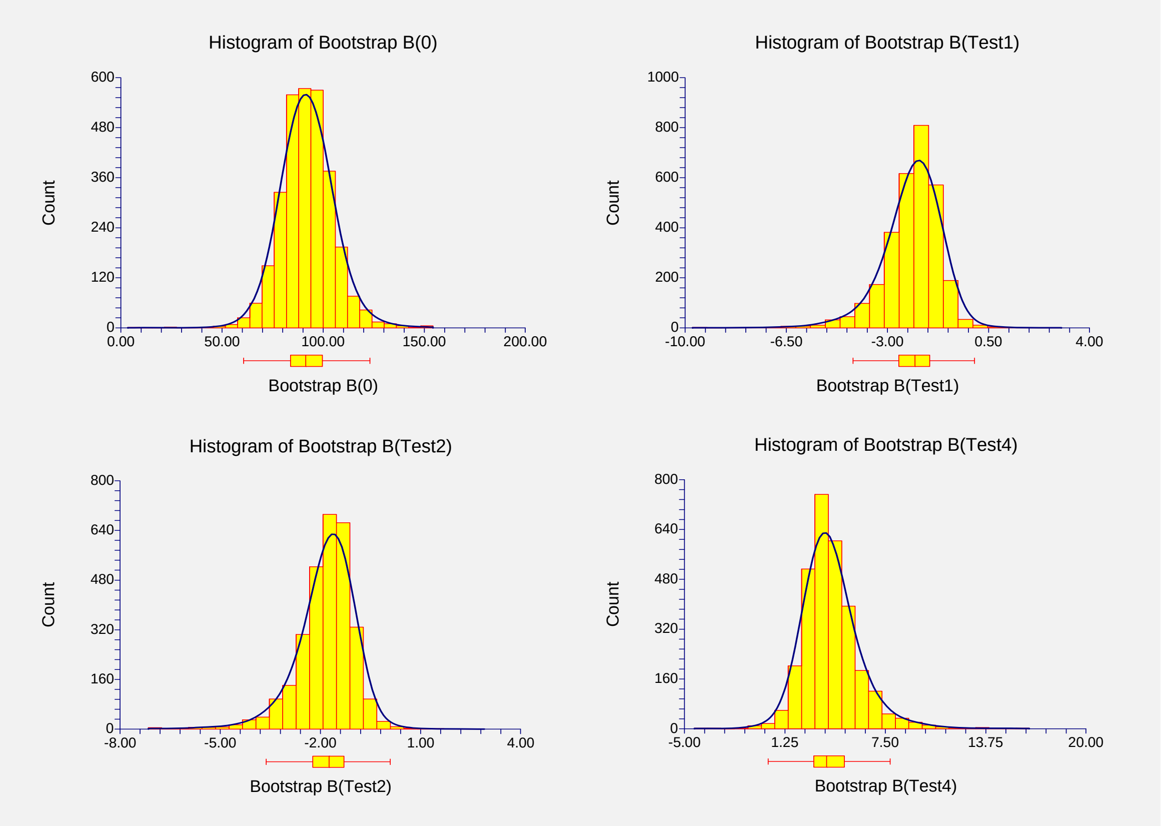
This is the confidence coefficient of the bootstrap confidence interval given to the right.

Bootstrap Confidence Limits - Lower and Upper

These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

Bootstrap Histograms Section



Histogram of Bootstrap B(0)

Histogram of Bootstrap B(Test1)

Histogram of Bootstrap B(Test2)

Histogram of Bootstrap B(Test4)

Each histogram shows the distribution of the corresponding parameter estimate.

Note that the number of decimal places shown in the horizontal axis is controlled by which histogram style file is selected. In this example, we selected Bootstrap2, which was created to provide two decimal places.

Example 3 – Robust Regression

This section presents an example of how to generate bootstrap confidence intervals with a multiple regression analysis. The tutorial will use the data are in the IQ database. This example will run a regression of IQ on Test1 through Test5.

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Multiple Regression window.

1 Open the IQ dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **IQ.s0**.
- Click **Open**.

2 Open the Multiple Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **IQ**.
- Set the **X's: Numeric Independent Variables** box to **Test1-Test5**.
- Check the **Perform Robust Regression** box.

4 Specify the reports.

- Select the **Reports tab**.
- Set the **Select a Group of Reports and Plots** to **Display only those items that are CHECKED BELOW**.
- Check the **Equation** box**.**
- Check the **Robust Coefficients** box.
- Check the **Robust Percentiles** box.
- Check the **Robust Residuals** box.

5 Specify the robust regression parameters.

- Select the **Robust tab**.
- Set the **Robust Method** to **Huber's Method**.
- Set the **Minimum % Beta Change** to **1.0**.
- You may change any of the other parameters as you see fit**.**

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	61.6716	16.5684	3.722	0.0048	Yes	0.9105
Test1	-1.4481	0.6706	-2.159	0.0591	No	0.4873
Test2	-1.2148	0.5713	-2.126	0.0624	No	0.4756
Test3	0.1945	0.1463	1.329	0.2165	No	0.2216
Test4	2.9261	1.1947	2.449	0.0368	Yes	0.5888
Test5	0.1186	0.1389	0.854	0.4152	No	0.1195

This report gives the robust regression coefficients as well as *t*-tests. Note that the statistical tests are approximate because we are using robust regression. You could generate bootstrap robust confidence intervals to supplement these results.

Robust Regression Coefficient Section

Robust Iteration	Max % Change in any Beta	Robust B(0)	Robust B(1)	Robust B(2)	Robust B(3)
0	100.000	85.2404	-1.9336	-1.6599	0.1050
1	136.537	77.6730	-1.7920	-1.5307	0.1384
2	209.779	73.4249	-1.7126	-1.4582	0.1571
3	33.306	71.3356	-1.6735	-1.4225	0.1663
4	24.365	69.2981	-1.6354	-1.3877	0.1753
5	30.744	66.1007	-1.5756	-1.3331	0.1894
6	18.776	61.8967	-1.4569	-1.2239	0.1960
7	0.794	61.8156	-1.4534	-1.2196	0.1944

This report shows the largest percent change in any of the regression coefficients as well as the first four regression coefficients. The first iteration always shows the ordinary least squares estimates on the full dataset so that you can compare these value with those that occur after a few robust iterations.

This report allows you to determine if enough iterations have been run for the coefficients to have stabilized. In this example, the coefficients have stabilized. If they had not, we would decrease the value of the Minimum % Beta Change and rerun the analysis.

Robust Percentiles of Residuals Section

Iter.	Max % Change	-------------------- Percentiles of Absolute Residuals ------------------------			
No.	in any Beta	25th	50th	75th	100th
0	100.000	2.767	5.073	9.167	22.154
1	136.537	2.693	4.225	8.314	25.177
2	209.779	1.917	4.493	7.834	26.874
3	33.306	1.695	4.341	7.599	27.709
4	24.365	1.641	3.714	7.369	28.523
5	30.744	1.555	3.118	7.008	29.800
6	18.776	1.526	2.529	7.115	30.702
7	0.794	1.508	2.460	7.137	30.677

The purpose of this report is to highlight the maximum percentage changes among the regression coefficients and to show the convergence of the absolute value of the residuals after a selected number of iterations.

Iter. No.

This is the robust iteration number.

Max % Change in any Beta

This is the maximum percentage change in any of the regression coefficients from one iteration to the next. This quantity can be used to determine if enough iterations have been run. Once this value is less than five percent, little is gained by further iterations.

Percentiles of Absolute Residuals

The absolute values of the residuals for this iteration are sorted and the percentiles are calculated. We want to terminate the iteration process when there is little change in median of the absolute residuals.

Robust Residuals and Weights Section

```
                                                   Absolute
             Actual   Predicted                     Percent        Robust
Row              IQ          IQ     Residual          Error        Weight
1           106.000     104.752        1.248          1.177        1.0000
2            92.000      97.256       -5.256          5.713        1.0000
3           102.000      99.878        2.122          2.080        1.0000
4           121.000     121.532       -0.532          0.440        1.0000
5           102.000      98.350        3.650          3.578        1.0000
6           105.000      99.843        5.157          4.911        1.0000
7            97.000      98.237       -1.237          1.275        1.0000
8            92.000      94.433       -2.433          2.644        1.0000
9            94.000      96.407       -2.407          2.561        1.0000
10          112.000     104.221        7.779          6.945        0.7702
11          130.000      99.319       30.681         23.601        0.1927
12          115.000     113.485        1.515          1.318        1.0000
13           98.000     105.163       -7.163          7.310        0.8318
14           96.000     104.776       -8.776          9.142        0.6842
15          103.000     104.767       -1.767          1.715        1.0000
```

The predicted values, the residuals, and the robust weights are reported for the last iteration. These robust weights can be saved for use in a weighted regression analysis, or they can be used as a filter to delete observations with a weight less than some number, say 0.20, in an ordinary least squares regression analysis.

Note that in this analysis, row 11 appears to be an outlier.

Row

This is the number of the row. Rows whose weight is less than 0.1 are starred.

Actual

This is the actual value of the dependent variable.

Predicted

This is the predicted value of *Y* based on the robust regression equation from the final iteration.

Residual

The residual is the difference between the Actual and Predicted values of *Y*.

Robust Weight

Once the convergence criteria for the robust procedure have been met, these are the final weights for each observation.

These weights will range from zero to one. Observations with a low weight make a minimal contribution to the determination of the regression coefficients. In fact, observations with a weight of zero have been deleted from the analysis. These weights can be saved and used again in a weighted least squares regression.

Example 4 – Variable Subset Selection

This section presents an example of how to select a subset of the available IV's that are the most useful in predicting *Y*. The tutorial will use the data are in the IQ database. In this example, we will select a subset from the five IV's available.

You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Multiple Regression window.

1 Open the IQ dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **IQ.s0**.
- Click **Open**.

2 Open the Multiple Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **IQ**.
- Set the **X's: Numeric Independent Variables** box to **Test1-Test5**.

4 Specify the subset selection method.

- On the Multiple Regression window, select the **Model tab**.
- Set the **Subset Selection** box to **Hierarchical Forward with Switching**.
- Set the **Max Terms in Subset** box to **6**.
- Set the **Which Model Terms** box to **Up to 2-Way**.

5 Specify the reports.

- Select the **Reports tab**.
- Set the **Select a Group of Reports and Plots** to **Display items appropriate for a STANDARD ANALYSIS**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Subset Selection Summary Section

No. Terms	No. X's	R-Squared Value	R-Squared Change
1	1	0.1379	0.1379
2	2	0.1542	0.0163
3	3	0.2466	0.0924
4	4	0.3587	0.1121
5	5	0.5681	0.2094
6	6	0.5877	0.0196

This report shows the number of terms, number of IV’s, and *R*-squared values for each subset size. This report is used to determine an appropriate subset size for a second run. You search the table for a subset size after which the *R*-squared increases only slightly as more variables are added.

In this example, there appears to be two places where a break occurs: from 1 to 2 terms and from 5 to 6 terms. Under normal circumstances, we would pick from a subset size of 5 for a second run. However, because the sample size in this example is only 15, we would not want to go above a subset size of 3 (our rule of thumb is *N*/#IV’s > 5).

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	R2	Term Entered	Term Removed
0	Add	0	0	0.0000	Intercept	
1	Add	1	1	0.1379	Test4	
2	Add	2	2	0.1542	Test3	
3	Add	3	3	0.2466	Test3*Test3	
4	Add	4	4	0.3587	Test4*Test4	
5	Add	5	5	0.4149	Test2	
6	Switch	5	5	0.4203	Test1	Test3*Test3
7	Switch	5	5	0.5681	Test2*Test2	Test4*Test4
8	Add	6	6	0.5877	Test1*Test1	

This report shows the details of which variables were added or removed at each step in the search procedure. The final model for three IV’s would include Test4, Test3, and Test3*Test3.

Because of the restrictions due to our use of hierarchical models, you might run an analysis using the Forward with Switching option as well as a search without 2-way interactions. Because of the small sample size, these options produce models with much larger *R*-squared values. However, it is our feeling that this larger *R*-squared values occur because the extra variables are actually fitting random error rather than a reproducible pattern.

Example 5 – Sales Price Prediction

This section presents an example of using multiple regression to construct an equation that predicts the sales price of a home based on a few basic IV's such as square footage, lot size, and so on. The RESALE database contains several variables relating to the sales price of a house. These include year built, number of bedrooms, number of bathrooms, size of garage, number of fireplaces, overall quality rating, amount of building with brick, finished square footage, total square footage, and lot size.

The RESALE database contains data on 150 sales that took place recently. Our task is to develop a mathematical model that relates sales price to the IV's listed about and then use this model to predict the eventual sales price for two additional properties.

Step 1 – View Scatter Plots

The starting point in such an analysis is to view individual scatter plots of sales price versus each of the potential IV's looking for outliers, curvilinear patterns, and other anomalies. Although we could create these scatter plots in other procedures, we will use the Multiple Regression procedure to do so.

You may follow along here by making the appropriate entries or load the completed template **Example5-1** from the Template tab of the Multiple Regression window.

1 **Open the RESALE dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **RESALE.s0**.
- Click **Open**.

2 **Open the Multiple Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **Price**.
- Set the **X's: Numeric Independent Variables** box to **Year-Lotsize**.

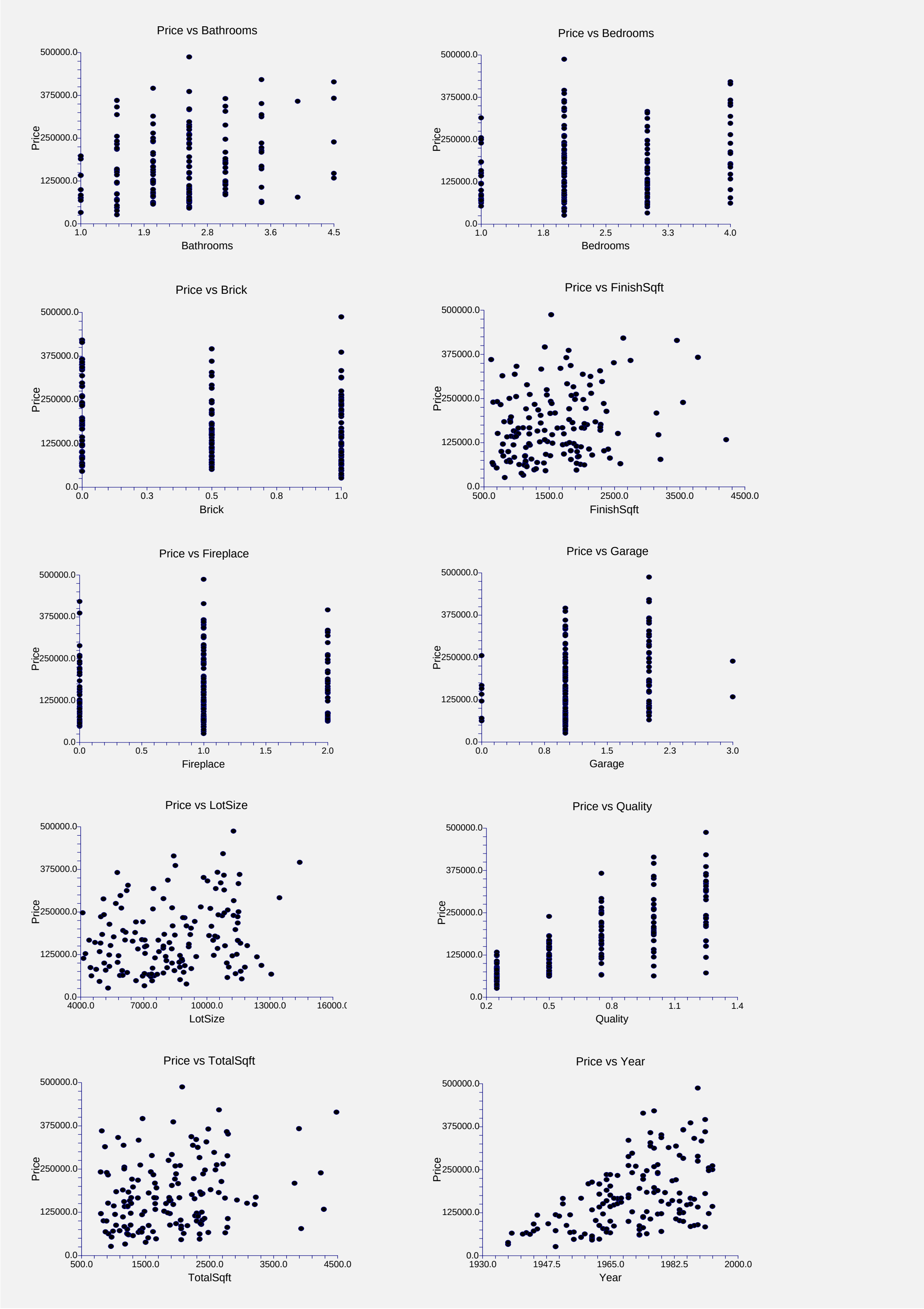
4 **Specify the reports.**

- Select the **Reports tab**.
- Set the **Select a Group of Reports and Plots** to **Display only those items that are CHECKED BELOW**.
- Check the **Y-X's Plots** box under **Select Plots**.

5 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Scatter Plot Output



Looking at these plots, we notice that Bathrooms, Quality, and Year appear to have the most direct relationship with price. We cannot spot any outliers, so we procedure to the next step.

Step 2 – Use Robust Regression to Find Outliers

Although we could not spot any outliers on the scatter plots, it is important to make sure that we have not missed any. To do this, we run a robust regression analysis and search the robust weights for values less that 0.20 (which we define as an outlier).

This analysis assumes that you have just completed Example 5-1. You may follow along here by making the appropriate entries or load the completed template **Example5-2** from the Template tab of the Multiple Regression window.

1 **On the Variables tab.**

- Check the **Perform Robust Regression** box.

2 **On the Reports tab.**

- Check the **Robust Coefficients** box.
- Check the **Robust Residuals** box.

3 **On the Robust tab.**

- Set the **Minimum % Beta Change** to **0.1**.
- Set the **Cutoff for Weight Report** to **0.40**.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Robust Regression Output

```
Robust Regression Coefficients Section
Robust     Max % Change         Robust         Robust       Robust       Robust
Iteration   in any Beta           B(0)           B(1)         B(2)         B(3)
0               100.000  -6975033.8132      -377.5098    9068.0709    8419.6302
1               165.840  -6924967.1956       248.5524    8082.0650    5634.3540
2                13.563  -6909107.4943       279.5934    8041.3988    5375.8683
3                13.681  -6903260.9959       278.5935    8053.7521    5328.9806
4                 4.179  -6902010.4388       281.6196    8059.4888    5310.6893
5                 1.071  -6901686.6682       283.5483    8060.9746    5305.8389
6                 0.269  -6901619.0601       284.3109    8061.4158    5304.5386
7                 0.089  -6901609.4729       284.5639    8061.5290    5304.1372

Robust Residuals and Weights
                                                              Absolute
                 Actual      Predicted                         Percent       Robust
Row               Price          Price        Residual           Error       Weight
55            32900.000     -70304.342      103204.342         313.691       0.3468
120          117800.000     210523.459      -92723.459          78.713       0.3860
150          487200.000     373867.349      113332.651          23.262       0.3158
```

From a perusal of these reports, we learn that there are three potential outliers: rows 55, 120, and 150. However, their robust weights are much larger than the cutoff value of 0.200 which we set as an indicator of when an observation is an outlier. So, even though these three observations are predicted poorly, we decide to leave them in the dataset for the rest of the analysis.

Step 3 – Variable Selection

The next step is to search for the most useful subset of the IV's. To do this, we made an initial search for each subset up to ten IV's. We will study the R-squared values to determine a reasonable subset size.

This analysis assumes that you have just completed Example 5-2. You may follow along here by making the appropriate entries or load the completed template **Example5-3** from the Template tab of the Multiple Regression window.

1 On the Variables tab.

- Make sure the **Perform Robust Regression** box is not checked.

2 On the Model tab.

- Set the **Subset Selection** box to **Hierarchical Forward with Switching**.
- Set the **Max Terms in Subset** to **10**.
- Set the **Which Model Terms** to **Up to 2-Way**.

3 On the Reports tab.

- Check the **Subset Summary** box.
- Check the **Subset Detail** box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Variable Selection Output

```
Subset Selection Summary Section
No.                  No.     R-Squared    R-Squared
Terms                X's         Value       Change
1                      1        0.5212       0.5212
2                      2        0.7676       0.2464
3                      3        0.8440       0.0764
4                      4        0.8929       0.0489
5                      5        0.8956       0.0027
6                      6        0.8969       0.0014
7                      7        0.9009       0.0039
8                      8        0.9020       0.0011
9                      9        0.9031       0.0011
10                    10        0.9037       0.0006

Subset Selection Detail Section
                   No. of   No. of            Term                        Term
Step     Action     Terms      X's       R2   Entered                     Removed
0           Add         0        0   0.0000   Intercept
1           Add         1        1   0.5212   Quality
2           Add         2        2   0.7676   Year
3           Add         3        3   0.8440   TotalSqft
4           Add         4        4   0.8929   LotSize
5           Add         5        5   0.8956   Bedrooms
6           Add         6        6   0.8968   Brick
7        Switch         6        6   0.8969   Brick*Brick                 Bedrooms
8           Add         7        7   0.9009   Bedrooms
9           Add         8        8   0.9020   Fireplace
10          Add         9        9   0.9031   Fireplace*Fireplace
11          Add        10       10   0.9037   Brick*Fireplace
```

Scanning down the *R*-squared values, it is easy to see that the appropriate subset size is four. With four IV's, an *R*-squared of 0.8929 is achieved which is impressive for this type of data. From the Subset Selection Detail report, we learn that the four IV's are Quality, Year, TotalSqrt, and LotSize. These seem to be a reasonable basis for sales price estimation.

Step 4 – Standard Regression

The next step is to generate a standard regression analysis using the four IV's that were selected in the last step.

This analysis assumes that you have just completed Example 5-3. You may follow along here by making the appropriate entries or load the completed template **Example5-4** from the Template tab of the Multiple Regression window.

1 **On the Variables tab.**

- Set the **X's: Numeric Independent Variables** box to **YEAR, QUALITY, TOTALSQRT, LOTSIZE**.

2 **On the Model tab.**

- Set the **Subset Selection** box to **None**.
- Set the **Which Model Terms** to **Up to 1-Way**.

3 **On the Reports tab.**

- Set the **Select a Group of Reports and Plots** option to **Display items appropriate for a Standard Analysis**.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Standard Regression Output

```
Run Summary Section
Parameter                Value               Parameter                    Value
Dependent Variable       Price               Rows Processed               150
Number Ind. Variables    4                   Rows Filtered Out            0
Weight Variable          None                Rows with X's Missing        0
R2                       0.8929              Rows with Weight Missing     0
Adj R2                   0.8899              Rows with Y Missing          0
Coefficient of Variation 0.1858              Rows Used in Estimation      150
Mean Square Error        1.049649E+09        Sum of Weights               150.000
Square Root of MSE       32398.29            Completion Status            Normal Completion
Ave Abs Pct Error        22.636
```

We have only included the Run Summary report here. You can look at the complete output when you run this example. We note that the final *R*-squared value is 0.8929, which is pretty good, but the average absolute percent error is 22.636%, which is disturbing.

This completes this analysis. If you wanted to use these results to predict the sales price of additional properties, you would simple add the data to the bottom of the database, leaving the Price variable blank. The Predicted Individuals report will give the estimates and prediction limits for these additional properties.

Example 6 – Checking the Parallel Slopes Assumption in Analysis of Covariance

An example of how to test the parallel slopes assumption is given in the General Linear Models chapter. Unfortunately, hand calculations and extensive data transformations are required to complete this test. This example will show you how to run this test without either transformations or hand calculations.

The ANCOVA database contains three variables: State, Age, and IQ. The researcher wants to test for IQ differences across the three states while controlling for each subjects age. An analysis of covariance should include a preliminary test of the assumption that the slopes between age and IQ are equal across the three states. Without parallel slopes, differences among mean state IQ's depend on age.

It turns out that a test for parallel slopes is a test for an Age by State interaction. All that needs to be done is to include this term in the model and the appropriate test will be generated.

You may follow along here by making the appropriate entries or load the completed template **Example6** from the Template tab of the Multiple Regression window.

1 Open the ANCOVA dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **ANCOVA.s0**.
- Click **Open**.

2 Open the Multiple Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **IQ**.
- Set the **X's: Numeric Independent Variables** box to **AGE**.
- Set the **X's: Categorical Independent Variables** box to **STATE**.
- Set the **Default Contrast Type** to **Standard Set**.

4 Specify the model.

- Select the **Model tab**.
- Set the **Which Model Terms** box to **Full Model**. This will cause the State by Age interaction to be added to the model.

5 Specify which reports.

- Select the **Reports tab**.
- Set the **Select a Group of Reports and Plots** window to **Display only those items that are CHECKED BELOW.**
- Check the box next to **ANOVA Detail**.

6 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Analysis of Variance Detail Section

```
Model                          Sum of        Mean                Prob     Power
Term            DF     R2      Squares       Square    F-Ratio   Level    (5%)
Intercept        1             313345.2      313345.2
Model            5   0.2438    80.15984      16.03197    1.547   0.2128   0.4472
Age              1   0.0296    9.740934      9.740934    0.940   0.3419   0.1537
State            2   0.1417    46.57466      23.28733    2.248   0.1274   0.4123
Age*State        2   0.1178    38.72052      19.36026    1.869   0.1761   0.3500
Error           24   0.7562    248.6402      10.36001
Total(Adjusted)29    1.0000       328.8      11.33793
```

The F-Value for the Age*State interaction term is 1.869. This matches the result that was obtained by hand calculations in the General Linear Model example. Since the probability level of 0.1761 is not significant, we cannot reject the assumption that the three slopes are equal.

Example 7 – Analyzing Pre-Post Data with both Categorical and Numeric IV's

The PREPOST database contains the results of a study involving 144 subjects that were divided into three groups. The first group (Control) received a placebo, the second group (Dose20) received a small dose of the drug of interest, and the third group (Dose40) received a large dose of the drug of interest. Each subject response was measured before (Pre) and after (Post) the drug was administered, and the gain from Pre to Post was calculated. Also, each subject's propensity score was measured. This Propensity is a combined index created from several demographic variables. The age group (Age) of each subject was also recorded.

The goal of the research is to build a regression model from this data that will allow the gain scores to be predicted. The model should include all significant interaction terms.

Step 1 – Scan for Outliers Using Robust Regression

The first step is to scan for outliers using robust regression. Of course, you should also look at the scatter plots of *Y* versus each IV. The robust regression is useful because it provides a list of potential outliers even when interactions are included. It is often difficult to find true outliers when interactions are included.

You may follow along here by making the appropriate entries or load the completed template **Example7-1** from the Template tab of the Multiple Regression window.

1 **Open the PREPOST dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **PREPOST.s0**.
- Click **Open**.

2 Open the Multiple Regression window.

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Multiple Regression**. The Multiple Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Multiple Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **Gain**.
- Set the **X's: Numeric Independent Variables** box to **Pre, Propensity**.
- Set the **X's: Categorical Independent Variables** box to **Group, Age**.
- Check the **Perform Robust Regression** box.

4 Specify the model.

- Select the **Model tab**.
- Set the **Which Model Terms** box to **Up to 2-Way**. This will cause the interactions and powers to be added to the model.

5 On the Reports tab.

- Set the **Select a Group of Reports and Plots** box to **Display only those items that are CHECKED BELOW**.
- Check the **Run Summary** box.
- Check the **Robust Coefficients** box.
- Check the **Robust Percentiles** box.
- Check the **Robust Residuals** box.

6 On the Robust tab.

- Set the **Minimum % Beta Change** to **1.0**.
- Set the **Maximum Iterations** to **20**.
- Set the **Cutoff for Weight Report** to **0.50**.

7 Run the procedure.

- From the Run menu, select **Run Procedure**.

Robust Regression Output

```
Robust Regression Coefficients Section
Robust       Max % Change              Robust       Robust     Robust     Robust
Iteration      in any Beta               B(0)         B(1)       B(2)       B(3)
0                100.000             -20.2420      -1.2082    -2.8321    33.2946
1               1428.034             -20.8166      -0.5774    -2.7455    32.6539
2                 33.511             -20.8870      -0.4605    -2.7529    32.4686
3                 13.860             -20.9568      -0.5066    -2.8545    32.1934
4                 11.311             -20.9678      -0.5639    -2.9666    31.9287
5                  5.393             -20.9945      -0.5826    -3.0258    31.6974
6                  6.463             -21.0137      -0.5903    -3.0500    31.5122
7                  5.753             -21.0157      -0.5951    -3.0674    31.3985
8                  4.314             -21.0250      -0.5964    -3.0758    31.3085
9                  3.430             -21.0304      -0.5968    -3.0831    31.2475
10                 2.585             -21.0323      -0.5970    -3.0894    31.2096
11                 1.809             -21.0334      -0.5972    -3.0940    31.1859
12                 1.215             -21.0340      -0.5972    -3.0972    31.1710
13                 0.797             -21.0344      -0.5973    -3.0993    31.1616
```

```
Robust Residuals and Weights
                                                           Absolute
                    Actual       Predicted                  Percent      Robust
Row                  Price           Price    Residual        Error      Weight
9                  222.000         204.114      17.886        8.057      0.3178
16                 174.000         159.132      14.868        8.545      0.3824
45                 214.000         196.278      17.722        8.281      0.3207
54                  57.000          69.430     -12.430       21.807      0.4569
99                 260.000         232.567      27.433       10.551      0.2072
105                 73.000          85.346     -12.346       16.913      0.4598
116                204.000         187.501      16.499        8.088      0.3445
144                  6.000          -6.211      12.211      203.520      0.4650
```

There are only a few suspected outliers. Row 99 was especially suspicious since its weight is almost down to 0.20. We also looked at the Regression Diagnostics report and found that these rows also had large values RStudent and Dffits. However, since we could find nothing wrong with the data for these subjects and since we want our final equation to represent as wide of a population as possible, we decided to include these rows in the rest of the analysis.

Step 2 – Search for a Parsimonious Model

Once we have determined that our data is as free of large outliers as we wish, our next task is to conduct a variable selection phase to find a model with as few IV's as possible which still achieves a high *R*-squared value. The Run Summary report (not shown above) listed the an *R*-squared of 0.9894 with a total of 21 IV's. Our goal in this phase is to substantially decrease the number of IV's while achieving an *R*-squared near 0.9894. Because we are fitting interactions, we will conduct as hierarchical forward search with switching.

Note that the changes listed below assume that you have just completed Step 1. You may follow along here by making the appropriate entries or load the completed template **Example7-2** from the Template tab of the Multiple Regression window.

1 **Specify the variables.**

- On the Multiple Regression window, select the **Variables tab**.
- Make sure the **Perform Robust Regression** box is not checked.

2 **Specify the model.**

- Select the **Model tab**.
- Set the **Subset Selection** box to **Hierarchical Forward with Switching**.
- Set the **Max Terms in Subset** box to **10**.
- Set the **Which Model Terms** box to **Up to 2-Way**.

3 **On the Reports tab.**

- Set the **Select a Group of Reports and Plots** box to **Display only those items that are CHECKED BELOW**.
- Check the **Run Summary** box.
- Check the **Subset Summary** box.
- Check the **Subset Detail** box.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**.

Variable Selection Output

```
Subset Selection Summary Section
No.                       No.     R-Squared     R-Squared
Terms                     X's         Value        Change
1                           1        0.3514        0.3514
2                           3        0.7334        0.3821
3                           5        0.7433        0.0099
4                           9        0.7618        0.0185
5                           7        0.9854        0.2236
6                           8        0.9862        0.0008
7                          10        0.9879        0.0017
8                          11        0.9880        0.0001
9                          16        0.9885        0.0005
10                         18        0.9889        0.0003

Subset Selection Detail Section
                    No. of    No. of             Term                        Term
Step       Action    Terms      X's        R2    Entered                     Removed
0             Add        0        0    0.0000    Intercept
1             Add        1        1    0.3514    Propensity
2             Add        2        2    0.7290    Group
3          Switch        2        3    0.7334    Pre                         Propensity
4             Add        3        4    0.7433    Age
5             Add        4        6    0.7618    Age*Group
6             Add        5       10    0.7690    Pre*Pre
7          Switch        5        8    0.9822    Group*Pre                   Age*Group
8          Switch        5        7    0.9854    Propensity                  Age
9             Add        6        8    0.9862    Propensity*Propensity
10            Add        7        9    0.9879    Group*Propensity
11            Add        8       11    0.9880    Pre*Propensity
12            Add        9       13    0.9880    Age
13         Switch        9       15    0.9884    Age*Group                   Group*Propensity
14         Switch        9       16    0.9885    Age*Pre                     Pre*Propensity
15            Add       10       17    0.9889    Age*Propensity
```

We notice from the Subset Selection Summary report that the first five terms achieve an *R*-squared of 0.9854. After that, additional terms increase *R*-squared very little. We decide to include the first five terms in our model.

The Subset Selection Detail report shows that these five terms are: Group, Pre, Propensity, Pre*Pre, and Group*Pre.

Step 3 – Estimate the Model

The next step is to estimate the regression equation and evaluate the residual plots. There are two ways to create the model. The first way is to reset the maximum number of terms to five and rerun the subset selection. The second way is enter the final model in the Custom Model box. This has the advantage that you can run other analyses, such as robust regression, which are not possible during a variable search. So we setup the analysis using the second method.

Note that the changes listed below assume that you have just completed Step 2. You may follow along here by making the appropriate entries or load the completed template **Example7-3** from the Template tab of the Multiple Regression window.

1 Specify the variables.

- On the Multiple Regression window, select the **Variables tab**.
- Remove **Age** from the list of Categorical Independent Variables since it was not included in the final model.

2 **Specify the model.**

- Select the **Model tab**.
- Set the **Subset Selection** box to **None**.
- Set the **Which Model Terms** box to **Custom Model**.
- Set the **Custom Model** box to **Group Pre Pre*Pre Group*Pre Propensity**.

3 **On the Reports tab.**

- Set the **Select a Group of Reports and Plots** box to **Display only those items that are CHECKED BELOW**.
- Check the **Run Summary** box.
- Check the **Equation** box.
- Check the **Regression Coefficients** box.
- Check the **ANOVA Detail** box.
- Check the **Residuals vs X Plots** box.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**.

Standard Regression Output

```
Run Summary Section
Parameter                  Value                 Parameter                   Value
Dependent Variable         Gain                  Rows Processed              144
Number Ind. Variables      7                     Rows Filtered Out           0
Weight Variable            None                  Rows with X's Missing       0
R2                         0.9854                Rows with Weight Missing    0
Adj R2                     0.9847                Rows with Y Missing         0
Coefficient of Variation   0.1496                Rows Used in Estimation     144
Mean Square Error          38.70051              Sum of Weights              144.000
Square Root of MSE         6.220973              Completion Status           Normal Completion
Ave Abs Pct Error          47.269

Regression Equation Section
              Regression     Standard      T-Value                  Reject     Power
Independent   Coefficient    Error         to test       Prob       H0 at      of Test
Variable      b(i)           Sb(i)         H0:B(i)=0     Level      5%?        at 5%
Intercept     11.5547        2.5123        4.599         0.0000     Yes        0.9954
(Group="DOSE20")
              -5.1942        2.7863        -1.864        0.0645     No         0.4567
(Group="DOSE40")
              -35.5054       2.7570        -12.878       0.0000     Yes        1.0000
Pre           -2.0806        0.2045        -10.173       0.0000     Yes        1.0000
Propensity    0.7301         0.0818        8.924         0.0000     Yes        1.0000
(Group="DOSE20")*Pre
              0.6312         0.0708        8.915         0.0000     Yes        1.0000
(Group="DOSE40")*Pre
              3.2646         0.0730        44.724        0.0000     Yes        1.0000
Pre*Pre       0.0241         0.0019        12.591        0.0000     Yes        1.0000

Estimated Model
11.5547265061703-5.19417863950441*(Group="DOSE20")-35.5053879443792*(Group="DOSE40")-
2.08055379387359*Pre+ .730124848180777*Propensity+ .631245843237725*(Group="DOSE20")*Pre+
3.26462104811019*(Group="DOSE40")*Pre+ .024058147508664*Pre*Pre
```

```
Regression Coefficient Section
Independent   Regression          Standard        Lower        Upper      Standardized
Variable       Coefficient           Error     95% C.L.     95% C.L.       Coefficient
Intercept          11.5547          2.5123       6.5866      16.5229            0.0000
(Group="DOSE20")
                   -5.1942          2.7863     -10.7043       0.3159           -0.0489
(Group="DOSE40")
                  -35.5054          2.7570     -40.9575     -30.0532           -0.3345
Pre                -2.0806          0.2045      -2.4850      -1.6761           -0.7314
Propensity          0.7301          0.0818       0.5683       0.8919            0.3453
(Group="DOSE20")*Pre
                    0.6312          0.0708       0.4912       0.7713            0.2465
(Group="DOSE40")*Pre
                    3.2646          0.0730       3.1203       3.4090            1.1848
Pre*Pre             0.0241          0.0019       0.0203       0.0278            0.6285

Note that several residual plots are output, but not shown here.
```

This concludes the regression analysis. We have estimated a regression equation that contains only seven IV's, yet accounts for over 98% of the variability in the Gain score.

Note that the interpretation of the regression coefficients is difficult because of the inclusion of the Group*Pre interaction term. For example, the equation seems to indicate that the Gain is reduced by 5.1942 for the Dose20 group as compared to the Control group. However, the (Group=DOSE2)*Pre regression coefficient of 0.6312 will more than offset this value for most subjects because typical pretest values are greater than 10. That is, 10*0.6312 = 6.312 which is greater than 5.1942.

For example, a subject in the Dose20 group with a pretest score of 50 has an estimated gain score which is 26.3658 = -5.1942+0.6312(50) higher than a similar subject in the Control group.

As a final note, you may wish to adjust the structure of the Group variable. If you wanted to change the reference value to *DOSE40* rather than the default of *CONTROL*, you would change the Default Reference Value on the Variables tab to *Last Value after Sorting* or the X's: Categorical Independent Variables box from *Group* to *Group(DOSE40)* and rerun the analysis. This would yield the following table (you can generate this table by loading the completed template **Example7-4** from the Template tab of the Multiple Regression window).

Standard Regression Output

```
Regression Equation Section
            Regression     Standard      T-Value                  Reject      Power
Independent Coefficient    Error         to test       Prob       H0 at       of Test
Variable    b(i)           Sb(i)         H0:B(i)=0     Level      5%?         at 5%
Intercept   -23.9507       2.6621        -8.997        0.0000     Yes         1.0000
(Group="CONTROL")
            35.5054        2.7570        12.878        0.0000     Yes         1.0000
(Group="DOSE20")
            30.3112        2.8590        10.602        0.0000     Yes         1.0000
Pre         1.1841         0.2087        5.674         0.0000     Yes         0.9999
Propensity  0.7301         0.0818        8.924         0.0000     Yes         1.0000
(Group="CONTROL")*Pre
            -3.2646        0.0730        -44.724       0.0000     Yes         1.0000
(Group="DOSE20")*Pre
            -2.6334        0.0753        -34.989       0.0000     Yes         1.0000
Pre*Pre     0.0241         0.0019        12.591        0.0000     Yes         1.0000

Estimated Model
-23.9506614382099+ 35.5053879443799*(Group="CONTROL")+ 30.3112093048753*(Group="DOSE20")+
1.1840672542367*Pre+ .730124848180749*Propensity-3.2646210481102*(Group="CONTROL")*Pre-
2.63337520487247*(Group="DOSE20")*Pre+ 2.40581475086632E-02*Pre*Pre
```

Chapter 306

Multiple Regression with Serial Correlation

Introduction

The regular Multiple Regression routine assumes that the random-error components are independent from one observation to the next. However, this assumption is often not appropriate for business and economic data. Instead, it is more appropriate to assume that the error terms are positively correlated over time. These are called *autocorrelated* or *serially correlated* data.

Consequences of the error terms being serially correlated include inefficient estimation of the regression coefficients, under estimation of the error variance (MSE), under estimation of the variance of the regression coefficients, and inaccurate confidence intervals.

The presence of serial correlation can be detected by the Durbin-Watson test and by plotting the residuals against their lags.

Autoregressive Error Model

When serial correlation is detected, there are several remedies. Since autocorrelation is often caused by leaving important independent variables out of the regression model, an obvious remedy is to add other, appropriate independent variables to the model. When this is not possible, another remedy is to use an autoregressive model. The usual multiple regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + \varepsilon_t$$

is modified by adding the equation

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

where

$|\rho| < 1$ is the serial correlation

$u_t \sim N\left(0, \sigma^2\right)$

The subscript *t* represents the time period. In econometric work, these *u's* are often called the *disturbances*. They are the ultimate error terms. Further details on this model can be found in chapter 12 of Neter, Kutner, Nachtsheim, and Wasserman (1996).

Cochrane-Orcutt Procedure

Several methods have been suggested to estimate the autoregressive error model. We have adopted the Cochrane-Orcutt procedure as given in Neter, Kutner, Nachtsheim, and Wasserman (1996). This is an iterative procedure that involves several steps.

1. *Ordinary least squares.* The regression coefficients are estimated using ordinary least squares. The array of residuals is calculated.
2. *Estimation of* ρ. The serial correlation is estimated from the current residuals $\left(e_t = Y_t - \hat{Y}_t\right)$ using the formula

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$$

3. *Obtain transformed data.* A new set of data is created using the formulas.

$$\begin{aligned} Y_t' &= Y_t - \hat{\rho} Y_{t-1} \\ X_{1t}' &= X_{1t} - \hat{\rho} X_{1,t-1} \\ &\vdots \\ X_{pt}' &= X_{pt} - \hat{\rho} X_{p,t-1} \end{aligned}$$

4. *Fit model to transformed data*. Ordinary least squares is used to fit the following multiple regression to the transformed data.

$$Y_t' = b_0' + b_1' X_{1t} + b_2' X_{2t} + \cdots + b_p' X_{pt}$$

5. *Create the regression model for the untransformed data*. The regression equation of the untransformed data is created using the following equations.

$$\begin{aligned} b_0 &= \frac{b_0'}{1-\hat{\rho}} \\ b_1 &= b_1' \\ b_2 &= b_2' \\ &\vdots \\ b_p &= b_p' \end{aligned}$$

The estimated standard errors of the regression coefficients are given by

$$s(b_0) = \frac{s(b_0')}{1-\hat{\rho}}$$
$$s(b_1) = s(b_1')$$
$$s(b_2) = s(b_2')$$
$$\vdots$$
$$s(b_p) = s(b_p')$$

6. *Iterate until convergence is reached*. Steps 2 – 4 are then repeated until the value of P stabilizes. Usually, only four or five iterations are necessary.

7. *Calculate Durbin-Watson test on transformed residuals*. As a final diagnostic check, the Durbin-Watson test may be run on the residuals $\left(e_t' = Y_t' - \hat{Y}_t'\right)$ from the transformed regression model.

Durbin-Watson Test

The Durbin-Watson test is often used to test for positive or negative, first-order, serial correlation. It is calculated as follows

$$DW = \frac{\sum_{j=2}^{N}\left(e_j - e_{j-1}\right)^2}{\sum_{j=1}^{N} e_j^2}$$

The distribution of this test is difficult because it involves the *X* values. Originally, Durbin-Watson (1950, 1951) gave a pair of bounds to be used. However, there is a large range of ‘inclusion’ found when using these bounds. Instead of using these bounds, we calculate the exact probability using the beta distribution approximation suggested by Durbin-Watson (1951). This approximation has been shown to be accurate to three decimal places in most cases which is all that are needed for practical work.

Forecasts

The predicted value for a specific set of independent variable values is given by

$$\hat{Y}_t = \hat{b}_0 + \hat{b}_1 X_{1t} + \hat{b}_2 X_{2t} + \cdots + \hat{b}_p X_{pt}$$

For forecasts *j* periods into the future after the end of the series (period *n* is the final period on which we have data), the formula is

$$F_{n+j} = \hat{b}_0 + \hat{b}_1 X_{1,n+j} + \hat{b}_2 X_{2,n+j} + \cdots + \hat{b}_p X_{p,n+j} + \hat{\rho}^j e_n$$

where e_n is the residual from the final observation. That is,

$$e_n = Y_n - \hat{Y}_n$$

The approximate $1-\alpha$ prediction interval for this forecast is

$$F_{n+j} \pm t_{1-\alpha/2,n-3} s_F$$

where s_F is the standard error of the prediction interval based on the transformed data.

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown below. These data give the annual values for several economic statistics. Later in this chapter, these data will be used in an example in which Housing is forecast from Mort5Yr and DispInc. These data are stored in a dataset called HOUSING.S0. Note that only two decimal places are displayed here, while on the database, more decimal places are stored.

HOUSING dataset (subset)

Year	Housing	Mort5Yr	DispInc	TBill	Unemp_rt
1981	403.34	18.25	27006.90	17.72	7.57
1982	407.92	17.93	26896.58	13.66	10.97
1983	446.87	13.17	26582.63	9.31	11.94
1984	457.22	13.54	27662.41	11.06	11.30
1985	485.25	12.08	28710.13	9.43	10.65
1986	475.87	11.17	29057.02	8.97	9.64
1987	491.30	11.12	29626.58	8.15	8.82
1988	493.23	11.61	31070.52	9.48	7.75
1989	487.14	12.01	32417.38	12.05	7.55
1990	491.00	13.31	32683.10	12.81	8.12
1991	512.39	11.07	31980.30	8.73	10.32
1992	523.07	9.50	32224.67	6.59	11.16
1993	533.20	8.76	32412.84	4.84	11.36
1994	497.75	9.53	32789.41	5.54	10.36
1995	502.59	9.14	33242.99	6.89	9.45
1996	522.73	7.91	33256.65	4.21	9.64
1997	538.72	7.05	33839.28	3.26	9.10
1998	533.61	6.92	34915.04	4.73	8.29
1999	531.89	7.54	35971.46	4.72	7.57
2000	528.09	8.32	37566.34	5.49	6.81
2001	544.91	7.38	38228.92	3.77	7.20
2002	547.70	6.99	38806.22	2.59	7.66
2003	561.19	6.36	38896.05	2.87	7.63
2004	581.54	5.38	39870.12	2.30	7.45
2005		6.00	41000.00		
2006		6.25	42000.00		

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable(s)

This option specifies one or more dependent (*Y*) variables. If more than one variable is specified, a separate analysis is run for each.

Numeric Independent Variables

X's: Numeric Independent Variable(s)

Specify any numeric independent variables in this box. Numeric variables are those whose values are numeric and are at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are more appropriately analyzed when you specify them as Categorical Independent Variables.

If you want to create powers and cross-products of these variables, specify an appropriate model in the 'Custom Model' field under the Model tab.

If you want to create predicted values of *Y* for values of *X* not in your database, add the *X* values to the bottom of the database. These rows will not be used during estimation phase, but predicted values will be generated for them on the reports.

Categorical Independent Variables

X's: Categorical Independent Variable(s)

Specify categorical (nominal) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories. For further details about categorical variables, see the discussion on this topic in the Multiple Regression chapter.

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. For further details about this value, see the discussion on this topic in the Multiple Regression chapter.

Default Contrast Type

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them.

This option allows you to specify what type of numeric variable will be created. For further details about this option, see the discussion on this topic in the Multiple Regression chapter.

Estimation Options

Maximum Cochrane-Orcutt Iterations

This is the maximum number of iterations that the procedure will cycle through. Some authors recommend only one iteration. Others recommend stopping once the Durbin-Watson test is not significant. This option lets you stop after a specific number of iterations. Usually, four or five iterations should be plenty.

Minimum Rho Change

If the change is rho (serial correlation) from one iteration to the next is less than this amount, the algorithm will stop iterating. We suggest you use a small amount such as 0.00001.

Alpha Levels

Alpha of C.I.'s and Tests

The value of alpha for the statistical tests and confidence intervals is specified here. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your particular study.

Alpha of Assumptions

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests use a larger alpha such as 0.10, 0.15, or 0.20.

We recommend 0.20.

Model Tab

These options control the regression model.

Model Specification

Which Model Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a time-series regression model, select *Up to 1-Way*.

The other options on this tab are covered in detail in the Multiple Regression chapter. We refer you to that chapter for further details.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Run Summary ... Residuals

Each of these options specifies whether the indicated report is calculated and displayed. Note that since some of these reports provide results for each row, they may be too long for normal use when requested on large databases.

Select Plots

Histogram ... Residuals vs X Plot

Indicate whether to display these plots.

Report Options

Show All Rows

This option makes it possible to display predicted values for only a few designated rows.

When checked predicted values, residuals, and other row-by-row statistics, will be displayed for all rows used in the analysis.

When not checked, predicted values and other row-by-row statistics will be displayed for only those rows in which the dependent variable's value is missing.

Format Tab

These options specify the number of decimal places shown when the indicated value is displayed in a report. The number of decimal places shown in plots is controlled by the Tick Label Settings buttons on the Axes tabs.

Report Options

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Skip Line After

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Report Options – Decimal Places

Probability ... Mean Square Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'All' to display all digits available. The number of digits displayed by this option is controlled by whether the PRECISION option is SINGLE or DOUBLE.

Plot Options Tab

These options control the titles and style files used on each of the plots.

Plot Titles and Style Files

Plot Titles

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Plot Style Files

Designate various plot style files. These files set all plot options that are not set directly by this procedure. Unless you choose otherwise, the default style file (Default) is used. These files are created in the various graphics procedures, depending on the plot type.

Plotting Symbol

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Histogram Options

Number of Bars

Specify the number of intervals, bins, or bars used in the histogram. Select ‘0 - Automatic’ to have the program select an appropriate number based on the number of residuals.

Axes Tabs

The options on these panels control the appearance of the *X* variables, *Y* variable, residuals, RStudent, Hat Diagonal, Rows Numbers, Counts, and Expected axes whenever they are included on a plot. This makes it easy to give a consistent look to all of your plots without modifying them individually.

Y-Variable ... Expected Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by the names of the corresponding variables. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the axis associated with this variable. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the associated axis.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on the associated axes.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful.

Data Storage Options – Select Items to Store

Predicted Y ... Upper C.L. Individual

Indicate whether to store these row-by-row values, beginning at the variable indicated by the *Store First Variable In* option.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Generating Forecasts (All Reports)

This section presents an example of how to generate forecasts for housing data that was presented earlier in this chapter. This data is stored in the HOUSING.S0 database. We suggest that you open it now.

This example will run an adjusted multiple regression of *Housing* on *Mort5Yr* and *DispInc*. The adjustment will use the Cochrane-Orcutt procedure. The data for housing ends in 2004. Forecasts will be generated for the years 2005 and 2006.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Multiple Regression with Serial Correlation window.

1 Open the Housing dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** folder of your NCSS folder.
- Click on the file **Housing.s0**.
- Click **Open**.

2 **Open the Multiple Regression with Serial Correlation window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines**, then **Multiple Regression with Serial Correlation**. The Multiple Regression with Serial Correlation procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Multiple Regression with Serial Correlation window, select the **Variables tab**.
- Set the **Y: Dependent Variables** box to **Housing**.
- Set the **X's: Numeric Independent Variables** box to **Mort5Yr - DispInc**.
- Set the **Maximum Cochrane-Orcutt Iterations** to **1**.

4 **Specify the reports.**

- Select the **Reports tab**.
- Make sure all reports and plots are checked.

5 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Section

```
Parameter                Value           Parameter                   Value
Dependent Variable       Housing         Rows Processed              32
Number Ind. Variables    2               Rows Filtered Out           0
Weight Variable          None            Rows with X's Missing       0
R2                       0.8941          Rows with Weight Missing    0
Adj R2                   0.8860          Rows with Y Missing         2
Coefficient of Variation 0.0361          Rows Used in Estimation     30
Mean Square Error        77.15598        Sum of Weights              29.000
Square Root of MSE       8.783848        Completion Status           Normal Completion
Ave Abs Pct Error        1.801           Autocorrelation (Rho)       0.5121
```

This report summarizes the multiple regression results. It presents the variables used, the number of rows used, and the basic results. The estimated value of the autocorrelation (rho) has been added to this report. Otherwise, it is identical to the corresponding report in the regular Multiple Regression report.

Note that values such as R2, Mean Square Error, etc., are calculated on the transformed data.

Descriptive Statistics Section

```
                                        Standard
Variable      Count    Mean        Deviation     Minimum     Maximum
DispInc       29       31000.94    5157.438      21780.1     39870.13
Mort5Yr       29       10.53919    3.183494      5.380194    18.25095
Housing       29       491.0214    48.53722      403.3378    581.5398
```

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

	DispInc	Mort5Yr	Housing
DispInc	1.0000	-0.5962	0.7913
Mort5Yr	-0.5962	1.0000	-0.8874
Housing	0.7913	-0.8874	1.0000

Pearson correlations are given for all variables.

Regression Equation Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0:B(i)=0	Prob Level	Reject H0 at 5%?
Intercept	445.1365	35.4690	12.550	0.0000	Yes
DispInc	0.0044	0.0009	5.118	0.0000	Yes
Mort5Yr	-8.5371	1.0522	-8.113	0.0000	Yes

Estimated Model

445.136489079996+ 4.4434007069797E-03*DispInc-8.53714263704248*Mort5Yr

This section reports the values and significance tests of the regression coefficients. Note that the intercept has been corrected by dividing by 1-rho. Other than this, the report has the same definitions as in regular Multiple Regression.

Regression Coefficient Section

Independent Variable	Regression Coefficient	Standard Error	Lower 95% C.L.	Upper 95% C.L.	Standardized Coefficient
Intercept	445.1365	35.4690	372.2289	518.0441	0.0000
DispInc	0.0044	0.0009	0.0027	0.0062	0.4068
Mort5Yr	-8.5371	1.0522	-10.7000	-6.3743	-0.6449

Note: The T-Value used to calculate these confidence limits was 2.056.

The report has the same definitions as in regular Multiple Regression.

Analysis of Variance Section

Source	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		1720724	1720724			
Model	2	0.8941	16943.61	8471.806	109.801	0.0000	
Error	26	0.1059	2006.055	77.15598			
Total(Adjusted)	28	1.0000	18949.67	676.7738			

This section reports the analysis of variance table. Note it was calculated from the transformed data on the last iteration. Other than this, the report has the same definitions as in regular Multiple Regression.

Serial-Correlation and Durbin-Watson Test

```
Serial Correlation of Residuals from Uncorrected Model
             Serial                 Serial                 Serial
Lag     Correlation      Lag   Correlation      Lag   Correlation
1            0.5090        9       -0.4075       17       -0.1140
2            0.1980       10       -0.5085       18       -0.0147
3            0.0802       11       -0.3018       19        0.1512
4            0.0505       12       -0.1962       20        0.1290
5            0.2072       13       -0.1042       21        0.0519
6            0.2165       14       -0.1067       22        0.0275
7           -0.0649       15       -0.3178       23        0.0457
8           -0.0979       16       -0.2177       24        0.0875
Above serial correlations significant if their absolute values are greater than 0.365148

Serial Correlation of Residuals from Corrected Model
             Serial                 Serial                 Serial
Lag     Correlation      Lag   Correlation      Lag   Correlation
1            0.0261        9       -0.2371       17        0.0817
2           -0.0349       10       -0.3626       18        0.0420
3            0.0972       11       -0.0584       19        0.0314
4           -0.1182       12        0.0042       20        0.0473
5            0.1002       13       -0.0671       21        0.0248
6            0.2071       14       -0.0042       22        0.0761
7           -0.3095       15       -0.2443       23        0.0038
8            0.1301       16        0.0617       24        0.0388
Above serial correlations significant if their absolute values are greater than 0.371391

Durbin-Watson Test For Serial Correlation of Uncorrected Model
                                                      Did the Test Reject
Parameter                                    Value      H0: Rho(1) = 0?
Durbin-Watson Value                         0.9234
Prob. Level: Positive Serial Correlation    0.0002                  Yes
Prob. Level: Negative Serial Correlation    0.9974                   No

Durbin-Watson Test For Serial Correlation of Corrected Model
                                                      Did the Test Reject
Parameter                                    Value      H0: Rho(1) = 0?
Durbin-Watson Value                         1.9221
Prob. Level: Positive Serial Correlation    0.3273                   No
Prob. Level: Negative Serial Correlation    0.4923                   No
```

This section reports the autocorrelation structure of the residuals both before and after the model is corrected for serial correlation. It has the same definitions as in the regular Multiple Regression report.

Predicted Values with Confidence Limits of Means

Row	Actual Housing	Predicted Housing	Standard Error Of Predicted	95% Lower Conf. Limit Of Mean	95% Upper Conf. Limit Of Mean
1	420.722	445.738			
2	431.522	447.504	3.273	440.776	454.232
3	448.085	462.874	4.140	454.364	471.384
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
26	528.086	541.005	3.211	534.406	547.605
27	544.913	552.012	2.883	546.086	557.938
28	547.703	557.854	3.010	551.667	564.041
29	561.186	563.635	2.950	557.571	569.698
30	581.540	576.364	3.536	569.095	583.633
31		578.744	3.901	570.725	586.762
32		579.760	4.192	571.143	588.377

Confidence intervals for the mean response of Y given specific levels for the IV's are provided here.

Predicted Values with Prediction Limits of Individuals

Row	Actual Housing	Predicted Housing	Standard Error Of Predicted	95% Lower Pred. Limit Of Individual	95% Upper Pred. Limit Of Individual
1	420.722	445.738			
2	431.522	447.504	9.374	428.235	466.772
3	448.085	462.874	9.711	442.914	482.835
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
26	528.086	541.005	9.352	521.782	560.229
27	544.913	552.012	9.245	533.009	571.015
28	547.703	557.854	9.285	538.768	576.940
29	561.186	563.635	9.266	544.588	582.681
30	581.540	576.364	9.469	556.900	595.828
31		578.744	9.611	558.988	598.500
32		579.760	9.733	559.753	599.766

A prediction interval for the individual response of *Y* given specific values of the IV's is provided here for each row. Note that the forecasts start where the actual housing values are blank.

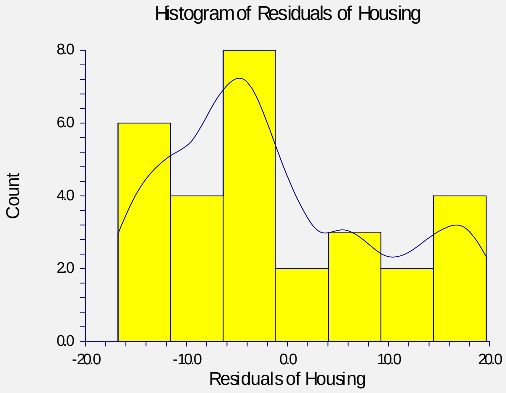
Residual Report

```
                                                            Absolute
            Actual     Predicted                            Percent
Row        Housing       Housing           Residual           Error
1          420.722       445.738
2          431.522       447.504            -15.982           3.704
3          448.085       462.874            -14.789           3.300
4          447.923       464.696            -16.773           3.745
5          451.401       454.340             -2.939           0.651
6          432.474       438.269             -5.795           1.340
7          403.338       409.328             -5.990           1.485
8          407.922       411.549             -3.627           0.889
.                .             .                  .               .
.                .             .                  .               .
.                .             .                  .               .
```

This section reports on the sample residuals, or e_i’s.

Histogram

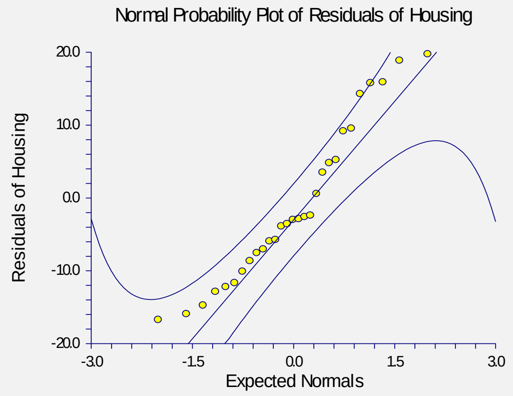
The purpose of the histogram and density trace of the residuals is to evaluate whether they are normally distributed. A dot plot is also given that highlights the distribution of points in each bin of the histogram. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating normality of the residuals. The better choice would be the normal probability plot.



Probability Plot of Residuals

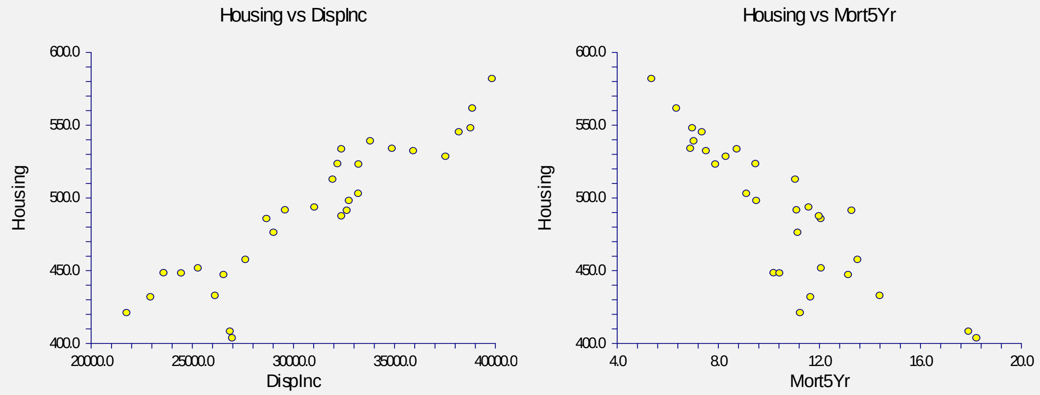
If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.

If the residuals are not normally distributed, then the t-tests on regression coefficients, the F-tests, and any interval estimates are not valid. This is a critical assumption to check.



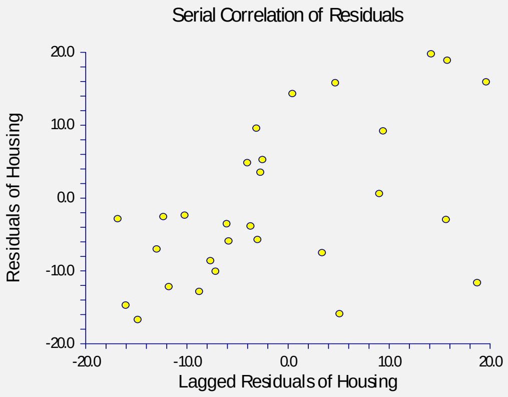
Plots of Y versus each IV

Actually, a regression analysis should always begin with a plot of Y versus each IV. These plots often show outliers, curvilinear relationships, and other anomalies.



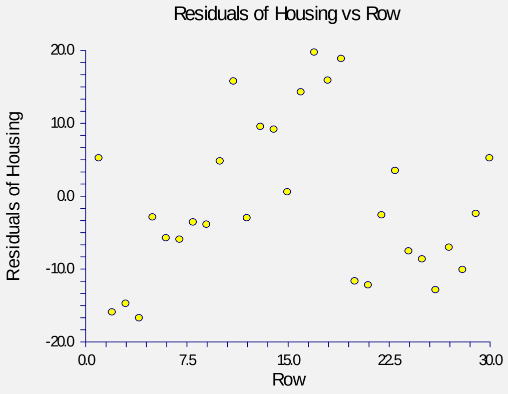
Serial Correlation of Residuals Plot

This is a scatter plot of the *j*th residual versus the *j*th-1 residual. The purpose of this plot is to check for first-order autocorrelation. Positive autocorrelation or serial correlation means that the residual in time period *j* tends to have the same sign as the residual in time period (*j*-1). On the other hand, a strong negative autocorrelation means that the residual in time period *j* tends to have the opposite sign as the residual in time period (*j*-1).



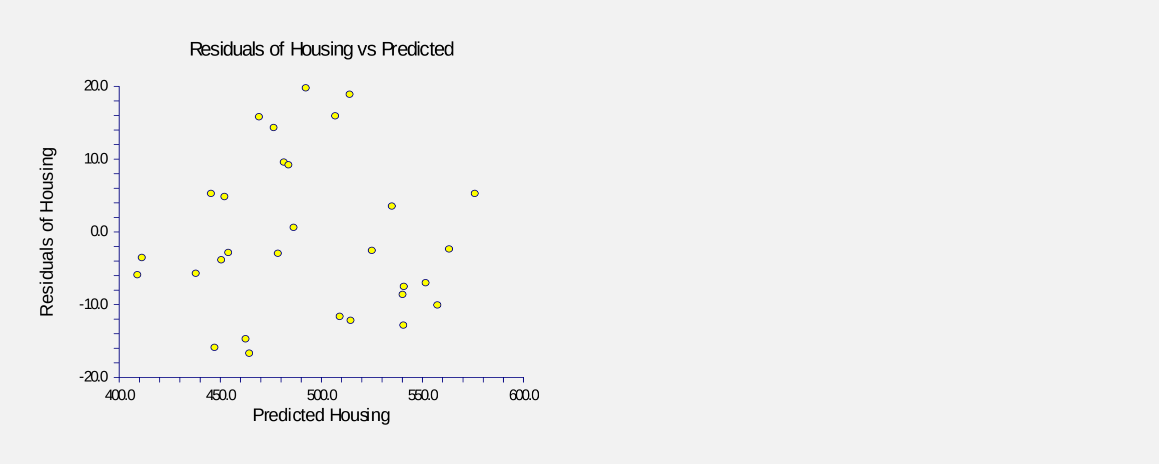
Sequence Plot

Sequence plots may be useful in finding variables that are not accounted for by the regression equation. They are especially useful if the data were taken over time.



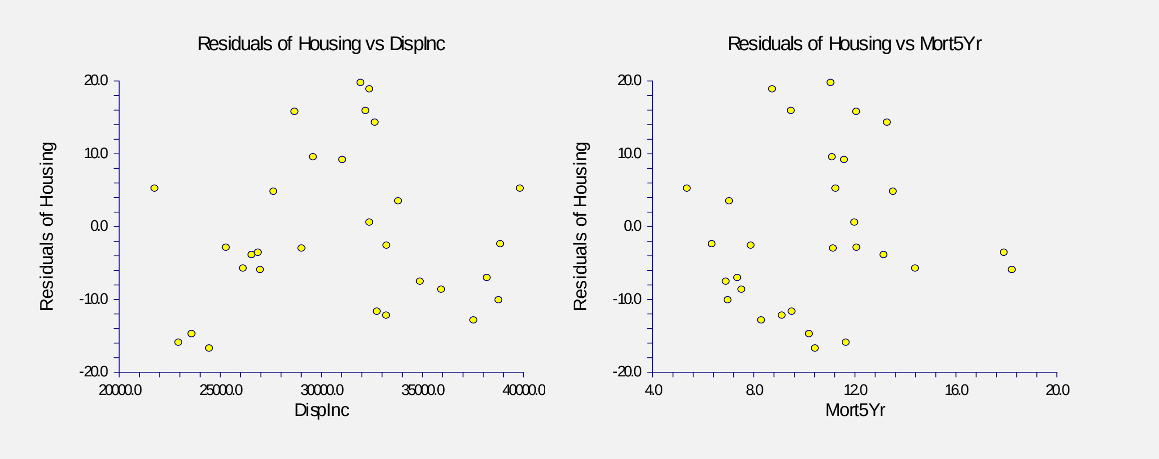
Residual vs Predicted Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance, a violation of a critical regression assumption. The sloping or curved band signifies inadequate specification of the model. The sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.



Residual vs Predictor(s) Plot

This is a scatter plot of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.



Chapter 310

Variable Selection for Multivariate Regression

Introduction

Often theory and experience give only general direction as to which of a pool of candidate variables should be included in the regression model. The actual set of predictor variables used in the final regression model must be determined by analysis of the data. Determining this subset is called the variable selection problem.

Finding this subset of regressor (independent) variables involves two opposing objectives. First, we want the regression model to be as complete and realistic as possible. We want every regressor that is even remotely related to the dependent variable to be included. Second, we want to include as few variables as possible because each irrelevant regressor decreases the precision of the estimated coefficients and predicted values. Also, the presence of extra variables increases the complexity of data collection and model maintenance. The goal of variable selection becomes one of parsimony: achieve a balance between simplicity (as few regressors as possible) and fit (as many regressors as needed).

There are many different strategies for selecting variables for a regression model. If there are no more than fifteen candidate variables, the All Possible Regressions procedure should be used since it will always give as good or better models than the stepping procedures available in this procedure. On the other hand, when there are more than fifteen candidate variables, the search procedure contained in this procedure is an excellent choice.

While studying at Texas A&M University, Dr. Claude McHenry (1978) developed a heuristic algorithm that usually yields the same subset as the all possible regressions routine, but with a lot less work. The algorithm is of a more general nature than the other variable selection procedures in ***NCSS*** because it allows more than one dependent variable to be studied. Hence, it is useful for variable selection in multivariate multiple regression and in discriminant analysis.

McHenry’s Select Algorithm

The algorithm seeks a subset that provides a maximum value of R-Squared (or a minimum Wilks’ lambda in the multivariate case). The algorithm first finds the best single variable. To find the best pair of variables, it tries each of the remaining variables and selects the one that adds the most. It then omits the first variable and determines if any other variable would add more. If a better variable is found, it is kept and the worst variable is removed. Another search is now made through the remaining variables. This switching process continues until no switching will result in a better subset.

Once the optimal pair of variables is found, the best three variables is searched for in much the same manner. First, the best third variable is found to add to the optimal pair of variables from the last step. Next, each of the first two variables is omitted and another, even better, variable is searched for. The algorithm continues until no switching improves R-Squared.

This algorithm is extremely fast. It seems to find the best (or very near best) subset in most situations. An interesting feature is the ability to specify more than one dependent variable. This is useful in discriminant analysis where each group may be considered as a binary (0, 1) variable. It is also useful when you want to predict several dependent variables using a minimum number of independent variables.

Assumptions and Limitations

The same assumptions and qualifications apply here as applied to multiple regression. We refer you to the Assumptions section in the Multiple Regression chapter for a discussion of these assumptions. We will here mention a couple of restrictions necessary for this algorithm to work.

Number of Observations

The number of observations must be at least one greater than the number of candidate regressors. A popular rule-of-thumb when using any variable selection procedure is that you have at least five observations for each candidate variable.

No Linear Dependencies

This algorithm begins by fitting the full model with all candidate variables. In order to solve this full model, no linear dependencies may exist in the data. A linear dependency occurs when one variable is a weighted average of the rest. For example, if one variable is the total of several others, it cannot be included in the candidate pool.

This same restriction applies to the set of dependent variables.

Using This Procedure

This procedure performs one portion of a regression analysis: it obtains a set of independent variables from a pool of candidate variables. Once the set of variables is obtained, you should proceed to the Multiple Regression procedure to estimate the regression coefficients, study the residuals, and so on.

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown below. These data are contained in the SAMPLE database.

SAMPLE dataset

Test1	Test2	Test3	Test4	Test5	IQ
83	34	65	63	64	106
73	19	73	48	82	92
54	81	82	65	73	102
96	72	91	88	94	121
84	53	72	68	82	102
86	72	63	79	57	105
76	62	64	69	64	97
54	49	43	52	84	92
37	43	92	39	72	94
42	54	96	48	83	112
71	63	52	69	42	130
63	74	74	71	91	115
69	81	82	75	54	98
81	89	64	85	62	96
50	75	72	64	45	103

Missing Values

Rows with missing values in the variable pool are ignored. This may cause differences in the results between this procedure and regression analysis. Suppose that through the selection process, none of the variables with missing values end up in the final subset. When a regression analysis is run on the subset, the rows with missing values will not be deleted (since those variables are no longer active). This will obviously change the estimated values.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Dependent Variables

Y's: Dependent Variable

Specify one or more dependent (Y) variables. If more than one variable is specified, the analysis becomes *multivariate* multiple regression subset selection. The best subset for fitting the block of dependent variables is found.

Independent Variables

X's: Independent Variables

Specify the independent (X) variables.

Forced X Variables

List any independent variables that are to be forced into the model. These variables will be kept in the model, even if they are not significant.

Model Selection

Maximum Variables

The largest subset that you would like to find. Under normal conditions, a maximum subset size of ten is reasonable.

Reports Tab

The following options control which reports and plots are displayed.

Select Additional Reports

Coded Report and Long Variable Names Report

Specifies whether these reports are displayed.

Select Plots

R-Squared Plot

Specifies whether to show the R-Squared Plot.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you specify whether to display only variable names, variable labels, or both.

R-Squared Plot Tab

A scatter plot comparing the value of R-Squared to the subset size is available to help you select the appropriate model size.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Variable Selection Analysis

This section presents an example of how to run a variable selection analysis of the data contained in the SAMPLE database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Variable Selection for Multivariate Regression window.

1 Open the SAMPLE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 Open the Variable Selection for Multivariate Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Variable Selection Routines**, then **Variable Selection for Multivariate Regression**. The Variable Selection for Multivariate Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Variable Selection for Multivariate Regression window, select the **Variables tab**.
- Double-click in the **Y's: Dependent Variables** text box. This will bring up the variable selection window.
- Select **IQ** from the list of variables and then click **Ok**. "IQ" will appear in the Y's: Dependent Variable box.
- Double-click in the **X's: Independent Variables** text box. This will bring up the variable selection window.

- Select **Test1** through **Test5** from the list of variables and then click **Ok**. "Test1-Test5" will appear in the X's: Independent Variables.
- Enter **5** in the **Alternative Models** box.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Code Cross-Reference Section

```
Code Cross-Reference Section

Code            Variable            Count         Mean
                IQ                  15            104.3333
A               Test1               15            67.93333
B               Test2               15            61.4
C               Test3               15            72.33334
D               Test4               15            65.53333
E               Test5               15            69.93333
```

The code, count, and mean are displayed for each variable. This report is particularly useful for checking that the correct variables were selected. The letters A to Z are assigned to each of the independent variables involved in the regression model. These are used to specify which variables are in each subset.

Selection Results Section

```
Selection Results Section

Model                    R-Squared     Coded
Size        R-Squared    Change        Variables
1           0.137941     0.137941      D
2           0.154246     0.016305      CD
3           0.383854     0.229608      ABD
4           0.396353     0.012499      ABCD
5           0.399068     0.002715      ABCDE
```

This report presents the results of the search procedure. The model for each subset (model) size is presented. To use this report, you scan down the R-Squared values, looking for the subset size where R-Squared stabilizes. In this example, the R-Squared value for the best three-variable model is 0.383854, and the R-Squared for the best four-variable model is 0.396353. This is a minor increase. We would select the three-variable model as our final model.

If more the one dependent variable is specified, the R-Squared column will be replaced by a Wilks' Lambda column.

Model Size

This is the number of independent variables in the model.

R-Squared

This is the value of R-Squared achieved for this subset. Note that if multiple dependent variables are specified, this column will be labeled Wilks' Lambda. Wilks' Lambda is the multivariate extension of R-Squared. It behaves like 1-(R-Squared). Hence, when you have multiple dependent variables, you look for a value close to zero, rather than close to one as you do with R-Squared.

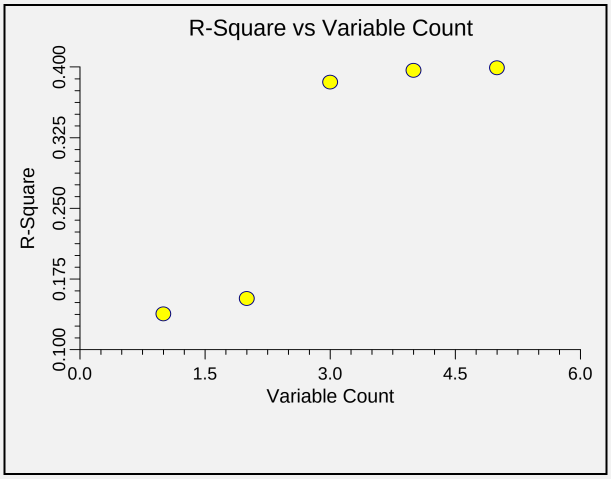
R-Squared Change

This is the amount that is added to R-Squared (or Wilks' Lambda) when an additional variable is added to the model.

Coded Variables

This is a list of the variables in the model. The variable which each letter represents is listed in the Code Cross-Reference Section.

R-Squared vs Variable Count Plot



This plot displays the values of R-Squared on the vertical axis and the subset size on the horizontal axis for the data displayed in the Selection Results Section, above. Note the large jump between the two-variable model and the three-variable model. We quickly see that the four and five variable models do not do much better. Hence, our conclusion is to use the three-variable model. The three-variable model is ABD, which translates to variables Test1, Test2, and Test4.

Chapter 311

Stepwise Regression

Introduction

Often, theory and experience give only general direction as to which of a pool of candidate variables (including transformed variables) should be included in the regression model. The actual set of predictor variables used in the final regression model must be determined by analysis of the data. Determining this subset is called the *variable selection* problem.

Finding this subset of regressor (independent) variables involves two opposing objectives. First, we want the regression model to be as complete and realistic as possible. We want every regressor that is even remotely related to the dependent variable to be included. Second, we want to include as few variables as possible because each irrelevant regressor decreases the precision of the estimated coefficients and predicted values. Also, the presence of extra variables increases the complexity of data collection and model maintenance. The goal of variable selection becomes one of parsimony: achieve a balance between simplicity (as few regressors as possible) and fit (as many regressors as needed).

There are many different strategies for selecting variables for a regression model. If there are no more than fifteen candidate variables, the *All Possible Regressions* procedure (discussed in the next chapter) should be used since it will always give as good or better models than the stepping procedures available in this procedure. On the other hand, when there are more than fifteen candidate variables, the four search procedures contained in this procedure may be of use.

These search procedures will often find very different models. Outliers and collinearity can cause this. If there is very little correlation among the candidate variables and no outlier problems, the four procedures should find the same model.

We will now briefly discuss each of these procedures.

Variable Selection Procedures

Forward (Step-Up) Selection

This method is often used to provide an initial screening of the candidate variables when a large group of variables exists. For example, suppose you have fifty to one hundred variables to choose from, way outside the realm of the all-possible regressions procedure. A reasonable approach would be to use this forward selection procedure to obtain the best ten to fifteen variables and then apply the all-possible algorithm to the variables in this subset. This procedure is also a good choice when multicollinearity is a problem.

The forward selection method is simple to define. You begin with no candidate variables in the model. Select the variable that has the highest R-Squared. At each step, select the candidate variable that increases R-Squared the most. Stop adding variables when none of the remaining variables are significant. Note that once a variable enters the model, it cannot be deleted.

Backward (Step-Down) Selection

This method is less popular because it begins with a model in which all candidate variables have been included. However, because it works its way down instead of up, you are always retaining a large value of R-Squared. The problem is that the models selected by this procedure may include variables that are not really necessary. The user sets the significance level at which variables can enter the model.

The backward selection model starts with all candidate variables in the model. At each step, the variable that is the least significant is removed. This process continues until no nonsignificant variables remain. The user sets the significance level at which variables can be removed from the model.

Stepwise Selection

Stepwise regression is a combination of the forward and backward selection techniques. It was very popular at one time, but the Multivariate Variable Selection procedure described in a later chapter will always do at least as well and usually better.

Stepwise regression is a modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model.

Stepwise regression requires two significance levels: one for adding variables and one for removing variables. The cutoff probability for adding variables should be less than the cutoff probability for removing variables so that the procedure does not get into an infinite loop.

Min MSE

This procedure is similar to the Stepwise Selection search procedure. However, instead of using probabilities to add and remove, you specify a minimum change in the root mean square error. At each step, the variable whose status change (in or out of the model) will decrease the mean square error the most is selected and its status is reversed. If it is currently in the model, it is removed. If it is not in the model, it is added. This process continues until no variable can be found that will cause a change larger than the user-specified minimum change amount.

Assumptions and Limitations

The same assumptions and qualifications apply here as applied to multiple regression. Note that outliers can have a large impact on these stepping procedures, so you must make some attempt to remove outliers from consideration before applying these methods to your data.

The greatest limitation with these procedures is one of sample size. A good rule of thumb is that you have at least five observations for each variable in the candidate pool. If you have 50 variables, you should have 250 observations. With less data per variable, these search procedures may fit the randomness that is inherent in most datasets and spurious models will be obtained.

This point is critical. To see what can happen when sample sizes are too small, generate a set of random numbers for 20 variables with 30 observations. Run any of these procedures and see what a magnificent value of R-Squared is obtained, even though its theoretical value is zero!

Using This Procedure

This procedure performs one portion of a regression analysis: it obtains a set of independent variables from a pool of candidate variables. Once the subset of variables is obtained, you should proceed to the Multiple Regression procedure to estimate the regression coefficients, study the residuals, and so on.

Data Structure

An example of data appropriate for this procedure is shown in the table below. This data is from a study of the relationships of several variables with a person’s IQ. Fifteen people were studied. Each person’s IQ was recorded along with scores on five different personality tests. The data are contained in the SAMPLE database. We suggest that you open this database now so that you can follow along with the example.

SAMPLE dataset

Test1	Test2	Test3	Test4	Test5	IQ
83	34	65	63	64	106
73	19	73	48	82	92
54	81	82	65	73	102
96	72	91	88	94	121
84	53	72	68	82	102
86	72	63	79	57	105
76	62	64	69	64	97
54	49	43	52	84	92
37	43	92	39	72	94
42	54	96	48	83	112
71	63	52	69	42	130
63	74	74	71	91	115
69	81	82	75	54	98
81	89	64	85	62	96
50	75	72	64	45	103

Missing Values

Rows with missing values in the active variables are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Dependent Variable

Y: Dependent Variable

Specifies a dependent (Y) variable. If more than one variable is specified, a separate analysis is run for each.

Weight Variable

Weight Variable

Specifies a variable containing observation (row) weights for generating weighted regression analysis. These weights might be those saved during a robust regression analysis.

Independent Variables

X's: Independent Variables

Specify the independent (X or candidate) variables.

Model Selection

Selection Method

This option specifies which of the four search procedures should be used: Forward, Backward, Stepwise, or Min MSE.

Prob to Enter

Sometimes call PIN, this is the probability required to enter the equation. This value is used by the Forward and the Stepwise procedures. A variable, not currently in the model, must have a t-test probability value less than or equal to this in order to be considered for entry into the regression equation. You must set PIN < POUT.

Prob to Remove

Sometimes call POUT, this is the probability required to be removed from the equation. This value is used by the Backward and the Stepwise procedures. A variable, currently in the model, must have a t-test probability value greater than this in order to be considered for removal from the regression equation. You must set PIN < POUT.

Min RMSE Change

This value is used by the Minimum MSE procedure to determine when to stop. The procedure stops when the maximum relative decrease in the square root of the mean square error brought about by changing the status of a variable is less than this amount.

Maximum Iterations

This is the maximum number of iterations that will be allowed. This option is useful to prevent the unlimited looping that may occur. You should set this to a high value, say 50 or 100.

Remove Intercept

Unchecked indicates that the intercept term is to be included in the regression. Checked indicates that the intercept should be omitted from the regression model. Note that deleting the intercept distorts most of the diagnostic statistics (R-Squared, etc.).

Reports Tab

These options control the reports that are displayed.

Select Reports

Descriptive Statistics and Selected Variables Reports

This option specifies whether the indicated report is displayed.

Report Options

Report Format

Two output formats are available: brief and verbose. The *Brief* output format consists of a single line for each step (scan through the variables). The *Verbose* output format gives a complete table of each variable’s statistics at each step. If you have many variables, the Verbose option can produce a lot of output.

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Stepwise Regression Analysis

This section presents an example of how to run a stepwise regression analysis of the data presented in the SAMPLE dataset.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Stepwise Regression window.

1 Open the SAMPLE dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 Open the Stepwise Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Variable Selection Routines**, then **Stepwise Regression**. The Stepwise Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Stepwise Regression window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable** text box. This will bring up the variable selection window.
- Select **IQ** from the list of variables and then click **Ok**. "IQ" will appear in the Y: Dependent Variable box.
- Double-click in the **X's: Independent Variables** text box. This will bring up the variable selection window.
- Select **Test1** through **Test5** from the list of variables and then click **Ok**. "Test1-Test5" will appear in the X's: Independent Variables.
- In the **Selection Method** list box, select **Backward**.

4 Specify the reports.

- On the Stepwise Regression window, select the **Reports tab**.
- In the **Report Format** list box, select **Verbose**.
- Check the **Descriptive Statistics** checkbox.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Descriptive Statistics Section

```
Descriptive Statistics Section

Variable        Count    Mean         Standard Deviation
Test1           15       67.93333     17.39239
Test2           15       61.4         19.39735
Test3           15       72.33334     14.73415
Test4           15       65.53333     13.95332
Test5           15       69.93333     16.15314
IQ              15       104.3333     11.0173
```

For each variable, the Count, Mean, and Standard Deviation are calculated. This report is especially useful for making certain that you have selected the right variables and that the appropriate number of rows was used.

Iteration Detail Section (Verbose Version)

```
Iteration Detail Section

Iteration 0:  Unchanged

                            Standard.     R-Squared    R-Squared                  Prob        Pct Change
In          Variable        Coefficient   Increment    Other X's    T-Value       Level       Sqrt(MSE)
Yes         Test1           -3.0524       .2357        .9747        -1.8789       .0930       11.9387
Yes         Test2           -2.9224       .2414        .9717        -1.9016       .0897       12.3210
Yes         Test3           .1404         .0152        .2280        .4773         .6445       -3.9386
Yes         Test4           4.7853        .2832        .9876        2.0596        .0695       15.0741
Yes         Test5           -.0595        .0027        .2329        -.2017        .8447       -4.9176
R-Squared=0.3991            Sqrt(MSE)=10.65198

Iteration 1:  Removed Test5 from equation

                            Standard.     R-Squared    R-Squared                  Prob        Pct Change
In          Variable        Coefficient   Increment    Other X's    T-Value       Level       Sqrt(MSE)
Yes         Test1           -3.0612       .2373        .9747        -1.9825       .0756       12.5340
Yes         Test2           -2.9032       .2392        .9716        -1.9906       .0745       12.6640
Yes         Test3           .1163         .0125        .0752        .4550         .6588       -3.6717
Yes         Test4           4.7850        .2832        .9876        2.1660        .0555       15.5681
No          Test5                         .0027        .2329        .2017         .8447       5.1719
R-Squared=0.3964            Sqrt(MSE)=10.12816

Iteration 2:  Removed Test3 from equation

                            Standard.     R-Squared    R-Squared                  Prob        Pct Change
In          Variable        Coefficient   Increment    Other X's    T-Value       Level       Sqrt(MSE)
Yes         Test1           -3.1020       .2444        .9746        -2.0890       .0607       13.1519
Yes         Test2           -2.9024       .2391        .9716        -2.0659       .0632       12.7977
Yes         Test4           4.7988        .2849        .9876        2.2553        .0455       15.7808
No          Test3                         .0125        .0752        .4550         .6588       3.8116
No          Test5                         .0000        .0810        .0087         .9932       4.8805
R-Squared=.3839             Sqrt(MSE)=9.756291
```

```
Iteration 3:  Unchanged

                    Standard.    R-Squared    R-Squared                 Prob        Pct Change
In       Variable   Coefficient  Increment    Other X's    T-Value      Level       Sqrt(MSE)
Yes      Test1      -3.1020      .2444        .9746        -2.0890      .0607       13.1519
Yes      Test2      -2.9024      .2391        .9716        -2.0659      .0632       12.7977
Yes      Test4      4.7988       .2849        .9876        2.2553       .0455       15.7808
No       Test3                   .0125        .0752        .4550        .6588       3.8116
No       Test5                   .0000        .0810        .0087        .9932       4.8805
R-Squared=.3839     Sqrt(MSE)=9.756291
```

This report presents information about each step of the search procedures. You can scan this report to see if you would have made the same choice. Each report shows the statistics after the specified action (entry or removal) was taken.

For each iteration, there are three possible actions:

1. *Unchanged*. No action was taken because of the scan in this step. Because of the "backward look" in the stepwise search method, this will show up a lot when this method is used. Otherwise, it will usually show up as the first and last steps.
2. *Removal*. A variable was removed from the model.
3. *Entry*. A variable was added to the model.

Individual definitions of the items on the report are as follows:

In

A *Yes* means the variable is in the model. A *No* means it is not.

Variable

This is the name of the candidate variable.

Standard. Coefficient

Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is:

$$b_{j,\,std} = b_j\left(\frac{s_y}{s_{x_j}}\right)$$

where s_y and s_{x_j} are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

R-Squared Increment

This is the amount that R-Squared would be changed if the status of this variable were changed. If the variable is currently in the model, this is the amount the R-Squared value would be decreased if it were removed. If the variable is currently out of the model, this is the amount the overall R-Squared would be increased if it were added. Large values here indicate important independent variables.

You want to add variables that make a large contribution to R-Squared and to delete variables that make a small contribution to R-Squared.

R-Squared Other X’s

This is a collinearity measure, which should be as small as possible. This is the R-Squared value that would result if this independent variable were regressed on all of the other independent variables currently in the model.

T-Value

This is the t-value for testing the hypothesis that this variable should be added to, or deleted from, the model. The test is adjusted for the rest of the variables in the model. The larger this t-value is, the more important the variable.

Prob Level

This is the two-tail p-value for the above t-value. The smaller this p-value, the more important the independent variable is. This is the significance value that is compared to the values of PIN and POUT (see *Stepwise Method* above).

Pct Change Sqrt(MSE)

This is the percentage change in the square root of the mean square error that would occur if the specified variable were added to, or deleted from, the model. This is the value that is used by the *Min MSE* search procedure. This percentage change in root mean square error (RMSE) is computed as follows:

$$Percent\ change = \left[\frac{RMSE_{previous} - RMSE_{current}}{RMSE_{current}}\right]100$$

R-Squared

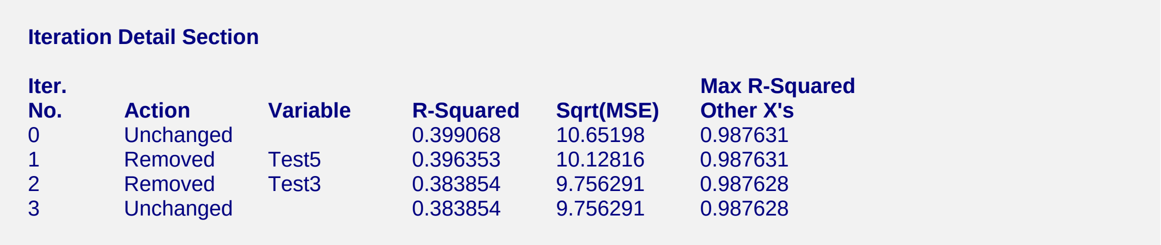
This is the R-Squared value for the current model.

Sqrt(MSE)

This is the square root of the mean square error for the current model.

Iteration Detail Section (Brief Version)

This report was not printed because the Report Format box was set to Verbose. If this option had been set to Brief, this is the output that would have been displayed.



```
Iteration Detail Section

Iter.                                                              Max R-Squared
No.      Action      Variable    R-Squared    Sqrt(MSE)    Other X's
0        Unchanged               0.399068     10.65198     0.987631
1        Removed     Test5       0.396353     10.12816     0.987631
2        Removed     Test3       0.383854     9.756291     0.987628
3        Unchanged               0.383854     9.756291     0.987628
```

This is an abbreviated report summarizing the statistics at each iteration. Individual definitions of the items on the report are as follows:

Iter. No.

The number of this iteration.

Action

For each iteration, there are three possible actions:

1. *Unchanged*. No action was taken because of the scan in this step. Because of the "backward look" in the stepwise search method, this will show up a lot when this method is used. Otherwise, it will show up at the first and last steps.
2. *Removed*. A variable was removed from the model.
3. *Added*. A variable was added to the model.

Variable

This is the name of the variable whose status is being changed.

R-Squared

The value of R-Squared for the current model.

Sqrt(MSE)

This is the square root of the mean square error for the current model.

Max R-Squared Other X's

This is the maximum value of R-Squared Other X's (see verbose report definitions) for all the variables in the model. This is a collinearity model. You want this value to be as small as possible. If it approaches 0.99, you should be concerned with the possibility that multicollinearity is distorting your results.

Chapter 312

All Possible Regressions

Introduction

Often, theory and experience give only general direction as to which of a pool of candidate variables (including transformed variables) should be included in the regression model. The actual set of predictor variables used in the final regression model must be determined by analysis of the data. Determining this subset is called the *variable selection* problem.

Finding this subset of regressor (independent) variables involves two opposing objectives. First, we want the regression model to be as complete and realistic as possible. We want every regressor that is even remotely related to the dependent variable to be included. The phrase “throw in the kitchen sink” takes on new meaning here. Second, we want to include as few variables as possible because each irrelevant regressor decreases the precision of the estimated coefficients and predicted values. Also, the presence of extra variables increases the complexity of data collection and model maintenance. The goal of variable selection becomes one of parsimony: achieve a balance between simplicity (as few regressors as possible) and fit (as many regressors as needed).

After a pool of candidate variables has been formed, the next task is to establish a basis for comparing two models. How do we decide if model A is better than model B? Three statistics have been found useful for selecting among various regression models. These are R-Squared, mean square error, and Cp. Other criteria have been suggested, but these three are the most popular.

Once we have a pool of variables and a selection criterion, the final task in variable selection is to plan a strategy to see how each of the possible models does on the criterion. The problem that now arises is that there are too many possible models to choose from. The number of possible models that can be formed from p regressors is 2 to the power p. If we have p = 4 regressors, there are 16 possible models to choose from. With 15 regressors, there are 32,768 possible models. With 20 regressors, there are 1,048,576 models. Obviously, the number of possible models grows exponentially with the number of regressors. However, with up to 15 regressors, the problem does seem manageable.

This procedure was programmed so that it will efficiently look at up to 32,768 models for up to 15 regressors. That is why it is called *all possible regressions*. It guarantees that you will find the “best” model, since it looks at all of them. Unfortunately, no automatic procedure will find the “best” model in every sense. It will, however, find the model that is best according to your selection criterion. It is still left up to you to determine if the model makes theoretical and practical sense.

All Possible Regressions

This algorithm fits all regressions involving one regressor, two regressors, three regressors, and so on. The selection criterion is recorded for each regression. Once the procedure finishes, the champion for each subset size is determined. You then determine which subset size is optimum for your case.

The All Possible Regressions solution is the target of the popular step-regression procedures. Although it takes longer to run, it guarantees the right answer. Hence, when you have 15 or fewer independent variables to choose from, this is the variable selection procedure you should use.

Assumptions and Limitations

The same assumptions and qualifications apply here as applied to multiple regression. We refer you to the Assumptions section in the Multiple Regression chapter (Chapter 15) for a discussion of these assumptions. We will here mention restrictions necessary for this algorithm to work.

Number of Regressor Variables

This procedure will work with up to fifteen regressor variables, not including the intercept. The intercept is always included in the regression model.

Number of Observations

The number of observations must be at least one greater than the number of candidate regressors. A popular rule-of-thumb when using any variable selection procedure is that you have at least five observations for each candidate variable.

No Linear Dependencies

Since one of the models that must be solved involves all of the candidate variables (the full model), no linear dependencies can exist among these variables. A linear dependency occurs when one variable is a weighted average of the rest. For example, if one variable is the total of several others, it cannot be included in the candidate pool.

Data Structure

An example of data appropriate for this procedure is shown in the table below. These data are from a study of the relationships of several variables with a person’s IQ. Fifteen people were studied. Each person’s IQ was recorded along with scores on five different personality tests. The data are contained in the SAMPLE database. We suggest that you open this database now so that you can follow along with the example.

SAMPLE dataset

Test1	Test2	Test3	Test4	Test5	IQ
83	34	65	63	64	106
73	19	73	48	82	92
54	81	82	65	73	102
96	72	91	88	94	121
84	53	72	68	82	102
86	72	63	79	57	105
76	62	64	69	64	97
54	49	43	52	84	92
37	43	92	39	72	94
42	54	96	48	83	112
71	63	52	69	42	130
63	74	74	71	91	115
69	81	82	75	54	98
81	89	64	85	62	96
50	75	72	64	45	103

Missing Values

Rows with missing values in the variable pool are ignored. The program does not run a separate analysis for each pattern of missing values. It is possible to get slightly different results when you analyze a subset because variables in the subset may not contain missing values. Without the missing values, the rows that were deleted in the original analysis are included in the subset analysis.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Dependent Variable

Y: Dependent Variable

Specifies a dependent variable. If more than one variable is specified, a separate analysis is run for each.

Weight Variable

Weight Variable

Specifies a variable containing observation (row) weights for generating weighted-regression analysis.

Independent Variables

X’s: Independent Variables

Specify the independent variables.

Model Selection

Best-Model Criterion

This option lets you specify whether the model selection is based on the model’s R-Squared or mean square error (MSE) value.

Alternative Models

While the algorithm is scanning for the “best” model for each subset size, it can retain information on the “near-best” models as well. This option specifies how many of these near-best models you want to see information on. It is often useful to see information reported on five to ten extra models.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Descriptive Statistics and Selected Variables Reports

Specifies whether these reports are displayed.

Select Plots

R-Squared Plot ... Cp Plot

Specifies whether these plots are displayed.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

R-Squared Plot ... Cp Plot Tabs

Various plots may be displayed to help you interpret the results of your search. Since all three of these panels have the same options, they will only be discussed once.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – All Possible Regressions Analysis

This section presents an example of how to run a all possible regressions analysis of the data contained in the SAMPLE database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the All Possible Regressions window.

1 **Open the SAMPLE dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sample.s0**.
- Click **Open**.

2 **Open the All Possible Regressions window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Variable Selection Routines**, then **All Possible Regressions**. The All Possible Regressions procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the All Possible Regressions window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable** text box. This will bring up the variable selection window.
- Select **IQ** from the list of variables and then click **Ok**. "IQ" will appear in the Y: Dependent Variable box.
- Double-click in the **X's: Independent Variable(s)** text box. This will bring up the variable selection window.

- Select **Test1** through **Test5** from the list of variables and then click **Ok**. “Test1-Test5” will appear in the X’s: Independent Variable(s).
- Enter **5** in the **Alternative Models** box.

4 Specify the reports.

- On the All Possible Regressions window, select the **Reports tab**.
- Check the **Descriptive Statistics** checkbox.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Descriptive Statistics Section

```
Descriptive Statistics Section

Variable      Count    Mean        Standard Deviation
Test1         15       67.93333    17.39239
Test2         15       61.4        19.39735
Test3         15       72.33334    14.73415
Test4         15       65.53333    13.95332
Test5         15       69.93333    16.15314
IQ            15       104.3333    11.0173
```

For each variable, the count of nonmissing values, the arithmetic mean of the nonmissing values, and the standard deviation of the nonmissing values are computed. This report is particularly useful for checking that the correct variables were selected.

All Possible Results Section

```
All Possible Results Section

Model                   Root
Size     R-Squared      MSE         Cp          Model
1        .137941        10.61539    1.910838    D  (Test4)
1        .057913        11.09719    3.109405    B  (Test2)
1        .050917        11.13832    3.214175    A  (Test1)
1        .005486        11.40179    3.894581    C  (Test3)
1        .003371        11.41391    3.926255    E  (Test5)

2        .154246        10.94386    3.666643    CD
2        .144790        11.00487    3.808266    AD
2        .139411        11.03943    3.888825    BD
2        .137980        11.0486     3.910256    DE
2        .098957        11.29591    4.494690    AB

3        .383854        9.756291    2.227864    ABD
3        .159103        11.39763    5.593906    BCD
3        .157158        11.4108     5.623033    ACD
3        .155707        11.42062    5.644768    CDE
3        .145431        11.48991    5.798660    ADE
```

```
4         .396353      10.12816     4.040666     ABCD
4         .383859      10.23245     4.227794     ABDE
4         .163351      11.92369     7.530276     BCDE
4         .157627      11.96441     7.616005     ACDE
4         .115826      12.25768     8.242057     ABCE

5         .399068      10.65198     6.000000     ABCDE
```

This report presents the results of the all possible regressions search procedure. The models for each subset (model) size are sorted from best to worst. To use this report, you scan down a criterion column, say R-Squared, for the subset size where this value stabilizes. In this example, the R-Squared value for the best three-variable model is 0.383854 and the R-Squared for the best four-variable model is 0.396353. This is a minor increase. We would select the three-variable model as our final model.

Model Size

This is the number of independent variables in the model. Model size will range from 1 to p. The option, Alternative Models, controls the number of models reported from each model subset size.

R-Squared

R-Squared is the ratio of the variation explained by the model to the total variation in the dependent variable. R-Squared ranges from zero to one. The larger the R-Squared, the better the model. A comprehensive definition of R-Squared is given in the Multiple Regression chapter.

Root MSE

This is the square root of the mean square error. The smaller this value is, the better the model.

Cp

Another criterion for variable selection and importance is Mallow’s Cp statistic. The optimum model will have a Cp value close to p+1, where p is the number of independent variables. A Cp greater than (p+1) indicates that the regression model is overspecified (contains too many variables and stands a chance of having collinearity problems). On the other hand, a model with a Cp less than (p+1) indicates that the regression model is underspecified (at least one important independent variable has been omitted). The formula for the Cp statistic is as follows, where k is the maximum number of independent variables available:

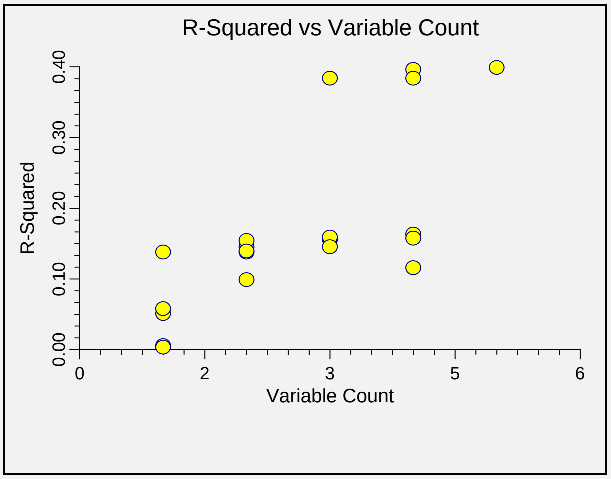
$$C_p = \left[\frac{MSE_p}{MSE_k}\right][n - p - 1] - [n - 2(p + 1)]$$

Model

This column labels the model whose statistics are being reported. Letters are given in a shorthand notation to represent the independent variables. The letter A is associated with the first variable, the letter B with the second, and so on. The letters and corresponding variable names are displayed for all variables in the first section of the report. Two-variable models are represented by two letters. Hence, in this example, the model CD represents the two-variable model consisting of variables Test3 and Test4.

R-Squared vs. Variable Count Plot

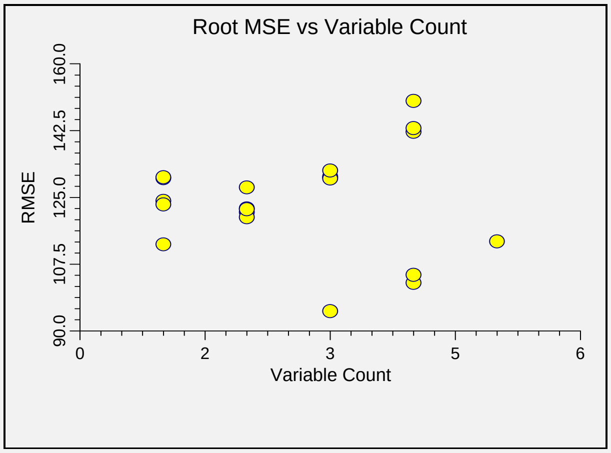
This plot displays the values of R-Squared on the vertical axis and the number of independent variables on the horizontal axis for the data displayed in the All Possible Results Section above. Note the large disparity in the three-variable models. There is one model that is way above the rest. We can also quickly see that the four- and five-variable models do not do much better.



Root MSE vs. Variable Count Plot

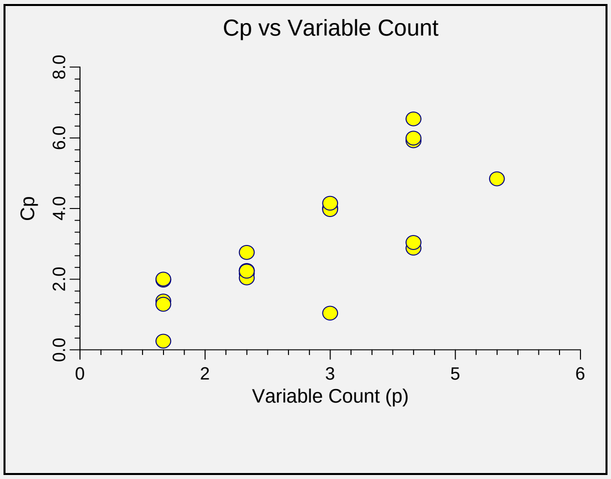
This plot displays the values of the square root of the mean square error on the vertical axis and the number of independent variables on the horizontal axis for the data displayed in the All Possible Results Section above. Note the large disparity in the three-variable models. There is one model that is way below the rest.

Root MSE is often considered a better criterion for choosing a best model than R-Squared. The root MSE decreases initially as p increases, stabilizes at some subset size, and eventually begins to increase with further increments of p. You should choose a best model based on the minimum MSE or a value of p near the point where the smallest MSE turns upward. The model subset that minimizes MSE will usually maximize R-Squared.



Cp vs. Variable Count Plot

This plot displays the values of Cp on the vertical axis and the number of independent variables on the horizontal axis for the data displayed in the All Possible Results Section above. The Cp plot is more difficult to interpret because we are looking for the model where Cp is closest to p+1. You will most likely want to stick with the numeric report when considering Cp.



Chapter 315

Nonlinear Regression

Introduction

Multiple regression deals with models that are linear in the parameters. That is, the multiple regression model may be thought of as a weighted average of the independent variables. A *linear* model is usually a good first approximation, but occasionally, you will require the ability to use more complex, nonlinear, models. Nonlinear regression models are those that are not linear in the parameters. Examples of nonlinear equations are:

Y = A + B EXP(-CX)

Y = (A +BX)/(1+CX)

Y = A + B/(C+X)

This program estimates the parameters in nonlinear models using the Levenberg-Marquardt nonlinear least-squares algorithm as presented in Nash (1987). We have implemented Nash's MRT algorithm with numerical derivatives. This has been a popular algorithm for solving nonlinear least squares problems, since the use of numerical derivatives means you do not have to supply program code for the derivatives.

Starting Values

Many people become frustrated with the complexity of nonlinear regression after dealing with the simplicity of multiple linear regression. Perhaps the biggest nuisance with the algorithm used in this program is the need to supply bounds and starting values. The convergence of the algorithm depends heavily upon supplying appropriate starting values.

Sometimes you will be able to use zeros or ones as starting values, but often you will have to come up with better values. One accepted method for obtaining a good set of starting values is to estimate them from the data. We will show you how this is done with the example that we will be using throughout this chapter.

Suppose you have 44 observations on X and Y (the data are shown below). Suppose further that you want to fit the specific nonlinear model:

Y = A + (0.49 - A) Exp (-B(X - 8)).

Since there are two unknown parameters, A and B, we select two observations. To make the estimates as representative as possible, we select observations from each end of the range of X values. The two observations we select are (10, 0.48) and (42, 0.39). Putting these two observations into our model yields two equations with two unknowns:

(1) $\quad$ **0.48 = A + (0.49 - A) Exp (-B(10 - 8))**

(2) $\quad$ **0.39 = A + (0.49 - A) Exp (-B(42 - 8))**.

Solving (1) for B yields

(3) $\quad$ **B={log((0.48 - A)/(0.49 - A)}/(-2).**

Putting this result into the second equation yields

(4) $\quad$ **{(0.39 - A)/(0.49 - A)}={(0.48 - A)/(0.49 - A)}^17.**

These equations appear difficult, but since we are only after starting values, we can analyze them for possible values of A and B. From (3), we see that A must be less than 0.48 and greater than 0. Suppose we pick a number in this range, say 0.1. Next, using (3), we calculate B as 0.013. These are our starting values.

Reviewing the steps we have taken:

1. Select one data value for each parameter.
2. Plug the selected data values into the model and solve for the parameters. If the model is too difficult, analyze the resulting equations for possible ranges of each parameter.
3. Try these starting values in the program. Remember, you do not have to be too accurate, just in the ball park.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in one dependent variable and one or more independent variables. An example of data appropriate for this procedure, taken from page 476 of Draper and Smith (1981), is shown below. These data are contained in the DS746 dataset. In this example, the dependent variable (Y) is the proportion of available chlorine in a certain quantity of chlorine solution and the independent variable (X) is the length of time in weeks since the product was produced. When the product is produced, the proportion of chlorine is 0.50. During the 8 weeks that it takes to reach the consumer, the proportion declines to 0.49. The hypothesized model for predicting Y from X is

$$\mathbf{Y = A + (0.49 - A)\ EXP(- B(X\text{-}8)) + e.}$$

Here, A and B are the parameters and e is the error or residual.

DS476 dataset

Row	X	Y
1	8	0.49
2	8	0.49
3	10	0.48
4	10	0.47
5	10	0.48
6	10	0.47
7	12	0.46
8	12	0.46
9	12	0.45
10	12	0.43
11	14	0.45
12	14	0.43
13	14	0.43
14	16	0.44
15	16	0.43
16	16	0.43
17	18	0.46
18	18	0.45
19	20	0.42
20	20	0.42
21	20	0.43
22	22	0.41
23	22	0.41
24	22	0.40
25	24	0.42
26	24	0.40
27	24	0.40
28	26	0.41
29	26	0.40
30	26	0.41
31	28	0.41
32	28	0.40
33	30	0.40
34	30	0.40
35	30	0.38
36	32	0.41
37	32	0.40
38	34	0.40
39	36	0.41
40	36	0.38
41	38	0.40
42	38	0.40
43	40	0.39
44	42	0.39

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Model Tab and Parameters - Cont Tab

These panels specify the model and variables used in the analysis.

Dependent Variable

Y: Dependent Variable

Specifies a single dependent (*Y*) variable from the current database.

Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your particular study.

Model

Model

This box contains the nonlinear equation.

This expression is made up of

1. Symbols: +, -, *, /, ^, <, >, =, (, and).
2. Functions:

ABS(X)	Absolute value of X
ASN(X)	Arc sine of X
ATN (X)	Arc tangent of X
COS(X)	Cosine of X
EXP(X)	Exponential of X
INT(X)	Integer part of X
LN(X)	Log base e of X
LOG(X)	Log base 10 of X
SGN(X)	Sign of X
SIN(X)	Sine of X
SQR(X)	Square root of X
TAN(X)	Tangent of X

3. One or more variables referenced by name.
4. Parameters, which are defined below.
5. Constants.

The syntax of the model expression follows that of the variable transformations, so we will not go into syntax here, but refer you to the Variable Transformations chapter. Note that only a subset of the functions available as transformations are also available here.

Model Parameters

Parameter

The model may contain up to thirty parameters. Each parameter used in the model must be defined in this section by entering a name, bounds, and starting value.

The parameter name is any combination of letters and numbers, except that the name must begin with a letter. You should not use symbols in the parameter name. All letters are converted to upper case internally, so it does not matter whether you use upper or lower case. The name cannot be one of the internal mathematical functions like SIN or TAN, as this will confuse the function parser. Also, the parameter name should not be the same as a variable name.

The name may be as long as you want, but, for readability, you should keep it short.

The model may contain up to thirty parameters.

Min Start Max

The minimum, starting value, and maximum are entered in this box. The three values must be separated with blanks or commas.

- **Minimum**

  This is the smallest value that the parameter can take on. The algorithm searches for a value between this and the Maximum Value. If you want to search in an unlimited range, enter a large negative number such as **-1E9**, which is -1000000000.

  Since this is a search algorithm, the narrower the range that you search in, the quicker the process will converge.

  Care should be taken to specify minima and maxima that keep calculations in range. Suppose, for example, that your equation includes the expression LOG(B*X) and that values of X are positive. Since you cannot take the logarithm of zero or a negative number, you should set the minimum of B as a positive number. This will insure that the estimation procedure will not fail because of impossible calculations.

- **Starting Value**

  This is the beginning value of the parameter. The algorithm searches for a value between the Minimum Value and the Maximum Value, beginning with this number. The closer this value is to the final value, the quicker the algorithm will converge.

  Although specific instructions for finding starting values are given at the beginning of this chapter, we would like to make the following suggestions here.

  1. Make sure that the starting values you supply are legitimate. For example, if the model you were estimating included the phrase 1/B, you would not want to start with B=0.

  2. Before you go to a lot of effort, make a few trial runs using starting values of 0.0, 0.5, and 1.0. Often, one of these values will converge.

  3. If you have a large number of observations, take a small sample of observations from your original database and work with this subset database. When you find a set of starting values that converges on this subset database, use the resulting parameter estimates as starting values with the complete database. Since nonlinear regression is iterative and each iteration must pass through the complete database, this can save a considerable amount of time while you are searching for starting values.

- **Maximum**

  This is the largest value that the parameter can take on. The algorithm searches for a value between the Minimum Value and this value, beginning at the Starting Value. If you want to search in an unlimited range, enter a large positive number such as **1E9**, which is 1000000000.

  Since this is a search algorithm, the narrower the range that you search in, the quicker the process will converge.

  Care should be taken to specify minima and maxima that keep calculations in check. Suppose, for example, that your equation includes the expression LOG(B*X) and that values of X are negative. Since you cannot take the logarithm of zero or a negative number, you should set the maximum of B as a negative number near zero. This will insure that the estimation procedure will not fail because of impossible calculations.

Options Tab

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc.

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec.

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the least squares algorithm. To remove the effects of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Additional Reports

Iteration Report ... Predicted Value and Residual Report

Each of these options specifies whether the indicated report is displayed.

Select Plots

Probability Plot ... Residuals vs X Plot

Each of these options specifies whether the indicated plot is displayed.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Prob Plot to Resid vs X Plot Tabs

Various plots may be displayed to help you validate the assumptions of your regression analysis as well as investigate the fit of your estimated equation. The actual uses of these plots will be described later. Each of these plots includes the following options.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tick marks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. An *{M}* is replaced by model expression. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that the variables you specify must already have been named on the current database.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Predicted Values

The predicted (Yhat) values.

Lower Prediction Limit

The lower confidence limit of the predicted value.

Upper Prediction Limit

The upper confidence limit of the predicted value.

Residuals

The residuals (Y-Yhat).

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Nonlinear Regression Analysis

This section presents an example of how to run a nonlinear regression analysis of the data that was presented above in the Data Structure section. In this example, we will fit the model

$$Y = A + (0.49 - A)\ EXP(- B(X-8))$$

to the data contained in the variables Y and X on the database DS476.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Nonlinear Regression window.

1 Open the DS476 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **DS476.s0**.
- Click **Open**.

2 Open the Nonlinear Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Nonlinear Routines**, then **Nonlinear Regression**. The Nonlinear Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Nonlinear Regression window, select the **Model tab**.
- Double-click in the **Y: Dependent Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**. "Y" will appear in the Y: Dependent Variable box.

- In the **Model** box, enter **A+(0.49-A)*EXP(-B*(X-8))**. Note that A and B are parameters to be defined below, X is a variable on the database, and EXP is the name of a function.
- Enter **A** in the **first Parameter** box.
- Enter **0 0.1 1** in the first **Min Start Max** box.
- Enter **B** in the **second Parameter** box.
- Enter **0 0.013 1** in the **second Min Start Max** box.

4 Specify the reports.

- Select the **Reports tab**.
- Check all reports and plots.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section

Itn     Error Sum
No.     of Squares          Lambda          A               B
0       1.643321E-02        0.00004         0.1             0.013
Stepsize reduced to .6032159 by bounds.
Stepsize reduced to .8185954 by bounds.
1       0.0147339           0.016           0.1464944       1.375224E-02
2       1.461316E-02        0.0064          0.2331648       1.792601E-02
3       1.278996E-02        0.0256          0.2486083       2.104608E-02
4       1.218322E-02        0.01024         0.3052482       2.783621E-02
5       0.0102341           0.04096         0.3141184       3.271746E-02
6       9.077431E-03        0.016384        0.3472015       0.0425624
7       7.713023E-03        0.065536        0.3519653       4.887892E-02
8       6.631856E-03        0.0262144       0.3685163       6.024325E-02
9       5.852748E-03        1.048576E-02    0.386985        8.150943E-02
10      5.045347E-03        4.194304E-03    0.3904939       9.880718E-02
11      5.001693E-03        1.677722E-03    0.390121        0.1015279
12      5.00168E-03         6.710886E-04    0.39014         0.101633
Convergence criterion met.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm's progress toward the solution.

Model Estimation Section

```
Model Estimation Section

Parameter           Parameter       Asymptotic      Lower           Upper
Name                Estimate        Standard Error  95% C.L.        95% C.L.
A                   0.39014         5.033759E-03    0.3799815       0.4002985
B                   0.101633        1.336168E-02    7.466801E-02    0.1285979

Model               Y = A+(0.49-A)*EXP(-B*(X-8))
R-Squared           0.873375
Iterations          12
Estimated Model
(.39014)+(0.49-(.39014))*EXP(-(.101633)*((X)-8))
```

This section reports the parameter estimates.

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Model

The model that was estimated. Use this to double check that the model estimated was what you wanted.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

$$\textit{R-Squared} = (\textit{ModelSS} - \textit{MeanSS})/(\textit{TotalSS}-\textit{MeanSS})$$

where

MeanSS is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

Estimated Model

This expression displays the estimated nonlinear-regression model. It is displayed in this format so that it may be copied to the clipboard and used elsewhere. For example, you could copy this expression here and paste it as a Variable Transformation.

Analysis of Variance Table

```
Analysis of Variance Table
                                        Sum of              Mean
Source                  DF              Squares             Square
Mean                    1               7.9475              7.9475
Model                   2               7.981998            3.990999
Model (Adjusted)        1               3.449832E-02        3.449832E-02
Error                   42              5.00168E-03         1.190876E-04
Total (Adjusted)        43              0.0395
Total                   44              7.987
```

The section presents an analysis of variance table.

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just SSE.

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters

                        A               B
A                       1.000000        0.887330
B                       0.887330        1.000000
```

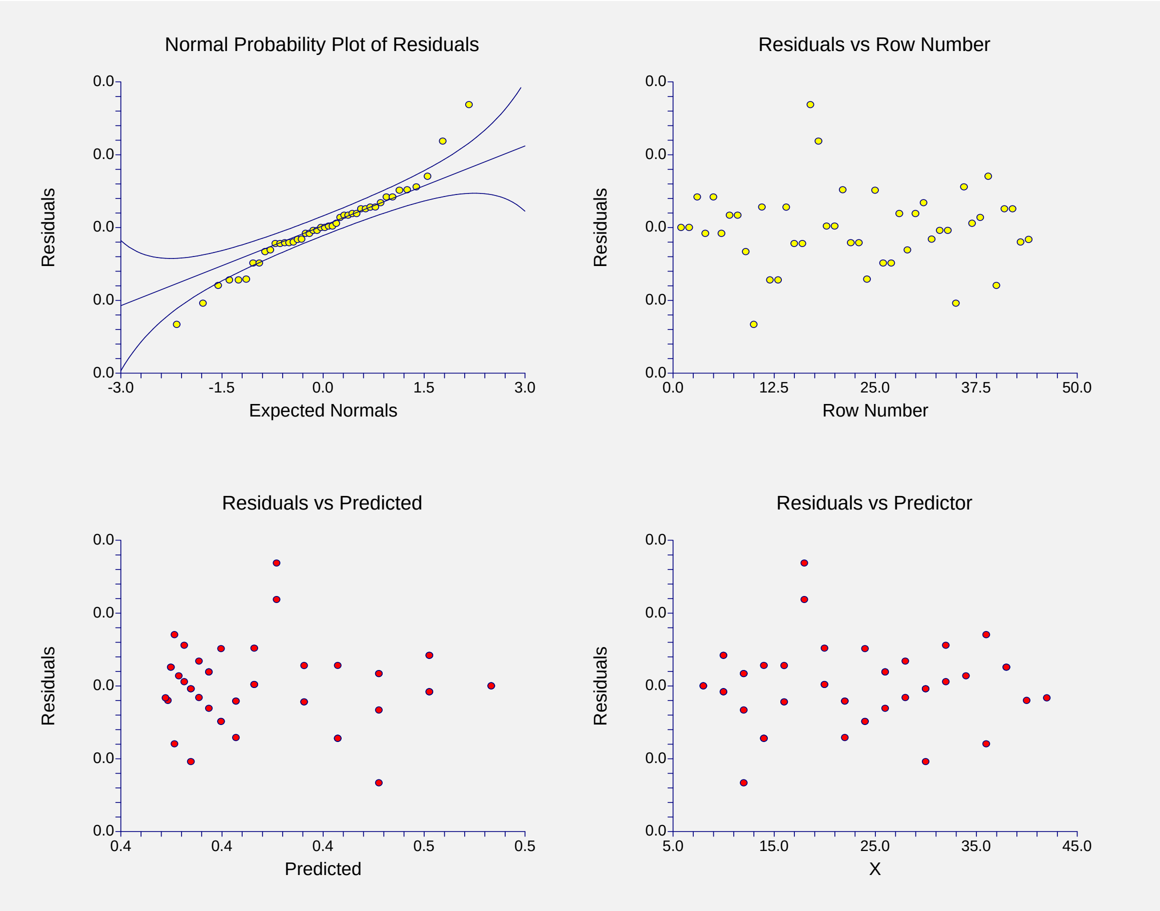
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section

Row                            Predicted        Actual
No.    Residual                Y                Y             X
1      0                       0.49             0.49          8
2      0                       0.49             0.49          8
3      -8.368287E-03           0.4716317        0.48          10
4      1.631713E-03            0.4716317        0.47          10
.      .                       .                .             .
.      .                       .                .             .
.      .                       .                .             .
```

This section shows the values of the residuals and predicted values. If you have observations in which the independent variables were given, but the dependent (Y) variable is missing, a predicted value will be generated and displayed in this report.

Residual Plots



Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps, plateaus, or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus Predicted Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance. A sloping or curved band signifies inadequate specification of the model. A sloping band with increasing or decreasing variability could suggest nonconstant variance and inadequate specification of the model.

Residual versus Independent Variable(s) Plot

This is a scatter plot of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

Predicting for New Values

You can use your model to predict Y for new values of the independent variables. Here is how. Add new rows to the bottom of your database containing the values of the independent variable(s) that you want to create predictions from. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows with a complete set of independent variables, regardless of whether the Y variable is available.

Chapter 320

Logistic Regression

Introduction

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is often used when the dependent variable has only two values. The name *multiple-group logistic regression* (MGLR) is usually reserved for the case when the dependent variable has three or more unique values. Multiple-group logistic regression is sometimes called *multinomial, polytomous, polychotomous,* or *nominal logistic regression.* Although the data structure is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing discrete response variables. In fact, the current feeling among many statisticians is that logistic regression is more versatile and better suited for most situations than is discriminant analysis because it does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes both regular (binary) logistic regression and multiple-group logistic regression on both numeric and categorical variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform a subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values, and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

The Logit and Logistic Transformations

In multiple regression, a mathematical model of a set of explanatory variables is used to predict the mean of the dependent variable. In logistic regression, a mathematical model of a set of explanatory variables is used to predict a transformation of the dependent variable. This is the *logit* transformation.

Suppose the numerical values of 0 and 1 are assigned to the two categories of a binary variable. Often, the 0 represents a negative response and the 1 represents a positive response. The mean of this variable will be the proportion of positive responses. Because of this, you might try to model the relationship between the probability (proportion) of a positive response and the explanatory variables.

If *p* is the proportion of observations with a response of 1, then 1-*p* is the probability of a response of 0. The ratio *p*/(1-*p*) is call the *odds* and the *logit* is the logarithm of the odds, or just *log odds*. Mathematically, the logit transformation is written

$$l = \operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

The following table shows the logit for various values of *p*.

P	Logit(P)	P	Logit(P)
0.001	-6.907	0.999	6.907
0.01	-4.595	0.99	4.595
0.05	-2.944	0.95	2.944
0.10	-2.197	0.90	2.197
0.20	-1.386	0.80	1.386
0.30	-0.847	0.70	0.847
0.40	-0.405	0.60	0.405
0.50	0.000		

Note that while *p* ranges between zero and one, the logit ranges between minus and plus infinity. Also note that the zero logit occurs when *p* is 0.50.

The *logistic* transformation is the inverse of the logit transformation. It is written

$$p = \operatorname{logistic}(l) = \frac{e^{l}}{1+e^{l}}$$

The Log Odds Ratio Transformation

The difference between two log odds can be used to compare two proportions, such as that of males versus females. Mathematically, this difference is written

$$\begin{aligned} l_1 - l_2 &= \operatorname{logit}(p_1) - \operatorname{logit}(p_2) \\ &= \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) \\ &= \ln\left(\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_2}{1-p_2}\right)}\right) \\ &= \ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right) \\ &= \ln\left(OR_{1,2}\right) \end{aligned}$$

This difference is often referred to as the *log odds ratio*. The odds ratio is often used to compare proportions across groups. Note that the logistic transformation is closely related to the odds ratio. The reverse relationship is

$$OR_{1,2} = \mathrm{e}^{(l_1 - l_2)}$$

The Logistic Regression and Logit Models

In multiple-group logistic regression, a discrete dependent variable *Y* having *G* unique values $(G \geq 2)$ is regressed on a set of *p* independent variables $X_1, X_2, ..., X_p$. *Y* represents a way of partitioning the population of interest. For example, *Y* may be presence or absence of a disease, condition after surgery, or marital status. Since the names of these partitions are arbitrary, refer to them by consecutive numbers. That is, in the discussion below, *Y* will take on the values 1, 2, …, *G*. In fact, ***NCSS*** allows *Y* to have both numeric and text values, but the notation is much simpler if integers are used.

In the discussion to follow, let

$$\mathrm{X} = (X_1, X_2, \cdots, X_p)$$

$$B_g = \begin{pmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{pmatrix}$$

The logistic regression model is given by the *G* equations

$$\begin{aligned} \ln\left(\frac{p_g}{p_1}\right) &= \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \\ &= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g \end{aligned}$$

Here, p_g is the probability that an individual with values $X_1, X_2, ..., X_p$ is in group *g*. That is,

$$p_g = \Pr(Y = g \mid \mathrm{X})$$

Usually $X_1 \equiv 1$ (that is, an intercept is included), but this is not necessary. The quantities $P_1, P_2, ..., P_G$ represent the prior probabilities of group membership. If these prior probabilities are assumed equal, then the term $\ln(P_g / P_1)$ becomes zero and drops out. If the priors are not assumed equal, they change the values of the intercepts in the logistic regression equation.

Group one is called the *reference group*. The regression coefficients $\beta_{11}, \beta_{12}, \cdots, \beta_{1p}$ for the reference group are set to zero. The choice of the reference group is arbitrary. Usually, it is the largest group or a control group to which the other groups are to be compared. This leaves *G*-1 logistic regression equations in the multinomial logistic model.

The $\beta' s$ are population regression coefficients that are to be estimated from the data. Their estimates are represented by *b*'s. The $\beta' s$ represents the unknown parameters, while the *b*'s are their estimates.

These equations are linear in the logits of *p*. However, in terms of the probabilities, they are nonlinear. The corresponding nonlinear equations are

$$\mathrm{p_g} = \mathrm{Prob}(Y = g \mid \mathrm{X}) = \frac{e^{\mathrm{XB}_g}}{1 + e^{\mathrm{XB}_2} + e^{\mathrm{XB}_3} + \cdots + e^{\mathrm{XB}_G}}$$

since $e^{\mathrm{XB}_1} = 1$ because all of its regression coefficients are zero.

A note on the names of the models. Often, all of these models are referred to as *logistic regression models*. However, when the independent variables are coded as ANOVA type models, they are sometimes called *logit models*.

A note about the interpretation of e^{XB} may be useful. Using the fact that $e^{a+b} = \left(e^a\right)\left(e^b\right)$, e^{XB} may be re-expressed as follows

$$\begin{aligned} e^{\mathrm{XB}} &= e^{\beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_p \mathrm{X}_p} \\ &= e^{\beta_1 \mathrm{X}_1} e^{\beta_2 \mathrm{X}_2} \cdots e^{\beta_p \mathrm{X}_p} \end{aligned}$$

This shows that the final value is the product of its individual terms.

Solving the Likelihood Equations

To improve notation, let

$$\begin{aligned} \pi_{gj} &= \mathrm{Prob}\left(Y = g \mid \mathrm{X}_j\right) \\ &= \frac{e^{\mathrm{X}_j \mathrm{B}_g}}{e^{\mathrm{X}_j \mathrm{B}_1} + e^{\mathrm{X}_j \mathrm{B}_2} + \cdots + e^{\mathrm{X}_j \mathrm{B}_G}} \\ &= \frac{e^{\mathrm{X}_j \mathrm{B}_g}}{\sum_{s=1}^{G} e^{\mathrm{X}_j \mathrm{B}_s}} \end{aligned}$$

The likelihood for a sample of *N* observations is then given by

$$l = \prod_{\mathrm{j}=1}^{\mathrm{N}} \prod_{g=1}^{G} \pi_{gj}^{\;y_{gj}}$$

where y_{gj} is one if the j^{th} observation is in group *g* and zero otherwise.

Using the fact that $\sum_{g=1}^{G} y_{gj} = 1$, the log likelihood, *L*, is given by

$$\begin{aligned} L = \ln(l) &= \sum_{j=1}^{\mathrm{N}} \sum_{g=1}^{G} \mathrm{y}_{gj} \ln\left(\pi_{gj}\right) \\ &= \sum_{j=1}^{\mathrm{N}} \sum_{g=1}^{G} \mathrm{y}_{gj} \ln\left(\frac{\mathrm{e}^{\mathrm{X}_j \mathrm{B}_g}}{\sum_{s=1}^{G} \mathrm{e}^{\mathrm{X}_j \mathrm{B}_s}} \right) \\ &= \sum_{\mathrm{j}=1}^{\mathrm{N}} \left[\sum_{\mathrm{g}=1}^{\mathrm{G}} y_{gj} \mathrm{X}_j \mathrm{B}_g - \ln\left(\sum_{g=1}^{G} \mathrm{e}^{\mathrm{X}_j \mathrm{B}_g} \right) \right] \end{aligned}$$

Maximum likelihood estimates of the β's are found by finding those values that maximize this log likelihood equation. This is accomplished by calculating the partial derivatives and setting them to zero. The resulting likelihood equations are

$$\frac{\partial L}{\partial \beta_{ik}} = \sum_{j=1}^{N} x_{kj}\left(y_{ig} - \pi_{ig}\right)$$

for g = 1, 2, …, G and k = 1, 2, …, p. Actually, since all coefficients are zero for g = 1, the range of g is from 2 to G.

Because of the nonlinear nature of the parameters, there is no closed-form solution to these equations and they must be solved iteratively. The Newton-Raphson method as described in Albert and Harris (1987) is used to solve these equations. This method makes use of the information matrix, $I(\beta)$, which is formed from the matrix of second partial derivatives. The elements of the information matrix are given by

$$\frac{\partial^2 L}{\partial \beta_{ik} \partial \beta_{ik'}} = -\sum_{j=1}^{N} x_{kj} x_{k'j} \pi_{ig}\left(1 - \pi_{ig}\right)$$

$$\frac{\partial^2 L}{\partial \beta_{ik} \partial \beta_{i'k'}} = \sum_{j=1}^{N} x_{kj} x_{k'j} \pi_{ig} \pi_{i'g}$$

The information matrix is used because the asymptotic covariance matrix of the maximum likelihood estimates is equal to the inverse of the information matrix. That is,

$$V\left(\hat{\beta}\right) = I(\beta)^{-1}$$

This covariance matrix is used in the calculation of confidence intervals for the regression coefficients, odds ratios, and predicted probabilities.

Interpretation of Regression Coefficients

The interpretation of the estimated regression coefficients is not as easy as in multiple regression. In multinomial logistic regression, not only is the relationship between X and Y nonlinear, but also, if the dependent variable has more than two unique values, there are several regression equations.

Consider the simple case of a binary dependent variable, Y, and a single independent variable, X. Assume that Y is coded so it takes on the values 0 and 1. In this case, the logistic regression equation is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Now consider impact of a unit increase in X. The logistic regression equation becomes

$$\begin{aligned} \ln\left(\frac{p'}{1-p'}\right) &= \beta_0 + \beta_1 (X+1) \\ &= \beta_0 + \beta_1 X + \beta_1 \end{aligned}$$

We can isolate the slope by taking the difference between these two equations. We have

$$\begin{aligned}\beta_1 &= \beta_0 + \beta_1(X+1) - \beta_0 + \beta_1 X \\ &= \ln\left(\frac{p'}{1-p'}\right) - \ln\left(\frac{p}{1-p}\right) \\ &= \ln\left(\frac{\frac{p'}{1-p'}}{\frac{p}{1-p}}\right) \\ &= \ln\left(\frac{odds'}{odds}\right)\end{aligned}$$

That is, β_1 is the log of the ratio of the odds at X+1 and X. Removing the logarithm by exponentiating both sides gives

$$e^{\beta_1} = \frac{odds'}{odds}$$

The regression coefficient β_1 is interpreted as the log of the odds ratio comparing the odds after a one unit increase in X to the original odds. Note that, unlike multiple regression, the interpretation of β_1 depends on the particular value of X since the probability values, the p's, will vary for different X.

Binary X

When X can take on only two values, say 0 and 1, the above interpretation becomes even simpler. Since there are only two possible values of X, there is a unique interpretation for β_1 given by the log of the odds ratio. In mathematical terms, the meaning of β_1 is then

$$\beta_1 = \ln\left(\frac{odds(X=1)}{odds(X=0)}\right)$$

To understand this equation further, consider first what the odds are. The odds is itself the ratio of two probabilities, p and 1-p. Consider the following table of odds values for various values of p. Note that 9:1 is read '9 to 1.'

Value of *p*	Odds of *p*
0.9	9:1
0.8	4:1
0.6	1.5:1
0.5	1:1
0.4	0.67:1
0.2	0.25:1
0.1	0.11:1

Now, using a simple example from horse racing, if one horse has 8:1 odds of winning and a second horse has 4:1 odds of winning, how do you compare these two horses? One obvious way is to look at the ratio of their odds. The first horse has twice the odds of winning as the second.

Consider a second example of two slow horses whose odds of winning are 0.1:1 and 0.05:1. Here again, their odds ratio is 2. The message here: the odds ratio gives a relative number. Even though the first horse is twice as likely to win as the first, it is still a long shot.

To completely understand β_1, we must take the logarithm of the odds ratio. It is difficult to think in terms of logarithms. However, we can remember that the log of one is zero. So a positive value of β_1 indicates that the odds of the numerator are large while a negative value indicates that the odds of the denominator are larger.

It is probability easiest to think in terms of e^{β_1} rather than β_1, because e^{β_1} is the odds ratio while β_1 is the log of the odds ratio. Both quantities are displayed in the reports.

Multiple Independent Variables

When there are multiple independent variables, the interpretation of each regression coefficient becomes more difficult, especially if interaction terms are included in the model. In general, however, the regression coefficient is interpreted the same as above, except that the caveat ‘holding all other independent variables constant’ must be added. That is, can the value of this independent variable be increased by one without changing any of the other variables. If it can, then the interpretation is as before. If not, then some type of conditional statement must be added that accounts for the values of the other variables.

Multinomial Dependent Variable

When the dependent variable has more than two values, there will be more than one regression equation. In fact, the number of regression equations is equal to one less than the number of values. This makes interpretation more difficult because there are several regression coefficients associated with each independent variable. In this case, care must be taken to understand what each regression equation is predicting. Once this is understood, interpretation of each of the *K*-1 regression coefficients for each variable can proceed as above.

Consider the following example in which there are two independent variables, X1 and X2, and the dependent variable has three groups: *A*, *B*, and *C*.

Row	Y	X1	X2	GA	GB	GC
1	A	3.2	5.8	1	0	0
2	A	4.7	6.1	1	0	0
3	B	2.8	3.5	0	1	0
4	B	3.3	4.6	0	1	0
5	B	3.9	5.2	0	1	0
6	C	4.2	3.7	0	0	1
7	C	7.3	4.4	0	0	1
8	C	5.3	5.1	0	0	1
9	C	6.8	4.5	0	0	1

Look at the three indicator variables: *GA*, *GB*, and *GC*. They are set to one or zero depending on whether *Y* takes on the corresponding value. Two regression equations will be generated corresponding to any two of these indicator variables. The value that is not used is called the *reference value*. Suppose the reference value is *C*. The two regression equations would be

$$\ln\left(\frac{p_A}{p_C}\right) = \beta_{A0} + \beta_{A1}X_1 + \beta_{A2}X_2$$

and

$$\ln\left(\frac{p_B}{p_C}\right) = \beta_{B0} + \beta_{B1}X_1 + \beta_{B2}X_2$$

The two coefficients for *X1* in these equations, β_{A1} and β_{B1}, give the change in the log odds of A versus C and B versus C for a one unit change in *X1*, respectively.

Statistical Tests and Confidence Intervals

Inferences about individual regression coefficients, groups of regression coefficients, goodness-of-fit, mean responses, and predictions of group membership of new observations are all of interest. These inference procedures can be treated by considering hypothesis tests and/or confidence intervals. The inference procedures in logistic regression rely on large sample sizes for accuracy.

Two procedures are available for testing the significance of one or more independent variables in a logistic regression: likelihood ratio tests and Wald tests. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation.

These two testing procedures will be described next.

Likelihood Ratio and Deviance

The *Likelihood Ratio* test statistic is -2 times the difference between the log likelihoods of two models, one of which is a subset of the other. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. The test is named as a ratio rather than a difference since the difference between two log likelihoods is equal to the log of the ratio of the two likelihoods. That is, if L_{full} is the log likelihood of the full model and L_{subset} is the log likelihood of a subset of the full model, the likelihood ratio is defined as

$$\begin{aligned} LR &= -2\left[L_{\text{subset}} - L_{\text{full}}\right] \\ &= -2\left[\ln\left(\frac{l_{\text{subset}}}{l_{\text{full}}}\right)\right] \end{aligned}$$

Note that the -2 adjusts *LR* so the chi-square distribution can be used to approximate its distribution.

The likelihood ratio test is the test of choice in logistic regression. Various simulation studies have shown that it is more accurate than the Wald test in situations with small to moderate sample sizes. In large samples, it performs about the same. Unfortunately, the likelihood ratio test requires more calculations than the Wald test, since it requires that two maximum-likelihood models must be fit.

Deviance

When the full model in the likelihood ratio test statistic is the saturated model, *LR* is referred to as the *deviance*. A saturated model is one which includes all possible terms (including interactions) so that the predicted values from the model equal the original data. The formula for the deviance is

$$D = -2\left[L_{\text{Reduced}} - L_{\text{Saturated}}\right]$$

The deviance may be calculated directly using the formula for the deviance residuals (discussed below). This formula is

$$D = 2\sum_{j=1}^{J}\sum_{g=1}^{G} w_{gj} \ln\left(\frac{w_{gj}}{n_j p_{gj}}\right)$$

This expression may be used to calculate the log likelihood of the saturated model without actually fitting a saturated model. The formula is

$$L_{\text{Saturated}} = L_{\text{Reduced}} + \frac{D}{2}$$

The deviance in logistic regression is analogous to the residual sum of squares in multiple regression. In fact, when the deviance is calculated in multiple regression, it is equal to the sum of the squared residuals. Deviance residuals, to be discussed later, may be squared and summed as an alternative way to calculate the deviance, *D*.

The change in deviance, ΔD, due to excluding (or including) one or more variables is used in logistic regression just as the partial *F* test is used in multiple regression. Many texts use the letter *G* to represent ΔD, but we have already used *G* to represent the number of groups in *Y*. Instead of using the *F* distribution, the distribution of the change in deviance is approximated by the chi-square distribution. Note that since the log likelihood for the saturated model is common to both deviance values, ΔD is calculated without actually estimating the saturated model. This fact becomes very important during subset selection. The formula for ΔD for testing the significance of the regression coefficient(s) associated with the independent variable *X*1 is

$$\begin{aligned}\Delta D_{X1} &= D_{\text{without } X1} - D_{\text{with } X1}\\ &= -2\left[L_{\text{without } X1} - L_{\text{Saturated}}\right] + 2\left[L_{\text{with } X1} - L_{\text{Saturated}}\right]\\ &= -2\left[L_{\text{without } X1} - L_{\text{with } X1}\right]\end{aligned}$$

Note that this formula looks identical to the likelihood ratio statistic. Because of the similarity between the change in deviance test and the likelihood ratio test, their names are often used interchangeably.

Wald Test

The Wald test will be familiar to those who use multiple regression. In multiple regression, the common *t*-test for testing the significance of a particular regression coefficient is a Wald test. In logistic regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$z_j = \frac{b_j}{s_{b_j}}$$

where s_{b_j} is an estimate of the standard error of b_j provided by the square root of the corresponding diagonal element of the covariance matrix, $V\left(\hat{\beta}\right)$.

With large sample sizes, the distribution of z_j is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as 'adequate.'

The Wald test is used in ***NCSS*** to test the statistical significance of individual regression coefficients.

Confidence Intervals

Confidence intervals for the regression coefficients are based on the Wald statistics. The formula for the limits of a $100(1-\alpha)\%$ two-sided confidence interval is

$$b_j \pm |z_{\alpha/2}| s_{b_j}$$

R-Squared

The following discussion summarizes the material on this subject in Hosmer and Lemeshow (1989). In multiple regression, R_M^2 represents the proportion of variation in the dependent variable accounted for by the independent variables. (The subscript "M" emphasizes that this statistic is for multiple regression.) It is the ratio of the regression sum of squares to the total sum of squares. When the residuals from the multiple regression can be assumed to be normally distributed, R_M^2 can be calculated as

$$R_M^2 = \frac{L_p - L_0}{L_0}$$

where L_0 is the log likelihood of the intercept-only model and L_p is the log likelihood of the model that includes the independent variables. Note that L_p varies from L_0 to 0. R_M^2 varies between zero and one.

This quantity has been proposed for use in logistic regression. Unfortunately, when R_L^2 (the R-squared for logistic regression) is calculated using the above formula, it does not necessarily range between zero and one. This is because the maximum value of L_p is not always 0 as it is in multiple regression. Instead, the maximum value of L_p is L_S, the log likelihood of the saturated model. To allow R_L^2 to vary from zero to one, it is calculated as follows

$$R_L^2 = \frac{L_p - L_0}{L_0 - L_S}$$

The introduction of L_S into this formula causes a degree of ambiguity with R_L^2 that does not exist with R_M^2. This ambiguity is due to the fact that the value of L_S depends on the configuration of independent variables. The following example will point out the problem.

Consider a logistic regression problem consisting of a binary dependent variable and a pool of four independent variables. The data for this example are given in the following table.

Y	X1	X2	X3	X4
0	1	1	2.3	5.9
0	1	1	3.6	4.8
1	1	1	4.1	5.6
0	1	2	5.3	4.1
0	1	2	2.8	3.1
1	1	2	1.9	3.7

(**Table Continued)**

Y	X1	X2	X3	X4
1	1	2	2.5	5.4
1	2	1	2.3	2.6
1	2	1	3.9	4.6
0	2	1	5.6	4.9
0	2	2	4.2	5.9
0	2	2	3.8	5.7
0	2	2	3.1	4.5
1	2	2	3.2	5.5
1	2	2	4.5	5.2

Notice that if only X1 and X2 are included in the model, the dataset may be collapsed because of the number of repeats. In this case, the value of L_S will be less than zero. However, if X3 or X4 are used there are no repeats and the value of L_S will be zero. Hence, the denominator of R_L^2 depends on which of the independent variables is used. This is not the case for R_M^2. This ambiguity comes into play especially during subset selection. It means that as you enter and remove independent variables, the target value L_S can change.

Hosmer and Lemeshow (1989) recommend against the use R_L^2 as a goodness of fit measure. However, we have included it in our output because it does provide a comparative measure of the proportion of the log likelihood that is accounted for by the model. Just remember than an R_L^2 value of 1.0 indicates that the logistic regression model achieves the same log likelihood as the saturated model. However, this does not mean that it fits the data perfectly. Instead, it means that it fits the data as well as could be hoped for.

Residual Diagnostics

Residuals are the discrepancies between the data values and the their predicted values from the fitted model. A residual analysis detects outliers, identifies influential observations, and diagnoses the appropriateness of the logistic model. An analysis of the residuals should be conducted before a regression model is used.

Unfortunately, the residuals are more difficult to define in logistic regression than in regular multiple regression because of the nonlinearity of the logistic model and because more than one regression equation is used. The discussion that follows provides an introduction to the residuals that are produced by the logistic regression procedure. Pregibon (1981) presented this material for the case of the two-group logistic regression. Extensions of Pregibon's results to the multiple-group case are provided in an article by Lesaffre and Albert (1989) and in the book by Hosmer and Lemeshow (1989). Lesaffre and Albert provide formulas for these extensions. On the other hand, Hosmer and Lemeshow recommend that individual logistic regressions be run in which the each group is treated separately. Hence, if you have three groups A, B, and C, you would run group A versus groups B and C, group B versus groups A and C, and group C versus groups and A and B. You would conduct a residual analysis for each of these regressions using Pregibon's two-group formulas. In ***NCSS***, we have adopted the approach of Hosmer and Lemeshow.

Data Configuration

When dealing with residuals, it is important to understand the data configuration. Often, residual formulations are presented for the case when each observation has a different combination of values of the independent variables. When some observations have identical independent variables or when you have specified a frequency variable, these observations are combined to form a single row of data. The N original observations are combined to form J unique rows. The response indicator variables y_{gj} for the original observations are replaced by two variables: w_{gj} and n_j. The variable n_j is the total number of observations with this independent variable configuration. The variable w_{gj} is the number of the n_j observations that are in group g.

NCSS automatically collapses the dataset of N observations into a combined dataset of J rows for analysis. The residuals are calculated using this last formula. However, the residuals are reported in the original observation order. Thus, if two identical observations have been combined, the residual is shown for each. If corrective action needs to be taken because a residual is too large, both observations must be deleted. Also, if you want to calculate the deviance or Pearson chi-square from the corresponding residuals, care must be taken that you use only the J collapsed rows, not the N original observations.

Simple Residuals

Each of the g logistic regression equations can be used to estimate the probabilities that an observation of independent variable values given by X_j belongs to the corresponding group. The actual values of these probabilities were defined earlier as

$$\pi_{gj} = \mathrm{Prob}\left(Y = g \mid \mathrm{X}_j\right)$$

The estimated values of these probabilities are called p_{gj}. If the hat symbol is used to represent an estimated parameter, then

$$p_{gj} = \hat{\pi}_{gj}$$

These estimated probabilities can be compared to the actual probabilities occurring in the database by subtracting the two quantities, forming a residual. The actual values were defined as the indicator variables y_{gj}. Thus, simple residuals may be defined as

$$r_{gj} = y_{gj} - p_{gj}$$

Note that, unlike multiple regression, there are g residuals for each observation instead of just one. This makes residual analysis much more difficult. If the logistic regression model fits an observation closely, all of its residuals will be small. Hence, when y_{gj} is one, p_{gj} will be close to one and when y_{gj} is zero, p_{gj} will be close to zero.

Unfortunately, the simple residuals have unequal variance equal to $n_j\pi_{gj}\left(1-\pi_{gj}\right)$, where n_j is the number of observations with the same values of the independent variables as observation j. This unequal variance makes comparisons among the simple residuals difficult and alternative types of residuals are necessary.

Pearson Residuals

One popular alternative to the simple residuals are the *Pearson residuals* which are so named because they give the contribution of each observation to the Pearson chi-square goodness of fit statistic. When the values of the independent variables of each observation are unique, the formula this residual is

$$\chi'_j = \pm\sqrt{\sum_{g=1}^{G}\frac{\left(y_{gj} - p_{gj}\right)^2}{p_{gj}}}, \quad j = 1,2,\cdots,N$$

The negative sign is used when $y_{gj} = 0$ and the positive sign is used when $y_{gj} = 1$.

When some of the observations are duplicates and the database has been collapsed (see Data Configuration above) the formula is

$$\chi_j = \pm\sqrt{\sum_{g=1}^{G}\frac{\left(w_{gj} - n_j p_{gj}\right)^2}{n_j p_{gj}}}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if w_{gj} / n_j is greater (less) than p_{gj}. Note that this is the formula used by ***NCSS***.

By definition, the sum of the squared Pearson residuals is the Pearson chi-square goodness of fit statistics. That is,

$$\chi^2 = \sum_{j=1}^{J}\chi_j^2$$

Deviance Residuals

Remember that the deviance is -2 times the difference between log likelihoods of a reduced model and the saturated model. The deviance is calculated using

$$\begin{aligned} D &= -2\left[L_{\text{Reduced}} - L_{\text{Saturated}}\right] \\ &= -2\left[\sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(p_{gj}\right) - \sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(y_{gj}\right)\right] \\ &= -2\left[\sum_{j=1}^{N}\sum_{g=1}^{G} y_{gj}\ln\left(p_{gj}\right)\right] \\ &= \sum_{j=1}^{N}\left[2\sum_{g=1}^{G} y_{gj}\ln\left(\frac{1}{p_{gj}}\right)\right] \end{aligned}$$

This formula uses the fact that the saturated model reproduces the original data exactly and that, in these sums, the value of 0 ln(0) is defined as 0 and that the ln(1) is also 0.

The deviance residuals are the square roots of the contribution of each observation to the overall deviance. Thus, the formula for the deviance residual is

$$d'_j = \pm\sqrt{2\sum_{g=1}^{G} y_{gj}\ln\left(\frac{1}{p_{gj}}\right)}, j = 1,2,\cdots,N$$

The negative sign is used when $y_{gj} = 0$ and the positive sign is used when $y_{gj} = 1$.

When some of the observations are duplicates and the database has been collapsed (see Data Configuration above) the formula is

$$d_j = \pm\sqrt{2\sum_{g=1}^{G} w_{gj} \ln\left(\frac{w_{gj}}{n_j p_{gj}}\right)}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if $w_{REF(g),j} / n_j$ is greater (less) than $p_{REF(g),j}$. Note that this is the formula used by ***NCSS***.

By definition, the sum of the squared deviance residuals is the deviance. That is,

$$D = \sum_{j=1}^{J} d_j^2$$

Hat Matrix Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. These are often called *leverage* design points. The larger the value of this statistic, the more the observation influences that estimates of the regression coefficients. An observation that is discrepant, but has low leverage, should not cause much concern. However, an observation with a large leverage and a large residual should be checked very carefully. The use of these hat diagonals is discussed further in the multiple regression chapter.

The formula for the hat diagonal associated with the *j*th observation and *g*th group is

$$h_{gj} = n_j p_{gj}\left(1 - p_{gj}\right)\sum_{i=1}^{p}\sum_{k=1}^{p} X_{ij} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix of the regression coefficients associated with the *g*th regression equation. The interpretation of this diagnostic is not as clear in logistic regression as in multiple regression because it involves the predicted values which in turn involve the dependent variable. In multiple regression, the hat diagonals only involve the independent variables.

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

DFBETA

One way to study the impact of an observation on each regression coefficient is to determine how much that coefficient changes when the observation is deleted. The DFBETA statistic is the standardized difference between a regression coefficient before and after the removal of the *j*th observation.

The formula for DFBETA is approximated by

$$\text{DFBETA}_{gij} = \left(\frac{w_{gj} - n_j p_{gj}}{\left(1 - h_{gj}\right)\sqrt{\hat{V}_{gii}}}\right)\sum_{k=1}^{p} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix associated with the *g*th regression equation.

Note that this formula matches Pregibon (1981) in the two-group case, but is different from Lesaffre (1989) in the multi-group case.

Cooks Distance: C and Cbar

C and *Cbar* are extensions of Cooks distance for logistic regression. Quoting from Pregibon (1981), page 719:

"Cbar measures the overall change in fitted logits due to deleting the *l*th observation for all points excluding the one deleted. Conversely, *C* includes the deleted point. Although *C* will usually be the preferred diagnostic to measure overall coefficients changes, in the examples examined to date, the one-step approximations were more accurate for *Cbar* than *C*."

The formulas for *C* and *Cbar* are

$$C_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1-h_{gj}\right)^2}, \quad j = 1,2,\cdots,J$$

$$\overline{C}_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1-h_{gj}\right)}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

DFDEV and DFCHI2

DFDEV and *DFCHI2* are statistics that measure the change in deviance and in Pearson's chi-square, respectively, that occurs when an observation is deleted from the dataset. Large values of these statistics indicate observations that have not been fitted well.

The formulas for these statistics are

$$DFDEV_{gj} = d_j^2 + \overline{C}_{gj}, \quad j = 1,2,\cdots,J$$

$$DFCHI2_{gj} = \frac{\overline{C}_{gj}}{h_{gj}}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

Predicted Probabilities

This section describes how to calculate the predicted probabilities of group membership and associated confidence intervals. Recall that the regression equation is linear when expressed in logit form. That is,

$$\begin{aligned}\ln\left(\frac{p_g}{p_1}\right) &= \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \\ &= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g\end{aligned}$$

The adjustment for the prior probabilities changes the value of the intercepts, so this expression may be simplified to

$$\begin{aligned}\ln\left(\frac{p_g}{p_1}\right) &= \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p \\ &= \mathrm{XB}_g\end{aligned}$$

if we assume that the intercepts have been appropriately adjusted. Assuming that the estimated matrix of regression coefficients is distributed asymptotically as a multivariate normal, the point estimates of this quantity for a specific set of X values is given by

$$l_j = \ln\left(\frac{p_g}{p_1} \mid \mathrm{X}_j\right) = \mathrm{X}_\mathrm{j}\hat{B}_g$$

and the corresponding confidence interval is given by

$$l_j \pm z_{\alpha/2}\left(\mathrm{X}_\mathrm{j}'\mathrm{V}_\mathrm{g}\mathrm{X}_\mathrm{j}\right)$$

where V_g is that portion of the covariance matrix $V\left(\hat{B}\right)$ that deals with the gth regression equation.

When there are only two groups, these confidence limits can be inverted to give confidence limits on the predicted probabilities as

$$\frac{1}{1+e^{\mathrm{X}_\mathrm{j}\hat{\mathrm{B}} \pm z_{\alpha/2}\sigma_B}} \text{ and } \frac{e^{\mathrm{X}_\mathrm{j}\hat{\mathrm{B}} \pm z_{\alpha/2}\sigma_B}}{1+e^{\mathrm{X}_\mathrm{j}\hat{\mathrm{B}} \pm z_{\alpha/2}\sigma_B}}$$

where

$$\sigma_B = \mathrm{X}_\mathrm{j}'\mathrm{V}_\mathrm{g}\mathrm{X}_\mathrm{j}$$

When there are more than two groups, the confidence limits on the logits are still given by

$$l_j \pm z_{\alpha/2}\left(\mathrm{X}_\mathrm{j}'\mathrm{V}_\mathrm{g}\mathrm{X}_\mathrm{j}\right)$$

However, this set of confidence limits of the logits cannot be inverted to give confidence limits for the predicted probabilities. We have found no presentation that gives an appropriate set of confidence limits. In order to provide an approximate answer, we provide approximate confidence limits by applying the inversion as if there were only two groups. This method ignores the correlation between the coefficients of the individual equations. However, we hope that it provides a useful approximation to the confidence intervals.

Subset Selection

Subset selection refers to the task of finding a small subset of the available independent variables that does a good job of predicting the dependent variable. Because logistic regression must be solved iteratively, the task of finding the best subset can be very time consuming. Hence, techniques that search all possible combinations of the independent variables are not feasible. Instead, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Before discussing the details of these two algorithms, it is important to comment on a couple of issues that can come up. First of all, since there is more than one regression equation when there are more than two categories in the dependent variable, it is possible that a variable is important in one of the equations and not in the others. The algorithms presented here are based on the overall likelihood. This means that if an independent variable is important in at least one of the regression equations, it will be kept.

A second issue is what to do with the individual-degree of freedom variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms search on model terms rather than on the individual binary variables. Thus, the whole set of binary variables associated with a given term are considered together for inclusion in, or deletion from, the model. It is all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and designate them as Numeric Variables.

Hierarchical Models

A third issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to consider only hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of the log likelihood. Enter this term into the model.
3. Continue adding terms until a target value for the log-likelihood is achieved or until a preset limit on the maximum number of terms in the model is reached. Note that these terms can be limited to those keeping the model hierarchical.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations and terms so that other, more time consuming, methods are not feasible.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of the log likelihood. If a switch can be found, it is made and the pool of terms is again searched to determine if another switch can be made. Note that this switching can be limited to those keeping the model hierarchical.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some F-to-enter and F-to-remove tests whose properties are not well understood to begin with.

Data Structure

The data given below are the first few rows of a set of data about leukemia patients published in Lee (1980). The response variable is whether leukemia remission occurred (REMISS). The independent variables are cellularity of the marrow clot section (CELL), smear differential percentage of blasts (SMEAR), percentage of absolute marrow leukemia cell infiltrate (INFIL), percentage labeling index of the bone marrow leukemia cells (LI), absolute number of blasts in the peripheral blood (BLAST), and the highest temperature prior to start of treatment (TEMP). This dataset is stored in the LEUKEMIA database in the Data directory.

LEUKEMIA dataset (subset)

REMISS	CELL	SMEAR	INFIL	LI	BLAST	TEMP
1	80	83	66	190	11.6	996
1	90	36	32	140	4.5	992
0	80	88	70	80	0.5	982
0	100	87	87	70	10.3	986
1	90	75	68	130	2.3	980
0	100	65	65	60	2.3	982
1	95	97	92	100	16.0	992
0	95	87	83	190	21.6	1020

Missing Values

If missing values are found in any of the independent variables being used, the row is omitted. If only the dependent variable is missing, the row will not be used in the formation of the coefficient estimates, but a predicted value will be generated for that row.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Group Variable

This discrete variable identifies the group to which each observation belongs. Values may be text or numeric. The variable should have only a few unique values, such as 0 and 1, Yes and No, or A, B, and C. When there are only two unique values, the analysis is called *logistic regression*. When there are three or more unique values, the analysis is called *multiple-group logistic regression*.

In a logistic regression with G groups, only G-1 logistic regression equations are needed. The group for which a regression equation is not created is called the *reference group*. This group is often the baseline group. In the analysis, the other groups are compared to this group.

The reference group may be designated within parentheses after the name of the variable; otherwise, the reference group is determined by the Default Reference Group setting. For example, suppose the group variable, CATEGORY, has three values: A, B, and C.

1. To designate A as the reference group, enter 'CATEGORY(A)' or change Default Reference Group to 'First Group after Sorting'.
2. To designate B as the reference group, enter 'CATEGORY(B)'.
3. To designate C as the reference group, enter 'CATEGORY(C)' or change Default Reference Group to 'Last Group after Sorting'.

Default Reference Group

This option specifies the default reference group for the logistic regression. The reference group is the group for which a regression equation is not created. In a logistic regression with G groups, only G-1 logistic regression equations are needed. This group is often the baseline group.

- **First Group after Sorting**

  Use the first group in alpha-numeric sorted order as the reference group.

- **Last Group after Sorting**

  Use the last group in alpha-numeric sorted order as the reference group.

The reference group may also be designated within parentheses after the name of the Y: Group Variable name, in which case the default reference group is ignored. Suppose the group variable, CATEGORY, has four values: A, B, C, and D.

1. If this option is set to 'First Group after Sorting' and the group variable is entered as 'CATEGORY', the reference group would be A.
2. If this option is set to 'Last Group after Sorting' and the group variable is entered as 'CATEGORY', the reference group would be D.

3. If the group variable is entered as 'CATEGORY(B)', the choice for this setting would be ignored, and the reference value would be B.

Frequency Variable

Frequency Variable

Specify an optional frequency (count) variable. This variable contains integers that represent the number of observations (or frequency) associated with each observation.

If left blank, each observation has a frequency of one. This variable lets you modify that frequency. This is especially useful when your data are already tabulated and you want to enter the counts.

Numeric Independent Variables

X’s: Numeric Independent Variables

Specify the numeric (continuous) independent variables. By numeric, we mean that the values are numeric and at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are better analyzed when you specify them as Categorical Independent Variables.

If you want to create powers and cross-products of these variables, specify an appropriate model in the ‘Custom Model’ field under the Model tab.

If you want to create predicted values of *Y* for values of *X* not in your database, add the *X* values to the bottom of the database. They will not be used during estimation, but predicted values will be generated for them.

Categorical Independent Variables

X’s: Categorical Independent Variable(s)

Specify categorical (nominal) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories.

The values in a categorical variable are not used directly in the regression analysis. Instead, a set of numeric variables is substituted for them. Suppose a categorical variable has *G* categories. ***NCSS*** automatically generates the *G*-1 indicator variables that are needed for the analysis. The type of indicator variable created is determined by the selection for the *Default Reference Value* and the *Default Contrast Type*. The type of indicator created can also be controlled by entering the reference value and contrast type directly according to the syntax below. See the Default Reference Value and Default Contrast Type sections below for a discussion of the reference value and contrast type options.

You can create the interactions among these variables automatically using the *Custom Model* field under the Model tab.

Syntax

The syntax for specifying a categorical variable is *VarName*(*RefValue*;*CType*) where *VarName* is the name of the variable, *RefValue* is the reference value, and *CType* is the type of numeric variables generated: B for binary, P for polynomial, R for contrast with the reference value, and S for a standard set of contrasts.

For example, suppose a categorical variable, STATE, has four values: Texas, California, Florida, and New York. To process this variable, the values are arranged in sorted order: California, Florida, New York, and Texas. Next, the reference value is selected. If a reference value is not specified, the default value specified in the *Default Reference Value* window is used. Finally, the method of generating numeric variables is selected. If such a method is not specified, the contrast type selected in the *Default Contrast Type* window is used. Possible ways of specifying this variable are

STATE	**RefValue = Default, CType = Default**
STATE(New York)	**RefValue = New York, CType = Default**
STATE(California;R)	**RefValue = California, CType = Contrast with Reference**
STATE(Texas;S)	**RefValue = Texas, CType = Standard Set**

More than one category variable may be designated using a list. Examples of specifying three variables with various options are shown next.

```
STATE  BLOODTYPE  GENDER
STATE(California;R)  BLOODTYPE(O)  GENDER(F)
STATE(Texas;S)  BLOODTYPE(O;R)  GENDER(F;B)
```

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting**

  Use the first value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**

  Use the last value in alpha-numeric sorted order as the reference value.

The reference value may also be designated within parentheses after the name of the categorical independent variable, in which case the default reference value is ignored. For example, suppose that the categorical independent variable, STATE, has four values: 1, 3, 4, and 5.

1. If this option is set to 'First Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 1.
2. If this option is set to 'Last Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 5.
3. If the categorical independent variable is entered as 'STATE(4)', the choice for this setting would be ignored, and the reference value would be 4.

Default Contrast Type

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to something other than 'Binary'.

- **Binary (This is the default)**

  Categories are converted to numbers using a set of binary indicator variables by assigning a '1' to the active category and a '0' to all other values. For example, suppose a categorical variable has G categories. ***NCSS*** automatically generates the G-1 binary (indicator) variables that are used in the regression. These indicator variables are set to 1 for those rows in which the value of this variable is equal to a certain value. They are set to 0 otherwise. The G-1 occurs because the Gth indicator variable is redundant (when all G-1 indicators are 0, wIfe know that the Gth indicator variable would be a 1). The value that is skipped is called the Reference Value.

  If your model includes interactions, using the binary indicator type may cause strange results.

  For the STATE variable, three binary variables would be generated. Suppose that the *Default Contrast Type* was 'Binary' and the statement used was 'STATE(Florida)'. The categories would be converted to numbers as follows:

STATE	B1	B2	B3
California	1	0	0
Florida	0	0	0
New York	0	1	0
Texas	0	0	1

- **Contrast with Reference**

  Categories are converted to numbers using a set of contrast variables by assigning a '1' to the active category, a '-1' to the reference value, and a '0' to all other values. A separate contrast is generated for each value other than the reference value.

  For the STATE variable, three numeric variables would be generated. Suppose the *Default Contrast Type* was 'Contrast with Reference', the *Default Reference Type* was 'Last Value after Sorting', and the variable was entered as 'STATE'. The categories would be converted to numbers as follows:

STATE	R1	R2	R3
California	1	0	0
Florida	0	1	0
New York	0	0	1
Texas	-1	-1	-1

- **Polynomial**

  If a variable has five or fewer categories, it can be converted to a set of polynomial contrast variables that account for the linear, quadratic, cubic, quartic, and quintic relationships. Note that these assignments are made after the values are sorted. Usually, the polynomial method is used on a variable for which the categories represent the actual values. That is, the values themselves are ordinal, not just category identifiers. Also, it is assumed that these values are equally spaced. Note that with this method, the reference value is ignored.

  For the STATE variable, linear, quadratic, and cubic variables are generated. Suppose that the *Default Contrast Type* was 'Polynomial' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	Linear	Quadratic	Cubic
California	-3	1	-1
Florida	-1	-1	3
New York	1	-1	-3
Texas	3	1	1

- **Standard Set**

  A variable can be converted to a set of contrast variables using a standard set of contrasts. This set is formed by comparing each value with those below it. Those above it are ignored. Note that these assignments are made after the values are sorted. The reference value is ignored.

  For the STATE variable, three numeric variables are generated. Suppose that the *Default Contrast Type* was 'Standard Set' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	S1	S2	S3
California	-3	0	0
Florida	1	-2	0
New York	1	1	-1
Texas	1	1	1

Validation

Validation Variable

This variable allows you to validate your logistic regression equations by forcing some observations to be ignored during the estimation phase and then predicted during the classification phase. This provides independent verification of your results.

The values in this variable determine whether the observation is used during the estimation of the logistic regression. If the value of this variable is one, the observation is used in estimating the logistic regression coefficients. If the value of this variable is zero, this observation is not used during the estimation phase. However, it is used during the validation run in which the estimated regression equations are used to classify these observations. The results are displayed in the Classification of Validation Data report.

Alpha Level

Alpha Level

This is the alpha level used in the confidence limits of the odds ratios.

Model Tab

These options control the logistic regression model.

Subset Selection

Subset Selection

This option specifies the subset selection algorithm used to reduce the number of independent variables used in the regression model. Note that since the solution algorithm is iterative, the selection process can be very time consuming. The Forward algorithm is much quicker than the Forward with Switching algorithm, but the Forward algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the generated individual binary variables. That is, either all numeric variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **None**

  No subset selection is attempted. All specified independent variables are used in the logistic regression equation.

- **(Hierarchical) Forward**

  With this algorithm, the term with the largest log likelihood is entered into the model. Next, the term that increases the log likelihood the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

  When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the log likelihood does not change significantly.

- **(Hierarchical) Forward with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest log likelihood is entered into the regression model. The term which increases the log likelihood the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, the likelihood function is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in the log likelihood. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Max Terms in Subset

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the Logistic Regression procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of the log likelihood.

Note that the intercept is counted in this number.

Estimation Options

The following options are used during the likelihood maximization process.

Maximum Iterations

The value specifies the maximum number of iterations allowed during the iteration procedure. If this number is reached, the procedure is terminated prematurely. Usually, no more than ten iterations are necessary for the algorithm to converge. If you reach this maximum before normal convergence occurs, you should try doubling this number. If the algorithm still does not converge before this maximum is reached, you should try omitting (or adding) other independent variables.

This value is used to prevent an infinite loop.

Iteration Termination

Unless the Maximum Iteration limit is reached, the maximum likelihood algorithm continues iterating until the relative change in the log likelihood from one step to the next is less than this amount. The smaller it is, the larger the average number of iterations that will be needed to solve the maximum likelihood equations.

Prior Probabilities

The prior probabilities are your estimates of the probabilities that a new individual falls in each group. Among other things, this value will change the estimated intercept(s).

- **Equal Priors**

  If this option is left blank, the prior probabilities of group membership are assumed equal and only the data values are used in the classification process.

- **Numeric List**

  Blanks or commas are used to separate the numbers in the list that represents the prior probabilities of group membership. You do not have to enter decimal points since the numbers you enter will be scaled so that they sum to one. For example, you could enter '4 4 2' or '2 2 1' when you have three groups whose population proportions are known to be 0.4, 0.4, and 0.2, respectively. Care must be taken that the number of entries matches the number of groups.

- **Ni/N**

  Enter 'Ni/N' when you want the priors to be estimated from group frequencies in the dataset. For example, say you have samples of 50, 100, and 250 from three groups and you select this option. The estimated priors would be 50/400=0.125, 100/400=0.25, and 250/400=0.625.

Model Specification

Which Model Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a straight-forward logistic regression model, select Up to 1-Way.

The options are:

- **Full Model**

  The complete, saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables).

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

  Note that the discussion of the Custom Model option discusses the interpretation of this model.

- **Up to 1-Way**

  This option generates a model in which each variable is represented by a single model term. No cross-products or interaction terms are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

  This is the option to select when you want to analyze the independent variables specified without adding any other terms.

For example, if you have three independent variables A, B, and C, this would generate the model:

A + B + C

- **Up to 2-Way**

  This option specifies that all main effects and two-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C

- **Up to 3-Way**

  All main effects, two-way interactions, and three-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

- **Up to 4-Way**

  All main effects, two-way interactions, three-way interactions, and four-way interactions are included in the model. For example, if you have four independent variables A, B, C, and D, this would generate the model:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D

- **Custom Model**

  The model specified in the *Custom Model* box is used.

Include Intercept

Check this box to include the intercept (B0) in the model. Under most circumstances, you will want to include an intercept in your model.

Write Model in Custom Model Field

When this option is checked, no data analysis is performed when the procedure is run. Instead, a copy of the full model is stored in the Custom Model box. You can then edit the model as desired. This option is useful when you have several variables and you want to be selective about which terms are used.

Note that the program will not do any calculations while this option is checked.

Model Specification – Custom Model

Max Term Order

This option specifies that maximum number of variables that can occur in an interaction term in a custom model. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would be excluded from the model.

This option is particularly useful when used with the bar notation of a custom model to allow a simple way to remove unwanted high-order interactions.

Custom Model

This options specifies a custom model. It is only used when the *Which Model Terms* option is set to *Custom Model*. A custom model specifies the terms (single variables and interactions) that are to be kept in the model.

Interactions

An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of indicator variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the indicator variables generated from the categorical variable.

Syntax

A model is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A + B + C + A*B + A*C + B*C + A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B + A*C.

Some examples will help to indicate how the model syntax works:

A|B = A + B + A*B

A|B A*A B*B = A + B + A*B + A*A + B*B

Note that you should only repeat numeric variables. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A + B + A*A + A*B + B*B

A|B|C = A + B + C + A*B + A*C + B*C + A*B*C

(A + B)*(C + D) = A*C + A*D + B*C + B*D

(A + B)|C = (A + B) + C + (A + B)*C = A + B + C + A*C + B*C

Reports Tab

The following options control which reports are displayed.

Select Reports – Summaries

Run Summary and Response Analysis

Each of these options specifies whether the indicated report is calculated and displayed.

Select Reports – Subset Selection

Subset Selection - Summary and Subset Selection - Detail

Indicate whether to display these subset selection reports.

Select Reports – Estimation

Parameter Significance Tests ... Write Estimated Model

Indicate whether to display these estimation reports.

Select Reports – Goodness-of-Fit

Analysis of Deviance and Log-Likelihood / R-Squared

Indicate whether to display these model goodness-of-fit reports.

Select Reports – Classification

Classification Matrix ... ROC Report

Indicate whether to display these classification reports.

Select Reports – Row-by-Row Lists

Row Classification Report ... Simple Residuals Report

This option specifies which rows, if any, are displayed on the row classification, row probabilities, and simple residuals reports. When you have a lot of data, you may wish to limit this report to only those rows that were classified incorrectly.

Note that Unused Rows are those that were not used during the parameter estimation phase. However, group probabilities are still generated for these rows.

Residuals ... Residual Diagnostics

Indicate whether to display these list reports. Note that since these reports provide results for each row, they may be too long for normal use when requested on large databases.

Select Plots

Y vs X Plot ... Pr(Correct) vs Cutoff Plot

Indicate whether to display these plots.

Format Tab

The following options control the format of the reports.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also, note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option applies to the Group Variable. It lets you select whether to display data values, value labels, or both. Use this option if you want the output to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying Value Labels elsewhere in this manual.

Skip Line After

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Report Options – Decimal Places

Probability ... DFBeta Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'All' to display all digits available. The number of digits displayed by this option is controlled by whether the PRECISION option is SINGLE or DOUBLE.

Report Options – ROC Curves and Prob(Correct) vs Cutoff Plot Options

Number Cutoffs

The probability range (0 to 1) is divided into this many cutoff points and a point for the ROC curve is generated for each. To accurately compute the area under the ROC curve a value of at least 29 should be used here. Values ending in 9, such as 19, 29, or 39, provide the best scales of the PC plot.

Y vs X, Residual vs X, ROC, and Prob vs Cutoff Plot Tabs

These options control the attributes of the various plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Show Row Number

Specify whether to display each point's row number next to the symbol on the plot.

Skip Reference Group (Residual vs X Plots only)

When checked, the residuals associated with the equation for the reference group are not displayed on the residual plots, since they are redundant and tend to clutter the plot. This option is most useful when the dependent variable has only two groups.

Plot Settings – Legend

Show Legend

Specify whether to display the legend.

Legend Text

Specify the text of the legend title. The characters {G} are replaced with an appropriate legend title, such as the group variable name.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Symbols/Lines Tab

These options specify the attributes of the symbols and lines used for each group in the various plots.

Plotting Symbols

Symbol 1 – 15

These options specify the attributes of the symbols used in the plots. The first symbol is used by the first group, the second symbol by the second group, and so on.

Clicking on a symbol box (or the small button to the right of the line box) will bring up a window that allows the attributes to be changed.

Plotting Lines

Line 1 – 15

These options specify the color, width, and pattern of the lines used in the plots. The first line is used by the first group, the second line by the second group, and so on. These line attributes are provided to allow the various groups to be indicated on black-and-white printers.

Clicking on a line box (or the small button to the right of the line box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**
  No data are stored even if some of the storage items are checked.

- **Store in empty columns**
  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in all columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful.

Data Storage Options – Select Items to Store on the Spreadsheet

Expanded X Values ... Covariance Matrix

Indicate whether to store these row-by-row values, beginning at the variable indicated by the *Store First Variable In* option. Note that several of these values include a different value for each group and so they require several columns when they are stored.

Expanded X Values

This option refers to the experimental design matrix. They include all binary and interaction variables generated.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Logistic Regression Analysis

This section presents an introductory example of how to run a logistic regression analysis. The data used are stored in the LEUKEMIA database. In this analysis, a logistic regression will be run to determine the relationship between CELL, LI, and TEMP on the binary dependent variable REMISS.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Logistic Regression window.

1 Open the LEUKEMIA dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LEUKEMIA.S0**.
- Click **Open**.

2 Open the Logistic Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Logistic Regression window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **REMISS** from the list of variables and then click **Ok**. “REMISS” will appear in the Y: Group Variable box.
- Click on the **Default Reference Group** box and select **Last Group after Sorting**.
- Double-click in the **X’s: Numeric Independent Variables** box. This will bring up the variable selection window.
- Select **CELL, LI, TEMP** from the list of variables and then click **Ok**. “CELL,LI,TEMP” will appear in the X’s: Numeric Independent Variables box. Remember to use the Ctrl key to select non-contiguous variables from the list.

4 Specify the reports.

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **All Rows**.
- Check **all reports except Subset Selection - Summary**, **Subset Selection - Detail**, and **Validation Matrix**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Run Summary Section

```
Parameter                  Value              Parameter              Value
Dependent Variable         REMISS             Rows Processed         29
Reference Group            1                  Rows Used              27
Number of Groups           2                  Rows for Validation    0
Frequency Variable         None               Rows X's Missing       0
Numeric Ind. Variables     3                  Rows Freq Miss. or 0   0
Categorical Ind. Variables 0                  Rows Prediction Only   2
Final Log Likelihood       -10.97669          Unique Row Patterns    28
Model R-Squared            0.36130            Sum of Frequencies     27
Actual Convergence         2.94261E-07        Likelihood Iterations  7
Target Convergence         0.000001           Maximum Iterations     20
Model D.F.                 4                  Max Like Message       Normal Completion
Model                      CELL|LI|TEMP
```

This report provides useful information about the reports to follow. It should be studied to make sure that the data were read in properly and that the logistic regression procedure terminated normally. We will only discuss those parameters that need special explanation.

Reference Group

The reference group is that category of the dependent variable that is defined implicitly in terms of the other categories. This is the category that is skipped on much of the output. If you did not specify the reference group with the Y Variable, the reference group is chosen according to the 'Default Reference Group' setting. This value is critical to interpretation of the rest of the output.

Number of Groups

This is the number of unique categories that were found for the dependent variable. Check this count to make certain it agrees with what you anticipated.

Final Log Likelihood

This is the log likelihood of the model that is reported on here.

Model R-Squared

This is the *R*-Squared that was achieved by your regression. Read the discussion of R-Squared that was given earlier to better understand how to interpret R-Squared in the case of logistic regression.

Actual and Target Convergence

The Target Convergence is the amount that is used to stop the iterative fitting of the maximum likelihood algorithm. If the Actual Convergence amount is larger than the Target amount, the algorithm ended before converging and care must be taken in using any of the results. If this happens, the usual remedy is to increase the maximum number of iterations. If this does not solve the problem, you will have to change the variables in the model.

Rows Processed, Used, etc.

These values record how many of each type of observation were encountered when the database was read. You should make sure that these amounts are what you expect.

Unique Row Patterns

This gives the number of unique patterns found in the variables. Both the dependent and independent variables are considered in forming this count.

Likelihood and Maximum Iterations

The Likelihood Iterations are the number of iterations necessary to solve the likelihood equations. Usually, fewer than ten iterations are necessary. If the number of Likelihood Iterations is equal to the Maximum Iterations, the maximum likelihood algorithm did not converge and you should take some remedial action such as increasing the Maximum Iterations or changing the regression model.

Max Like Message

This is the message that was returned when the maximum likelihood algorithm ended. Unless the message “Normal Completion” is received, you should take appropriate corrective action.

Model D.F.

This is the number of degrees of freedom in the *G*-1 logistic regression models.

Model

This is an abbreviated representation of the regression model that was fit to the data.

Response Analysis Section

```
REMISS                      Unique                 Act vs Pred   % Correctly
Categories      Count         Rows       Prior       R-Squared    Classified
0                  18           17     0.50000         0.17842        83.333
1                   9            9     0.50000         0.32704        77.778
Total              27           26                                    81.481
```

This report describes the dependent variable. Use it to understand the dependent variable and how well the regression model approximates it.

Categories

These are the unique values found for the dependent variable. Check to make sure that no unexpected categories were found.

Count

This is the sum of the frequencies (counts) for each category of the dependent variable.

Unique Rows

This is the number of unique rows in each category as determined by the values of the independent variables.

Prior

This is the prior probability of each category as given by the user in the Prior Probabilities option box.

Act vs Pred R-Squared

This is the R-Squared that is achieved when the indicator variable for this category is regressed on the predicted probability of being in this category.

% Correctly Classified

This is the percent of the observations from this category that were correctly classified as such by the multinomial logistic regression model.

Parameter Significance Tests Section

```
Parameter Significance Tests Section (Reference Group: REMISS = 1)
               Regression                          Wald         Wald          Odds
               Coefficient      Standard        Z-Value         Prob         Ratio
Parameter      (B or Beta)         Error         (Beta=0)       Level       Exp(B)
B0: Intercept    -68.32696      56.88604          -1.201     0.22970       0.00000
B1: CELL          -9.65213       7.75107          -1.245     0.21303       0.00006
B2: LI            -3.86710       1.77828          -2.175     0.02966       0.02092
B3: TEMP          82.07365      61.71233           1.330     0.18354        10000+
```

This report gives the estimated logistic regression equation and associated significance tests. The reference group of the dependent variable is shown in the title. If the dependent variable has more than two categories, the appropriate information is displayed for each of the *G*-1 equations.

Parameter

This is the variable from the model that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *B2: GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the Skip Line After option of the Format tab.

Regression Coefficient (B or Beta)

This is the estimated value of the corresponding regression coefficient, sometimes referred to as B or Beta. The interpretation of the regression coefficients is difficult. We refer you to the discussion given at that beginning of this chapter for more details.

Standard Error

This is s_{b_j} , the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is used as the denominator of the Wald test.

Wald Z-Value (Beta=0)

This is the *z* value of the Wald test used for testing the hypothesis that $\beta_{gj} = 0$ against the alternative $\beta_{gj} \neq 0$. The Wald test is calculated using the formula

$$z_{gj} = \frac{b_{gj}}{s_{b_{gj}}}$$

The distribution of the Wald statistic is closely approximated by the normal distribution in large samples. However, in small samples, the normal approximation may be poor. For small samples, the deviance tests should be used instead to test significance since they perform better.

One problem that occurs in multiple-group logistic regression is that the test may be significant for the regression coefficient associated with one category, but not for the same coefficient associated with another category. In this case, we recommend that the independent variable be kept in the model if it is significant in at least one of the *G*-1 regression equations.

Wald Prob Level

This is the significance level of the Wald test. If this value is less than some predefined alpha level, say 0.05, the variable is said to be statistically significant. Otherwise, the variable is not significant.

Odds Ratio Exp(B)

This is the estimated odds ratio associated with this regression coefficient. It is only useful for binary independent variables in which the two values are zero and one. These are the values that are generated for categorical independent variables. The formula used is

$$OR = e^b$$

Because of formatting limitations, the value is not displayed if it is larger than 10000.

Parameter Confidence Limits Section

```
Parameter Confidence Limits Section (Reference Group: REMISS = 1)
             Regression               Lower 95%   Upper 95%        Odds
             Coefficient   Standard  Confidence  Confidence       Ratio
Parameter    (B or Beta)      Error       Limit       Limit      Exp(B)
B0: Intercept  -68.32696   56.88604  -179.82155    43.16763     0.00000
B1: CELL        -9.65213    7.75107   -24.84394     5.53968     0.00006
B2: LI          -3.86710    1.77828    -7.35245    -0.38174     0.02092
B3: TEMP        82.07365   61.71233   -38.88029   203.02760      10000+
```

This report gives the estimated logistic regression equation and associated confidence limits. The reference group of the dependent variable is shown in the title. If the dependent variable has more than two categories, the information is displayed for each of the *G*-1 equations.

Parameter

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *B2: GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the Skip Line After option of the Format tab.

Regression Coefficient (B or Beta)

This is the estimated value of the regression coefficient, sometimes referred to as B or Beta. The interpretation of the regression coefficients is difficult. We refer you to the discussion given at that beginning of this chapter for more details.

Standard Error

This is s_{b_j}, the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is used as the denominator of the Wald test.

Confidence Limits

These are the lower and upper confidences limits for β_{gj} based on the Wald statistic. These confidence limits are use the formula

$$b_{gj} \pm z_{1-\alpha/2} s_{b_{gj}}$$

Since they are based on the Wald test, they are only valid for large samples.

Odds Ratio Exp(B)

This is the estimated odds ratio associated with this regression coefficient. It is only useful for binary independent variables in which the two values are zero and one. These are the values that are generated for categorical independent variables. The formula used is

$$OR = e^{b}$$

Because of formatting limitations, the value is not displayed if it is larger than 10000.

Odds Ratio Estimation Section

```
Odds Ratios Section (Reference Group: REMISS = 1)
                Regression         Odds     Lower 95%      Upper 95%
                Coefficient        Ratio   Confidence     Confidence
Parameter       (B or Beta)       Exp(B)        Limit          Limit
B0: Intercept     -68.32696      0.00000      0.00000         10000+
B1: CELL           -9.65213      0.00006      0.00000      254.59635
B2: LI             -3.86710      0.02092      0.00064        0.68267
B3: TEMP           82.07365       10000+      0.00000         10000+
```

This report presents estimates of the odds ratios and associated confidence limits associated with each variable in the model.

Parameter

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *B2: GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the Skip Line After option of the Format tab.

This is the estimated value of the corresponding regression coefficient, sometimes referred to as B or Beta. The interpretation of the regression coefficients is difficult. We refer you to the discussion given at that beginning of this chapter for more details.

Odds Ratio Exp(B)

This is the estimated odds ratio associated with this regression coefficient. It is only useful for binary independent variables in which the two values are zero and one. These are the values that are generated for categorical independent variables. The formula used is

$$OR = e^{b}$$

Because of formatting limitations, the value is not displayed if it is larger than 10000.

Confidence Limits

The lower and upper confidence limits yield an interval estimate of the odds ratio. The confidence coefficient is one minus alpha. Thus, when alpha is 0.05, the confidence coefficient is 0.95 or 95%. The formula used is

$$e^{(b_i \pm z_{1-\alpha/2} s_{b_i})}$$

Since these confidence limits are based on Wald statistics, they are only valid for large samples.

Estimated Logistic Regression Model(s)

```
Model For REMISS = 0
-68.3269603055054 -9.65212973757993*CELL -3.86709587172716*LI + 82.0736535775605*TEMP

Note that each model estimates B for a specific group, where Logit(Y) = XB.
To calculate a probability, transform the logit using Prob(Y=group) = 1/(1+Exp(-XB))
or Prob(Y<>group) = Exp(-XB)/(1+Exp(-XB)).

Transformation Note:
Regular transformations must be less the 255 characters. If this expression is longer the 255 characters,
copy this expression and paste it into a text file, then use the transformation FILE(filename.txt)
access the text file.
```

This report gives the logistic regression model in a regular text format that can be used as a transformation formula. A separate model is displayed for each of the *G*-1 categories of the dependent variable. The regression coefficients are displayed in double precision because a single-precision formula does not include the accuracy necessary to calculate the scores (logits) and predicted probabilities.

Note that a transformation must be less than 255 characters. Since these formulas are often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file, including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

Analysis of Deviance Section

```
                                              Increase
                                            From Model
Term                                          Deviance          Prob
Omitted          DF          Deviance     (Chi Square)         Level
All               3          34.37177         12.41839       0.00608
CELL              1          24.64782          2.69445       0.10070
LI                1          30.82856          8.87518       0.00289
TEMP              1          24.34072          2.38734       0.12232
None(Model)       3          21.95337
```

This report is the logistic regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms in the regression are statistically significant after adjusting for all other terms in the model.

This report is not produced during a subset selection run.

Note that this report requires that a separate logistic regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.

The “All” line refers to the intercept-only model. This line tests the significance of the full model. The “None(Model)” refers to the complete model with no terms removed.

Note that it is usually not advisable to include an interaction term in a model when one of the associated main effects is missing—which is what happens here. However, in this case, we believe this to be a useful test.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the chi-square test displayed on this line. DF is equal to (*G*-1)*DFt* where *DFt* is the degrees of freedom of the term.

Deviance

The deviance is equal to minus two times the log likelihood achieved by the model being described on this line of the report. See the discussion given earlier in this chapter for a technical discussion of the deviance. A useful way to interpret the deviance is as the analog of the residual sum of squares in multiple regression. This value is used to create the difference in deviance that is used in the chi-square test.

Increase From Model Deviance (Chi Square)

This is the difference between the deviance for the model described on this line and the deviance of the complete model. This value follows the chi-square distribution in medium to large samples. See the discussion given earlier in this chapter for a technical discussion of this value. This value can be thought of as the analog of the residual sum of squares in multiple regression. Thus, you can think of this value as the increase in the residual sum of squares that occurs when this term is removed from the model.

Another way to interpret this test is as a redundancy test because it tests whether this term is redundant after considering all of the other terms in the model.

Note that the first line gives a test for the whole model.

Prob Level

This is the significance level of the chi-square test. This is the probability that a chi-square value with degrees of freedom DF is equal to this value or greater. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

Log Likelihood & R-Squared Section

```
                                            R-Squared    Reduction       Reduction
Term(s)                             Log  Of Remaining   From Model  From Saturated
Omitted              DF      Likelihood       Term(s)    R-Squared       R-Squared
All                   1       -17.18588       0.00000
CELL                  1       -12.32391       0.28290      0.07839         0.71710
LI                    1       -15.41428       0.10308      0.25821         0.89692
TEMP                  1       -12.17036       0.29184      0.06946         0.70816
None(Model)           3       -10.97669       0.36130      0.00000         0.63870
None(Saturated)      28         0.00000       1.00000                      0.00000
```

This report provides the log likelihoods and R-squared values of various models. This report is not produced during a subset selection run.

Note that this report requires that a separate logistic regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the term that is omitted from the model. The “All” line refers to the intercept-only model. The “None(Model)” refers to the complete model with no terms removed. The “None(Saturated)” line gives the results for the saturated model.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the term displayed on this line. DF is equal to $(G\text{-}1)DFt$ where DFt is the degrees of freedom of the term.

Log Likelihood

This is the log likelihood of the model displayed on this line. Note that this is the log likelihood of the logistic regression without the term listed.

R-Squared of Remaining Term(s)

This is the *R*-squared of the model displayed on this line, R_L^2. Note that the model does not include the term listed at the beginning of the line.

This *R*-squared is analogous to the *R*-squared in multiple regression, but it is not the same. This value is discussed in detail under the heading *R*-Squared above. Refer to that section for more details about this statistic. We repeat the summary of the interpretation of *R*-squared in logistic regression.

Hosmer and Lemeshow (1989) recommend against the use R_L^2 as a goodness of fit measure. However, we have included it in our output because it does provide a comparative measure of the proportion of the log likelihood that is accounted for by the model. Just remember than an R_L^2 value of 1.0 indicates that the logistic regression model achieves the same log likelihood as the saturated model. However, this does not mean that it fits the data perfectly. Instead, it means that it fits the data as well as could be hoped for.

Reduction From Model R-Squared

This is amount that *R*-squared is reduced when the term is omitted from the regression model. This reduction is calculated from the *R*-squared achieved by the full model.

This quantity is used to determine if removing a term causes a large reduction in R-squared. If it does not, then the term can be safely removed from the model.

Reduction From Saturated R-Squared

This is the amount that *R*-squared is reduced when the term is omitted from the regression model. This reduction is calculated from the *R*-squared achieved by the saturated model. This item is included because it shows how removal of this term impacts the best *R*-squared that is possible.

Classification Table

```
                 Estimated
Actual                   0            1            Total
0                       15            3               18
1                        2            7                9
Total                   17           10               27
Percent Correctly classified = 81.5%
```

This table displays the results of classifying the data based on the logistic regression equations. The table presents the counts for each category.

The Percent Correctly Classified is also presented. This is the percent of the total count that fall on the diagonal of the table.

ROC Section

```
ROC Section for Value 0
Prob      N(1|1)   N(1|0)   N(0|1)  N(0|0)  Sensitivity  Specificity  Sensitivity   Proportion
Cutoff    A        B        C       D       A/(A+C)      D/(B+D)      +Specificity  Correct
0.05000   18       8        0       1       1.00000      0.11111      1.11111       0.70370
0.10000   17       8        1       1       0.94444      0.11111      1.05556       0.66667
0.15000   17       8        1       1       0.94444      0.11111      1.05556       0.66667
0.20000   17       5        1       4       0.94444      0.44444      1.38889       0.77778
0.25000   16       4        2       5       0.88889      0.55556      1.44444       0.77778
0.30000   15       3        3       6       0.83333      0.66667      1.50000       0.77778
0.35000   15       3        3       6       0.83333      0.66667      1.50000       0.77778
0.40000   15       2        3       7       0.83333      0.77778      1.61111       0.81481
0.45000   15       2        3       7       0.83333      0.77778      1.61111       0.81481
0.50000   15       2        3       7       0.83333      0.77778      1.61111       0.81481
0.55000   15       1        3       8       0.83333      0.88889      1.72222       0.85185
0.60000   12       0        6       9       0.66667      1.00000      1.66667       0.77778
0.65000   11       0        7       9       0.61111      1.00000      1.61111       0.74074
0.70000   11       0        7       9       0.61111      1.00000      1.61111       0.74074
0.75000   9        0        9       9       0.50000      1.00000      1.50000       0.66667
0.80000   9        0        9       9       0.50000      1.00000      1.50000       0.66667
0.85000   8        0        10      9       0.44444      1.00000      1.44444       0.62963
0.90000   7        0        11      9       0.38889      1.00000      1.38889       0.59259
0.95000   7        0        11      9       0.38889      1.00000      1.38889       0.59259
Area Under ROC Curve = 0.89198
```

One ROC report is generated for each category. Only the report for category 0 is displayed here. ROC curves can be used to determine appropriate cutoff values for classification by letting you compare the sensitivity and specificity of various cutoff values. When classifying, you usually classify a row into that category that has the highest membership probability. However, this is not always the optimum strategy. This table shows you what happens when various cutoff values are selected.

Classifying an observation can have any one of four possible results. An observation from the group can be correctly classified as being from that group (state A) or incorrectly classified as being from another group (state C). An observation from another group can be incorrectly classified as being from the group (state B) or correctly classified as being from another group (state D).

The number of observations in each state is computed for each cutoff value between zero and one. A number of measures can be calculated from these values. The measures used in ROC analysis are called *sensitivity* and *specificity*. Sensitivity is the proportion of those from this group that are correctly identified as such. In terms of the four states, sensitivity = *A/(A+C).* Specificity is the proportion of those from other groups that are correctly identified as such. In terms of four states, specificity = *D/(B+D).* Thus, the optimum cutoff value is that one for which the sum of sensitivity and specificity is the maximum. This may be found be investigating the report. An ROC plot is also generated for each report that gives a graphical display of this report.

An ROC analysis is most useful in the two-group case. In the multiple-group case, it is of only marginal usefulness, since a cutoff value is not specified. Rather, each observation is classified into that group which has the highest membership probability.

Prob Cutoff

This is the probability cutoff for classification into this group. If an observation’s predicted probability for membership in this group is greater than this amount, the observation is classified in this group. Otherwise, it is classified as being in some other group.

A B C D

The counts for each of the four states. These counts are represented using the notation *N(i|j)* where *i* is the classified group and *j* is the actual group.

Sensitivity

Sensitivity is the proportion of those from this group that are correctly identified as such. In terms of the four states, sensitivity = *A/(A+C).*

Specificity

Specificity is the proportion of those from other groups that are correctly identified as such. In terms of four states, specificity = *D/(B+D).*

Sensitivity + Specificity

A common rule for selecting an appropriate cutoff value is to choose the cutoff with the largest total of sensitivity and specificity. This column allows you to do this very quickly.

Proportion Correct

Another rule for selecting an appropriate cutoff value is to choose that cutoff which maximizes the number of observations that are correctly classified. This column of the report allows you to quickly find the optimum cutoff value. Unfortunately, when one group has many more rows than the others, this rule may not be useful since it will lead you to classify everyone into the most prevalent group.

Area Under ROC Curve

The area under the ROC curve is a popular measure associated with ROC curves. When applied to classification in logistic regression, its maximum value of one occurs when all rows are correctly classified. Its minimum value of zero occurs when all rows are incorrectly classified. Thus, the nearer this value is to one, the better the classification.

Row Classification Section

```
                                  Estimated     Lower 95%    Upper 95%
       Actual      Estimated     REMISS        Confidence   Confidence
Row    REMISS      REMISS        Probability   Limit        Limit
1      1           1             0.83900       0.31617      0.98326
2      1           1             0.73317       0.48928      0.88739
3      0           0             0.81061       0.24565      0.98253
4      0           0             0.55936       0.24511      0.83230
5      1           1             0.83326       0.44347      0.96908
6      0           0             0.57370       0.21384      0.86943
7*     1           0             0.51337       0.32143      0.70145
8*     0           1             0.75562       0.21175      0.97267
9      0           0             0.71480       0.31903      0.93059
10     0           0             0.99687       0.19043      1.00000
.      .           .             .             .            .
.      .           .             .             .            .
.      .           .             .             .            .
```

This report displays the actual and predicted group and membership probability for each row of the report. It also provides confidence limits for the predicted group-membership probability.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

Estimated Group

This is the group with the largest membership probability.

Estimated Probability

This is the estimated probability that the row belongs to the group listed in the Estimated Group column.

These values allow you to determine how certain the classification is. When the value is near one (above 0.7), the logistic regression is convinced that the observation belongs in the designated group. When the value is near 0.5 or less, the classification was not as clear.

Lower and Upper Confidence Limits

These values provide a confidence interval for the estimated membership probability. Note that this confidence interval is only approximate in the multiple-group case. Formulas and technical details are given above in the section entitled Predicted Probabilities.

Row Classification Probabilities

```
                     Estimated     Estimated
         Actual      Prob. in      Prob. in
Row      REMISS      0             1
1        1           0.16100       0.83900
2        1           0.26683       0.73317
3        0           0.81061       0.18939
4        0           0.55936       0.44064
5        1           0.16674       0.83326
6        0           0.57370       0.42630
7*       1           0.51337       0.48663
8*       0           0.24438       0.75562
9        0           0.71480       0.28520
10       0           0.99687       0.00313
.        .           .             .
.        .           .             .
.        .           .             .
```

This report displays the actual group and the membership probabilities for each group and each row. This allows you investigate how certain each classification is.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

Estimated Prob. In Group

This is the estimated probability that the row belongs in each group. These values allow you to determine how certain the classification is.

Simple Residual Report

```
                     Residual      Residual
         Actual     for Group     for Group
Row      REMISS             0             1
1        1           -0.16100       0.16100
2        1           -0.26683       0.26683
3        0            0.18939      -0.18939
4        0            0.44064      -0.44064
5        1           -0.16674       0.16674
6        0            0.42630      -0.42630
7*       1           -0.51337       0.51337
8*       0            0.75562      -0.75562
9        0            0.28520      -0.28520
10       0            0.00313      -0.00313
.        .                  .             .
.        .                  .             .
.        .                  .             .
```

This report displays the simple residuals for each group. Each of the *g* logistic regression equations can be used to estimate the probabilities that each observation belongs to the corresponding group.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

Residual for Group

These residuals are defined as

$$r_{gj} = y_{gj} - p_{gj}$$

where p_{gj} is the estimated membership probability and y_{gj} is an indicator variable that is one if the actual group is *g* and zero otherwise.

Note that, unlike multiple regression, there are *g* residuals for each observation instead of just one. This makes residual analysis much more difficult. If the logistic regression model fits an observation closely, all of its residuals will be small, but never zero.

Unfortunately, the simple residuals have unequal variance equal to $n_j \pi_{gj}\left(1-\pi_{gj}\right)$, where n_j is the number of observations with the same values of the independent variables as observation *j*. This unequal variance makes comparisons among the simple residuals difficult and alternative types of residuals are necessary.

Residual Report

```
      Actual         Pearson                  Deviance                    Maximum
Row   REMISS         Residual                 Residual                    Hat Diagonal
1     1             -0.43806  |.............  -0.59253 |||...........     0.20631  ||||..........
2     1             -0.60328  ||............  -0.78789 |||||.........     0.05654  |.............
3     0              0.48336  |.............   0.64802 ||||..........     0.26518  ||||||........
4     0              1.25520  |||||.........   1.52442 |||||||||.....     0.23855  |||||.........
5     1             -0.44733  |.............  -0.60400 |||...........     0.12192  ||............
6     0              0.86201  |||...........   1.05417 ||||||........     0.16277  |||...........
7*    1             -1.02710  ||||..........  -1.20021 |||||||.......     0.04169  |.............
8*    0              1.75843  |||||||.......   1.67872 ||||||||||....     0.28695  ||||||........
9     0              0.63166  ||............   0.81945 |||||.........     0.14925  |||...........
10    0              0.05607  |.............   0.07923 |.............     0.04227  |.............
.     .                     . .                      . .                  .        .
.     .                     . .                      . .                  .        .
.     .                     . .                      . .                  .        .
```

This report displays the Pearson residuals, the deviance residuals, and the hat diagonal for each row. These are the residuals that most textbooks on logistic regression recommend that you use.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

Pearson Residual

The *Pearson residuals* give the contribution of each row to the Pearson chi-square goodness of fit statistic. When the values of the independent variables of each observation are unique, the formula for this residual is

$$\chi_j = \pm\sqrt{\sum_{g=1}^{G}\frac{\left(w_{gj} - n_j p_{gj}\right)^2}{n_j p_{gj}}}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if w_{gj} / n_j is greater (less) than p_{gj}. By definition, the sum of the squared Pearson residuals is the Pearson chi-square goodness of fit statistics.

Deviance Residuals

Remember that the deviance is -2 times the difference between log likelihoods of a reduced model and the saturated model. The formula for a deviance residual is

$$d_j = \pm\sqrt{2\sum_{g=1}^{G} w_{gj}\ln\left(\frac{w_{gj}}{n_j p_{gj}}\right)}, \quad j = 1,2,\cdots,J$$

where the plus (minus) is used if $w_{REF(g),j} / n_j$ is greater (less) than $p_{REF(g),j}$. By definition, the sum of the squared deviance residuals is the deviance.

Maximum Hat Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. These are often called *leverage* design points. The larger the value of the hat diagonal, the more the observation influences estimates of the regression coefficients. There is a separate hat diagonal defined for each category. The value reported here is the maximum of all *G* of the hat diagonals for each row.

An observation that has a large residual, but has low leverage, does not cause much concern. However, an observation with a large leverage and a large residual should be checked very carefully. The formula for the hat diagonal associated with the *j*th observation and *g*th group is

$$h_{gj} = n_j p_{gj}\left(1 - p_{gj}\right)\sum_{i=1}^{p}\sum_{k=1}^{p} X_{ij} X_{kj} \hat{V}_{gik}, \quad j = 1,2,\cdots,J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix of the regression coefficients associated with the *g*th regression equation. The interpretation of this diagnostic is not as clear in logistic regression as in multiple regression because it involves the predicted values which in turn involve the dependent variable. In multiple regression, the hat diagonals only involve the independent variables.

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

DFBetas Report

```
DFBetas Report For REMISS = 0
      Actual       DFBeta                     DFBeta                     DFBeta
Row   REMISS       Intercept                  CELL                       LI
1     1            -0.05383  |.............   0.11561  |.............   -0.12403  |.............
2     1            -0.06191  |.............   0.03603  |.............   -0.07986  |.............
3     0            -0.03248  |.............  -0.29680  |||...........   -0.19367  ||............
4     0             0.07853  |.............   0.22408  ||............   -0.36761  |||||.........
5     1            -0.15954  ||............  -0.02455  |.............   -0.11640  |.............
6     0             0.10146  |.............   0.11173  |.............   -0.16597  ||............
7*    1             0.05201  |.............   0.05264  |.............   0.12518   |.............
8*    0            -0.83713  ||||||||||||||  -0.10576  |.............   0.19110   ||............
9     0            -0.20605  |||...........  -0.03081  |.............   -0.23153  |||...........
10    0            -0.01139  |.............  -0.00613  |.............   -0.01005  |.............
9     0             0.63166  ||............   0.81945  |||||.........   0.14925   |||...........
10    0             0.05607  |.............   0.07923  |.............   0.04227   |.............
.     .                    . .                       . .              .         .
.     .                    . .                       . .              .         .
.     .                    . .                       . .              .         .
```

One way to study the impact of an observation on each regression coefficient is to determine how much that coefficient changes when the observation is deleted. The DFBETA statistic is the standardized difference between a regression coefficient before and after the removal of the *j*th observation.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

DFBeta

The DFBeta statistic is the standardized difference between a regression coefficient before and after the removal of the *j*th observation.

The formula for DFBeta is approximated by

$$\mathrm{DFBeta}_{gij} = \left(\frac{w_{gj} - n_j p_{gj}}{\left(1 - h_{gj}\right)\sqrt{\hat{V}_{gii}}} \right) \sum_{k=1}^{p} X_{kj} \hat{V}_{gik}, \quad j = 1, 2, \cdots, J$$

where $\hat{V}_{gik}$ is the portion of the covariance matrix associated with the *g*th regression equation.

Note that this formula matches Pregibon (1981) in the two group case, but is different from Lesaffre (1989) in the multi-group case.

Influence Diagnostics Report

```
Influence Diagnostics Report For REMISS = 0
                                       Cook's                  Cook's
         Actual          Hat         Distance                  Distance
Row      REMISS      Diagonal             (C)                  (CBar)
1        1            0.20631  ||||..........   0.06285  |.............   0.04988  |.............
2        1            0.05654  |.............   0.02312  |.............   0.02181  |.............
3        0            0.26518  ||||||........   0.11474  |.............   0.08432  |.............
4        0            0.23855  |||||.........   0.64822  |||||.........   0.49359  |||||.........
5        1            0.12192  ||............   0.03164  |.............   0.02778  |.............
6        0            0.16277  |||...........   0.17254  |.............   0.14446  |.............
7*       1            0.04169  |.............   0.04790  |.............   0.04590  |.............
8*       0            0.28695  ||||||........   1.74508  |||||||||||||.   1.24433  ||||||||||||||
9        0            0.14925  |||...........   0.08228  |.............   0.07000  |.............
10       0            0.04227  |.............   0.00014  |.............   0.00014  |.............
.        .                  .  .                      .  .                .        .
.        .                  .  .                      .  .                .        .
.        .                  .  .                      .  .                .        .
```

This report gives two distance measures similar to Cook's distance in multiple regression.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

Hat Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. They are discussed in more detail in the Residual Report.

Cook's Distance (C) and (CBar)

C and *Cbar* are extensions of Cooks distance for logistic regression. Quoting from Pregibon (1981), page 719:

"Cbar measures the overall change in fitted logits due to deleting the *l*th observation for all points excluding the one deleted. Conversely, *C* includes the deleted point. Although *C* will usually be the preferred diagnostic to measure overall coefficients' changes, in the examples examined to date, the one-step approximations were more accurate for *Cbar* than *C*."

The formulas for *C* and *Cbar* are

$$C_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1 - h_{gj}\right)^2}, \quad j = 1,2,\cdots,J$$

$$\overline{C}_{gj} = \frac{\chi_j^2 h_{gj}}{\left(1 - h_{gj}\right)}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

Residual Diagnostics Report

```
Residual Diagnostics Report For REMISS = 0
                                        Deviance                 Chi-Square
       Actual              Hat          Change                   Change
Row    REMISS         Diagonal          (DFDev)                  (DFChi2)
1      1               0.20631  ||||..........   0.40098  |.............   0.24178   |.............
2      1               0.05654  |.............   0.64257  |.............   0.38576   |.............
3      0               0.26518  ||||||........   0.50425  |.............   0.31795   |.............
4      0               0.23855  |||||.........   2.81743  ||||||........   2.06910   ||............
5      1               0.12192  ||............   0.39260  |.............   0.22789   |.............
6      0               0.16277  |||...........   1.25574  ||............   0.88752   |.............
7*     1               0.04169  |.............   1.48639  |||...........   1.10084   |.............
8*     0               0.28695  ||||||........   4.06243  |||||||||.....   4.33640   ||||..........
9      0               0.14925  |||...........   0.74150  |.............   0.46899   |.............
10     0               0.04227  |.............   0.00642  |.............   0.00328   |.............
.      .                     .  .                      .  .                .         .
.      .                     .  .                      .  .                .         .
.      .                     .  .                      .  .                .         .
```

This report gives statistics that help detect observations that have not been fitted well by the model.

Row

This is the row from the database. Rows that are starred are misclassified.

Actual Group

This is the group to which this row belongs (if known).

Hat Diagonal

The diagonal elements of the hat matrix can be used to detect points that are extreme in the independent variable space. They are discussed in more detail in the Residual Report.

Deviance Change (DFDev) and Chi-Square Change (DFChi2)

DFDEV and *DFCHI2* are statistics that measure the change in deviance and in Pearson's chi-square, respectively, that occurs when an observation is deleted from the dataset. Large values of these statistics indicate observations that have not been fitted well.

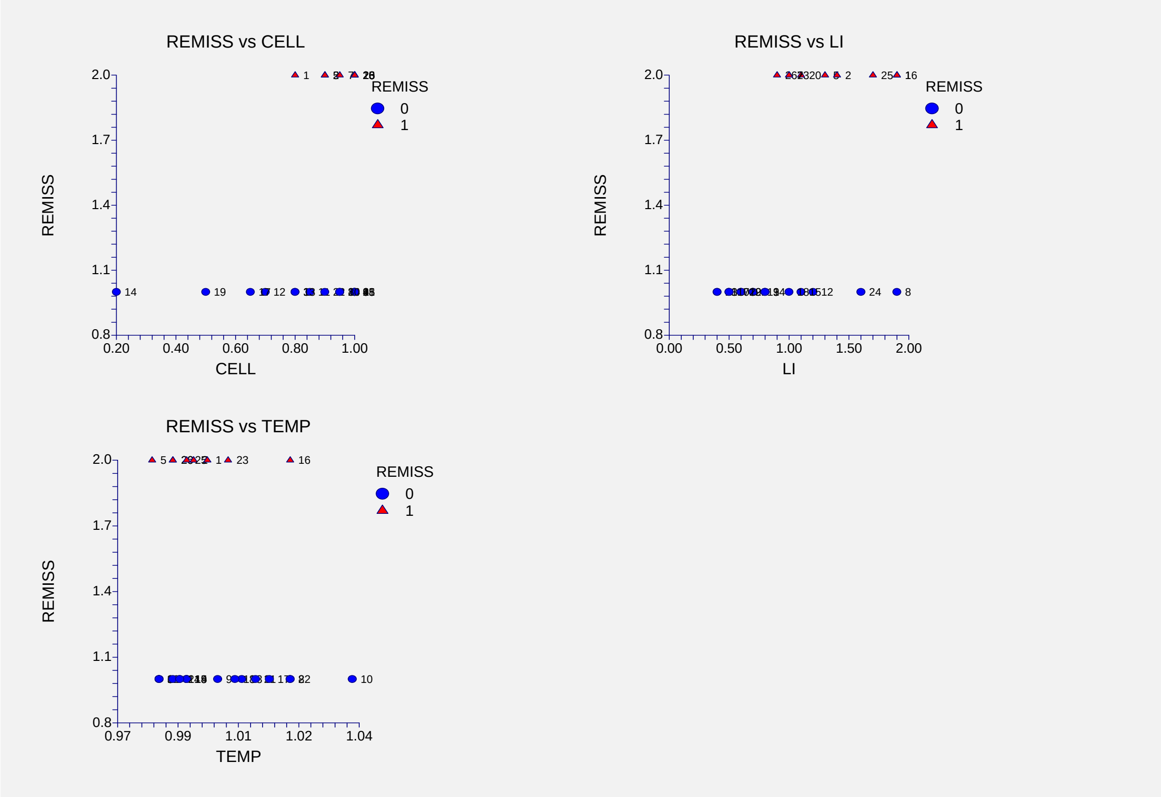
The formulas for these statistics are

$$DFDEV_{gj} = d_j^2 + \overline{C}_{gj}, \quad j = 1,2,\cdots,J$$

$$DFCHI2_{gj} = \frac{\overline{C}_{gj}}{h_{gj}}, \quad j = 1,2,\cdots,J$$

Note that this formula matches Pregibon (1981) in the two-group case. In the multiple-group case, the two-group formula is applied to each group.

Y versus X Plots



This section shows scatter plots with the dependent variable on the vertical axis and each of the independent variables on the horizontal axis. The plot is useful for finding typos, outliers, and other anomalies in that data.

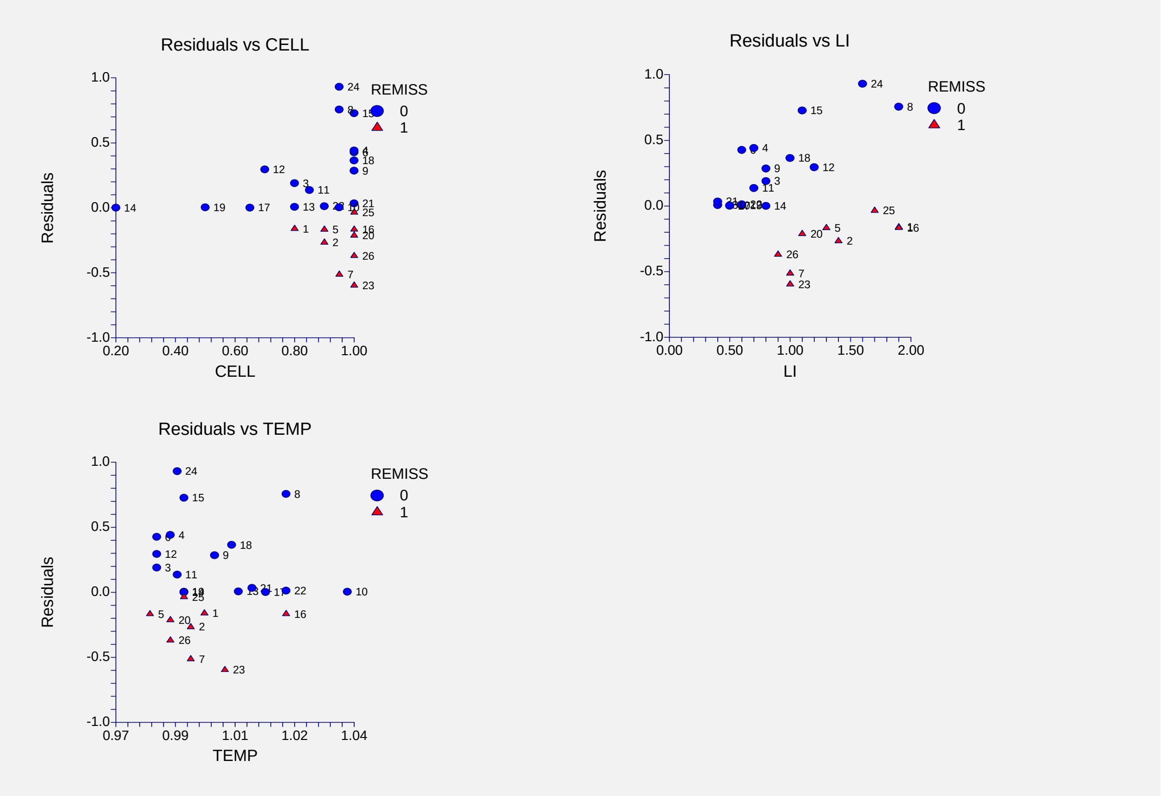
Vertical Axis

The categories of the dependent variable are shown on the vertical axis. Each category is assigned a whole number, beginning with the number one. The numbers are assigned in sorted order. Thus, if your dependent variable has values A, B, and C, it would be plotted on a numeric scale ranging from about 0.8 to 3.2. The groups would be plotted as the numbers 1, 2, and 3.

Horizontal Axis

The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

Simple Residuals versus X Plots



This section shows scatter plots with the simple residuals on the vertical axis and each of the independent variables on the horizontal axis. The plots are useful for finding outliers and other anomalies in the data.

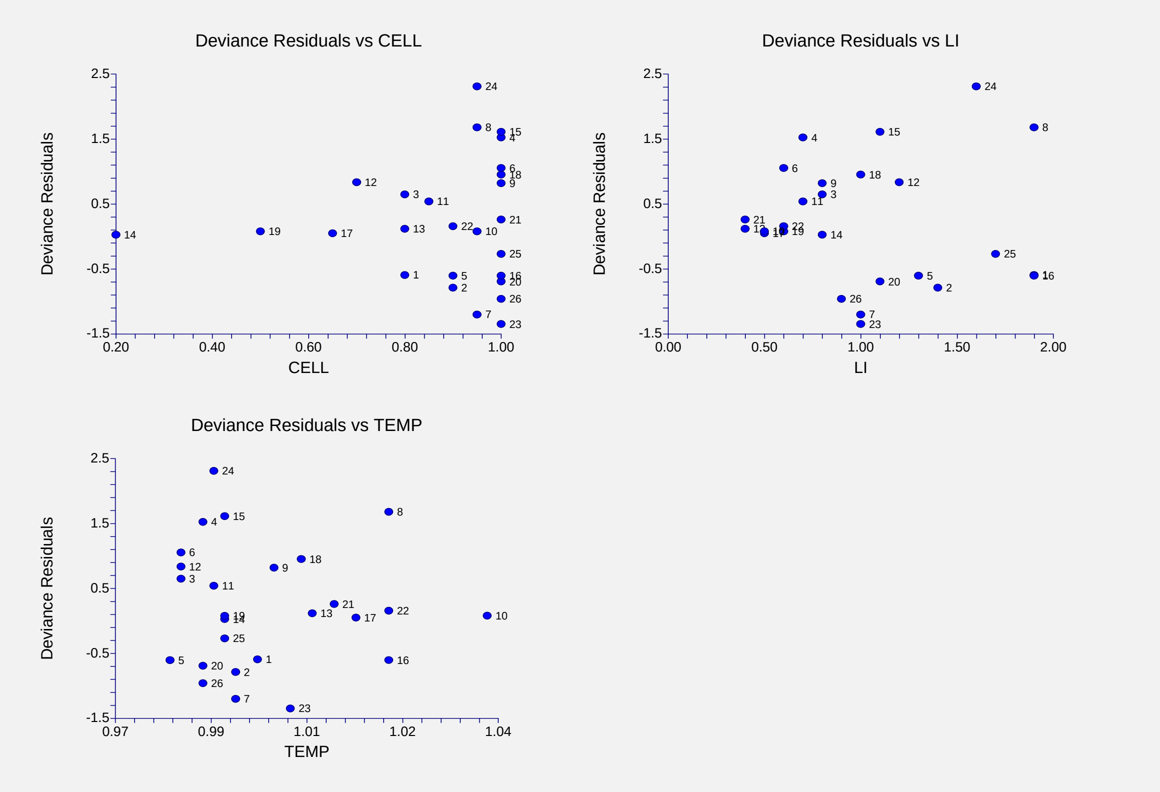
Vertical Axis

The residuals are displayed on the vertical axis. Note that the *G* residuals for each row corresponding to the simple residuals are displayed. Thus, if you have *N* rows, you will have *GN* points displayed on the plot.

Horizontal Axis

The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

Deviance Residuals versus X Plots



This section shows scatter plots with the deviance residuals on the vertical axis and each of the independent variables on the horizontal axis. The plots are useful for finding outliers and other anomalies in the data.

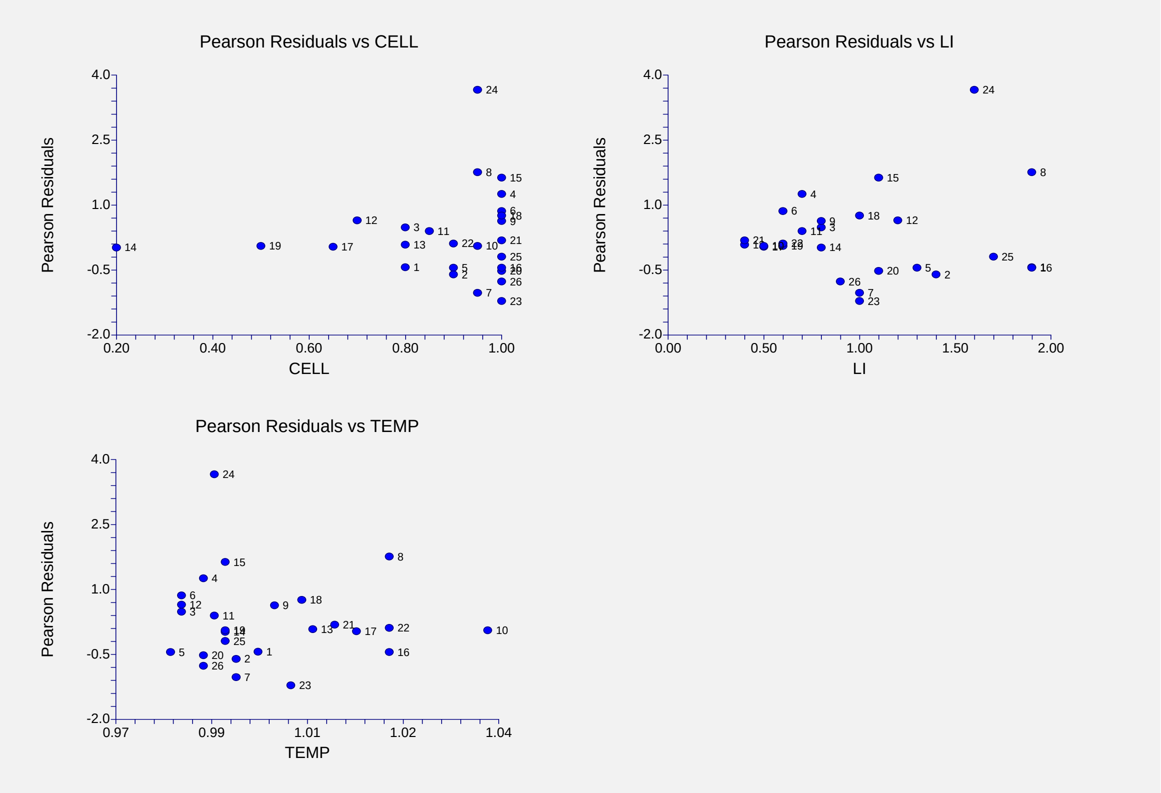
Vertical Axis

The deviance residuals are displayed on the vertical axis.

Horizontal Axis

The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

Pearson Residuals versus X Plots



This section shows scatter plots with the Pearson residuals on the vertical axis and each of the independent variables on the horizontal axis. The plots are useful for finding outliers and other anomalies in the data.

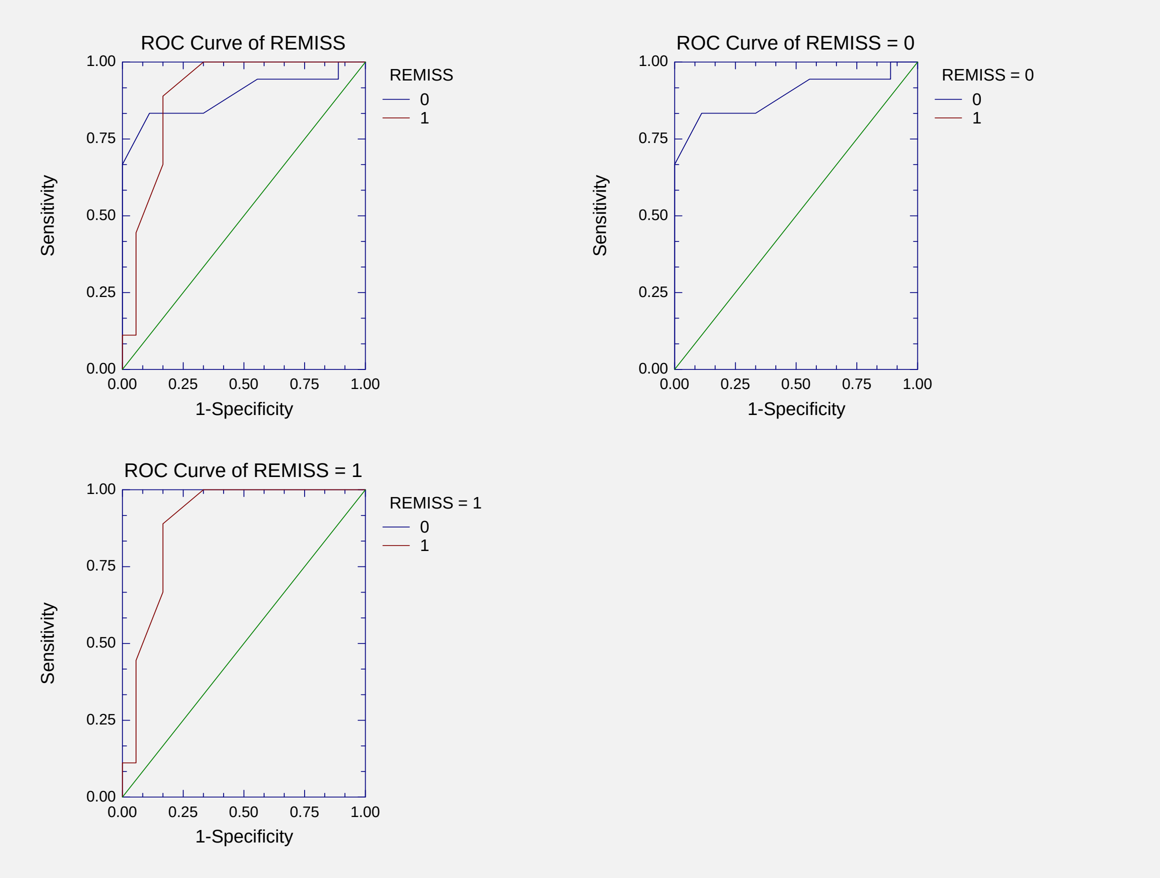
Vertical Axis

The Pearson residuals are displayed on the vertical axis.

Horizontal Axis

The independent variables are shown on the horizontal axis. When the independent variable is categorical, binary variables are generated for each of the categories and a separate scatter plot is generated for each binary variable.

ROC Curves - Combined and Separate



This section displays the ROC curves that can be used to help you find the best cutoff points to use for classification. The cutoff point nearest the top-left corner of the plot is the optimum cutoff. You will have to refer to the ROC Report to determine the exact value of the cutoff.

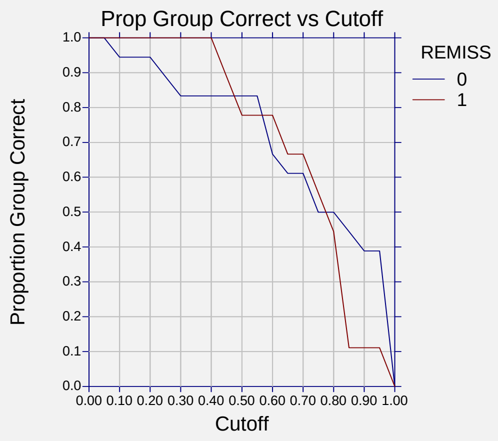
Vertical Axis

The sensitivity is displayed on the vertical axis.

Horizontal Axis

One minus the specificity is displayed on the horizontal axis.

Prob Correct versus Cutoff Plot



This section displays a plot that shows the proportion correct versus the cutoff. It is useful to help determine the cutoff point used in classification. This plot may be difficult to use with three or more categories because of the ambiguity in the plot.

Vertical Axis

The proportion correctly classified for various cutoff values are displayed on the vertical axis.

Horizontal Axis

The cutoff values are displayed on the horizontal axis. These cutoff values are in terms of the estimated group-membership probabilities. Thus a cutoff of 0.4 means that any rows with a group-membership probability of 0.4 or more are classified into this group.

Example 2 – Subset Selection

This section presents an example of how to conduct a subset selection. The data used are stored in the LEUKEMIA database. This analysis will search for the best model from among a pool of the six numeric variables.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Logistic Regression window.

1 Open the LEUKEMIA dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LEUKEMIA.S0**.
- Click **Open**.

2 Open the Logistic Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Logistic Regression window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **REMISS** from the list of variables and then click **Ok**. “REMISS” will appear in the **Y: Group Variable** box.
- Click on the **Default Reference Group** box and select **Last Group after Sorting**.
- Double-click in the **X’s: Numeric Independent Variables** box. This will bring up the variable selection window.
- Select the variables from **CELL** to **TEMP** from the list of variables and then click **Ok**. “CELL-TEMP” will appear in the X’s: Numeric Independent Variables box.

4 Specify the model.

- On the Logistic Regression window, select the **Model tab**.
- Select **Hierarchical Forward with Switching** in the Subset Selection box.
- Set the **Max Terms in Subset** to **6**.
- Set the **Which Model Terms** box to **Up to 1-Way**.

5 Specify the reports.

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **None**.
- Check the **Run Summary**, **Subset Selection - Summary**, **Subset Selection - Detail**, and **Parameter Significance Tests** reports. All other reports should be unchecked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**.

Run Summary Section

```
Parameter                  Value                Parameter              Value
Dependent Variable         REMISS               Rows Processed         29
Reference Value            1                    Rows Used              27
Number of Values           2                    Rows for Validation    0
Frequency Variable         None                 Rows X's Missing       2
Numeric Ind. Variables     6                    Rows Freq Miss. or 0   0
Categorical Ind. Variables 0                    Rows Prediction Only   0
Final Log Likelihood       -10.87752            Unique Row Patterns    27
Model R-Squared            0.36707              Sum of Frequencies     27
Actual Convergence         2.081623E-06         Likelihood Iterations  9
Target Convergence         0.000001             Maximum Iterations     20
Model D.F.                 6                    Max Like Message       Quasi-Separation
Model                      CELL|SMEAR|INFIL|LI|BLAST|TEMP

******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********
Your dataset had QUASI-COMPLETE SEPARATION which means that the maximum likelihood routine
did NOT converge so the statistical tests are not valid. Although the prediction equations
correctly classified much of your data, they may not do so for other observations.
Quasi-Complete Separation often occurs because your sample size is too small.
******** WARNING ******** WARNING ******** WARNING ******** WARNING ******** WARNING ********
```

The first thing we notice is the warning message about quasi-separation. If quasi-separation occurs, the maximum likelihood estimates do not exist and all results are suspect. We note that 9 likelihood iterations occurred and the Actual Convergence is near the Target Convergence. We decide to rerun the analysis after resetting the Max Terms in Subset box from 6 to 5. Note that this error message often occurs when a small set of data is fit with a model with too many terms.

At this point, either reset the value of Max Terms in Subset (on the Model tab) to 5 manually or load the template **Example2a**. Now, rerun the analysis.

Run Summary Section

```
Parameter                Value                 Parameter             Value
Dependent Variable       REMISS                Rows Processed        29
Reference Value          1                     Rows Used             27
Number of Values         2                     Rows for Validation   0
Frequency Variable       None                  Rows X's Missing      2
Numeric Ind. Variables   6                     Rows Freq Miss. or 0  0
Categorical Ind. Variables 0                   Rows Prediction Only  0
Final Log Likelihood     -10.92900             Unique Row Patterns   27
Model R-Squared          0.36407               Sum of Frequencies    27
Actual Convergence       7.136538E-07          Likelihood Iterations 7
Target Convergence       0.000001              Maximum Iterations    20
Model D.F.               5                     Max Like Message      Normal Completion
Model                    CELL|SMEAR|INFIL|LI|BLAST|TEMP
```

The warning message has disappeared and the algorithm finished normally.

Subset Selection Summary Section

```
No.      No.      Log           R-Squared     R-Squared
Terms    X's      Likelihood    Value         Change
1        1        -17.18588     0.00000       0.00000
2        2        -13.03648     0.24144       0.24144
3        3        -12.17036     0.29184       0.05040
4        4        -10.97669     0.36130       0.06946
5        5        -10.92900     0.36407       0.00277
```

This report shows the best log-likelihood value for each subset size. In this example, it appears that four terms (the intercept and three variables) provides the best model. Note that adding the fifth variable does not increase the R-squared value very much.

No. Terms

The number of terms. Note that this includes the intercept.

No. X’s

The number of *X*’s that were included in the model. Note that in this case, the number of terms matches the number of *X*’s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function evaluated at the maximum likelihood estimates. Our goal is to find a subset size above which little is gained by adding more variables.

R-Squared Value

This is the value of *R*-squared calculated using the formula

$$R_L^2 = \frac{L_p - L_0}{L_0 - L_S}$$

as discussed in the introduction. We are looking for the subset size at which this value does not increase by a meaningful amount.

R-Squared Change

This is the increase in *R*-squared that occurs when each new subset size is reached. Search for the subset size below which the *R*-squared value does not increase by more than 0.02 for small samples or 0.01 for large samples.

In this example, the optimum subset size appears to be four terms.

Subset Selection Detail Section

```
                No. of    No. of    Log          Term         Terms
Step   Action   Terms     X's       Likelihood   Entered      Removed
1      Add      1         1         -17.18588    Intercept
2      Add      2         2         -13.03648    LI
3      Add      3         3         -12.17036    CELL
4      Add      4         4         -10.97669    TEMP
5      Add      5         5         -10.92900    SMEAR
```

This report shows the highest log likelihood for each subset size. In this example, it appears that four terms (the intercept and three variables) provide the best model. Note that adding the fifth variable does not increase the *R*-squared value very much.

Action

This item identifies the action that was taken at this step. A term was added, removed, or two were switched.

No. Terms

The number of terms. Note that this includes the intercept.

No. X’s

The number of *X*’s that were included in the model. Note that in this case, the number of terms matches the number of *X*’s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function after the completion of this step. Our goal is to find a subset size above which little is gained by adding more variables.

Terms Entered and Removed

These columns identify the terms added, removed, or switched.

Discussion of Example 2

After considering these reports, it was decided to include CELL, LI, and TEMP in the final logistic regression model. Another run should now take place using only these independent variables. A complete residual analysis would be necessary before the equation is finally adopted.

Example 3 – One Categorical Variable

The independent variables in logistic regression may be categorical as well as numerical. This example is of the simplest categorical case of a binary response and a binary independent variable. More complicated examples will be shown below.

In this example, a simple yes-no question is asked of each member of two groups. The following two-by-two table presents the results. The analyst wants to understand the relationship between group membership and response to the question.

	Response		
Group	**Yes**	**No**	**Total**
A	91	9	100
B	93	27	120
Total	184	36	220

These data would normally be analyzed using the methods for comparing two proportions such as Fisher's exact test or the chi-square test for independence in a contingency table. The following table presents the results of this analysis.

Two Proportions Output

```
Data Section
               Sample   Number in   Number in   Proportion In   Proportion In
Sample           Size   Group One   Group Two       Group One       Group Two
One               100           9          91        0.090000        0.910000
Two               120          27          93        0.225000        0.775000
Total             220          36         184        0.163636        0.836364

Hypothesis Test Section
                 Fisher's Exact Test      Normal Approximation                  Yates Chi-Square Test
Alternative          Prob   Decision                  Prob   Decision        Chi-Square        Prob
Hypothesis          Level       (5%)     Z-Value     Level       (5%)             Value       Level
P1-P2<>0         0.009733  Reject Ho     -2.6951  0.007037  Reject Ho            6.3107    0.012001
P1-P2<0          0.005272  Reject Ho     -2.6951  0.003518  Reject Ho
P1-P2>0          0.998394  Accept Ho     -2.6951  0.996482  Accept Ho

Odds Ratio and Relative Risk Section
                          Common    Original    Iterated    Log Odds    Relative
Parameter             Odds Ratio  Odds Ratio  Odds Ratio       Ratio        Risk
Upper  95% C.L.                     0.779298    0.809907   -0.249362    0.838047
Estimate                0.340659    0.353005    0.353005   -1.041272    0.400000
Lower  95% C.L.                     0.159904    0.139852   -1.833182    0.180300
```

The conclusion of this analysis is to reject the null hypothesis that the two proportions are equal. The significance levels are 0.009733 using Fisher's exact test and 0.007037 using the normal approximation which is equivalent to the chi-square test for independence. Note that the odds ratio is 0.340659.

We will now see how to analyze these data using logistic regression. The data must be entered into a database so that they can be processed. The following table shows how these data are rearranged and entered. These data have been entered into a database named 2BY2.

2BY2 dataset (subset)

Group	Response	Count
A	No	9
A	Yes	91
B	No	27
B	Yes	93

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Logistic Regression window.

1 Open the 2BY2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **2BY2.S0**.
- Click **Open**.

2 Open the Logistic Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Logistic Regression window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **Response** from the list of variables and then click **Ok**. “Response” will appear in the Y: Group Variable box.
- Click on the **Default Reference Group** box and select **Last Group after Sorting**.
- Double-click in the **X’s: Categorical Independent Variables** box. This will bring up the variable selection window.
- Select the variable **Group** from the list of variables and then click **Ok**. “Group” will appear in the X’s: Categorical Independent Variables box.
- Click on the **Default Reference Value** box and select **Last Value after Sorting**.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select the variable **Count** from the list of variables and then click **Ok**. “Count” will appear in the Frequency Variable box.

4 Specify the reports.

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **None**.
- Check the **Run Summary**, **Response Analysis**, **Parameter Significance Tests**, **Odds Ratios**, **Analysis of Deviance**, and **Log-Likelihood and R-Squared** reports. All other reports should be unchecked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**.

Selected portions of the output reports are shown below.

Logistic Regression Output

```
Run Summary Section
Parameter                 Value              Parameter              Value
Dependent Variable        Response           Rows Processed         4
Reference Value           YES                Rows Used              4
Number of Values          2                  Rows for Validation    0
Frequency Variable        Count              Rows X's Missing       0
Numeric Ind. Variables    0                  Rows Freq Miss. or 0   0
Categorical Ind. Variables 1                 Rows Prediction Only   0
Final Log Likelihood      -94.23344          Unique Row Patterns    4
Model R-Squared           0.06908            Sum of Frequencies     220
Actual Convergence        2.559022E-11       Likelihood Iterations  6
Target Convergence        0.000001           Maximum Iterations     20
Model D.F.                2                  Max Like Message       Normal Completion
Model                     Group

Response Analysis Section
Response                         Unique                Act vs Pred   % Correctly
Categories        Count            Rows       Prior      R-Squared    Classified
NO                   36               2     0.50000        0.03302        75.000
YES                 184               2     0.50000        0.03302        49.457
Total               220               4                                   53.636

Parameter Significance Tests Section (Reference Value: Response = YES)
                 Regression                          Wald         Wald          Odds
                Coefficient      Standard         Z-Value         Prob         Ratio
Parameter       (B or Beta)         Error        (Beta=0)        Level        Exp(B)
B0: Intercept       0.39465       0.12074           3.269      0.00108       1.48387
B1: (Group=”A”)    -1.07687       0.41218          -2.613      0.00898       0.34066

Odds Ratios Section (Reference Value: Response = YES)
                 Regression          Odds     Lower 95%         Upper 95%
                Coefficient         Ratio    Confidence        Confidence
Parameter       (B or Beta)        Exp(B)         Limit             Limit
B0: Intercept       0.39465       1.48387       1.17119           1.88004
B1: (Group=”A”)    -1.07687       0.34066       0.15187           0.76413

Analysis of Deviance Section
                                                 Increase
                                               From Model
Term                                             Deviance          Prob
Omitted            DF        Deviance        (Chi Square)         Level
All                 1       196.08640             7.61951       0.00577
GROUP               1       196.08640             7.61951       0.00577
None(Model)         1       188.46689
```

```
Log Likelihood & R-Squared Section
                                           R-Squared      Reduction        Reduction
Term(s)                         Log     Of Remaining     From Model   From Saturated
Omitted              DF   Likelihood        Term(s)       R-Squared        R-Squared
All                   1    -98.04320        0.00000
GROUP                 1    -98.04320        0.00000         0.06908          1.00000
None(Model)           1    -94.23344        0.06908         0.00000          0.93092
None(Saturated)       4    -42.89226        1.00000                          0.00000
```

Although a casual comparison between this report and that of the Two Proportion procedure shows little in common, a more detailed report shows many similarities. First of all, notice that the significance level of the test of GROUP in the Analysis of Deviance Section of 0.00577 compares very closely with the 0.007037 from the chi-square test. Also notice that the odds ratios from both reports round to 0.34066. The confidence limits of these two reports are not exactly the same, but they are close.

To summarize the logistic regression analysis, we can conclude that there is a significant relationship between response and group.

This example has shown the similarities between these two approaches to the analysis of two proportions. Usually, you would analyze these data using the two proportions approach. However, that approach is not as easily extended to the case of several independent variables including a mixture of categorical and numeric.

Example 4 – Logit Model Validation with BMDP PR

This example will serve three purposes. First of all, it will be the first example of a dataset whose response variable has more than two outcomes. Second, it will be an example of what the output looks like when all of the independent variables are categorical. And finally, it will validate the procedure by allowing the comparison of the ***NCSS*** output with that of the ***BMDP PR*** program which also performs multiple-group logistic regression. This example comes from the ***BMDP*** manual. The database containing the data used in this example is named NC CRIMINAL.

The NC CRIMINAL database contains data that will be used to study the relationship between a cases verdict and three factors: race, county, and type of offense. The variables that are on the database are as follows.

Count contains the number of individuals with the characteristics specified on that row.

Verdict is the response variable. Three outcomes are given in the database: *G* for guilty, *NG* for not guilty, and *NP* for not prosecuted.

Race gives the race of the individual. It has two values: *A* and *B*.

County refers to county in North Carolina in which the offense was considered. The possible values are: *Durham* and *Orange*.

Offense contains the particular offense that the individual was accused of. These are *Drunk*, *Violence*, *Property*, *Major Traffic*, and *Speeding*.

You can view the data by loading the NC CRIMINAL database, so they will not be displayed here.

You may follow along here by making the appropriate entries or load the completed template **Example4** from the Template tab of the Logistic Regression window.

1 Open the NC CRIMINAL dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **NC CRIMINAL.S0**.
- Click **Open**.

2 Open the Logistic Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Logistic Regression window, select the **Variables tab**.
- Select **VERDICT(NP)** in the **Y: Group Variable** box. The *NP* value specifies that this category is to be used as the reference group.
- Enter the **RACE(A) COUNTY(DURHAM) OFFENSE(DRUNK)** in the **X's: Categorical Independent Variables** box. Note that the values in parentheses specify the reference value for each variable. These are specified so that the output will match that found in ***BMDP***.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select the variable **Count** from the list of variables and then click **Ok**. "Count" will appear in the Frequency Variable box.

4 Specify the Prior Probabilities.

- Select the **Model tab**.
- Set the **Prior Probabilities** to **Ni/N**. This indicates that the outcome frequencies found in the data will be used as the prior probabilities of group membership.

5 Specify the reports.

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **None**.
- Check the **Run Summary**, **Response Analysis**, **Parameter Significance Tests, Analysis of Deviance**, and **Log-Likelihood and R-Squared** reports. All other reports should be unchecked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**.

Selected portions of the output reports are shown next.

Logistic Regression Output

```
Run Summary Section
Parameter                    Value                Parameter               Value
Dependent Variable           Verdict              Rows Processed          60
Reference Value              NP                   Rows Used               57
Number of Values             3                    Rows for Validation     0
Frequency Variable           Count                Rows X's Missing        0
Numeric Ind. Variables       0                    Rows Freq Miss. or 0    3
Categorical Ind. Variables   3                    Rows Prediction Only    0
Final Log Likelihood         -408.29185           Unique Row Patterns     60
Model R-Squared              0.69779              Sum of Frequencies      615
Actual Convergence           4.751901E-11         Likelihood Iterations   6
Target Convergence           0.000001             Maximum Iterations      20
Model D.F.                   14                   Max Like Message        Normal Completion

Response Analysis Section
Verdict                             Unique                Act vs Pred     % Correctly
Categories          Count             Rows       Prior      R-Squared      Classified
G                     445               20     0.72358        0.17107          93.933
NG                    123               20     0.20000        0.10397          20.325
NP                     47               20     0.07642        0.06628           0.000
Total                 615               60                                     72.033

Parameter Significance Tests Section (Reference Value: Verdict = NP)
                  Regression                          Wald          Wald          Odds
                 Coefficient      Standard         Z-Value          Prob         Ratio
Parameter        (B or Beta)         Error        (Beta=0)         Level        Exp(B)
B0: Intercept
  G                  2.82983       0.44457           6.365       0.00000      16.94253
  NG                 1.24012       0.48781           2.542       0.01102       3.45604
B1: (County="ORANGE")
  G                 -0.89593       0.33719          -2.657       0.00788       0.40823
  NG                -0.12175       0.36036          -0.338       0.73547       0.88537
B2: (Offense="MJTRAFFIC")
  G                 -0.21380       0.62893          -0.340       0.73390       0.80751
  NG                 0.48012       0.67038           0.716       0.47387       1.61627
B3: (Offense="PROPERTY")
  G                 -0.91853       0.57784          -1.590       0.11193       0.39911
  NG                 0.00928       0.61911           0.015       0.98804       1.00932
B4: (Offense="SPEED")
  G                  0.49546       0.51245           0.967       0.33361       1.64126
  NG                -0.26697       0.57599          -0.463       0.64301       0.76570
B5: (Offense="VIOLENCE")
  G                 -2.23014       0.51372          -4.341       0.00001       0.10751
  NG                -0.57863       0.53748          -1.077       0.28168       0.56067
B6: (Race="B")
  G                  0.26083       0.33984           0.767       0.44279       1.29800
  NG                -0.10324       0.36248          -0.285       0.77579       0.90191

Analysis of Deviance Section                      Increase
                                                From Model
Term                                              Deviance          Prob
Omitted           DF          Deviance        (Chi Square)         Level
All               12         925.59805           109.01434       0.00000
COUNTY             2         832.03780            15.45409       0.00044
OFFENSE            8         898.18115            81.59744       0.00000
RACE               2         819.21845             2.63475       0.26784
None(Model)       12         816.58371
```

```
Log Likelihood & R-Squared Section
                                          R-Squared    Reduction      Reduction
Term(s)                         Log   Of Remaining   From Model  From Saturated
Omitted               DF  Likelihood      Term(s)    R-Squared       R-Squared
All                    2  -462.79903      0.00000
COUNTY                 2  -416.01890      0.59887      0.09892         0.40113
OFFENSE                8  -449.09057      0.17549      0.52230         0.82451
RACE                   2  -409.60923      0.68093      0.01686         0.31907
None(Model)           12  -408.29185      0.69779      0.00000         0.30221
None(Saturated)      120  -384.68551      1.00000                      0.00000
```

The output format is similar to previous examples. Notice in the analysis of deviance section that the variable race is not significant. That is, in these data, the race of the defendant is not related to the verdict.

The Parameter Significance Tests report combines the two logistic regression equations on one report. This makes it a bit more complicated to read, but it allows a quick comparison to be made of the corresponding regression coefficients. For each independent variable, the regression coefficient from each equation is shown. Thus, 2.82983 is the intercept for the *G* equation and 1.24012 is the intercept for the *NG* equation. Of course, no coefficient is show for *NP* because it is the reference value.

Also note that the definition of the binary variables is as before. Thus the parameter *B1: County=ORANGE* refers to a binary variable that was generated from the County variable. This binary variable is one when the county value is *ORANGE* and zero otherwise.

Validation

In order to validate this module, the estimated regression coefficients and the log likelihood generated by the ***BMDP*** (refer to page 1165 of version 7.0 of the ***BMDP*** manual) are displayed below.

Outcome: G	Coefficient	Std Error
1 RACE	0.2608	0.340
2 COUNTY	-0.8959	0.337
3 OFFENSE(1)	-2.230	0.514
4 OFFENSE(2)	-0.9185	0.578
5 OFFENSE(3)	-0.2138	0.629
6 OFFENSE(4)	0.4955	0.512
7 CONST1	2.830	0.445

Outcome: NG	Coefficient	Std Error
8 RACE	-0.1032	0.362
9 COUNTY	-0.1218	0.360
10 OFFENSE(1)	-0.5786	0.537
11 OFFENSE(2)	0.9281E-02	0.619
12 OFFENSE(3)	0.4801	0.670
13 OFFENSE(4)	-0.2670	0.576
14 CONST1	1.240	0.488

As you can see, these results match those displayed by ***NCSS*** exactly.

Example 5 – Logit Model with Interaction

This example continues with the analysis of the data given in Example 4. In that example, no interactions were included in the model. This example will include the two-way interactions in the model.

You may follow along here by making the appropriate entries or load the completed template **Example5** from the Template tab of the Logistic Regression window.

1 **Open the NC CRIMINAL dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **NC CRIMINAL.S0**.
- Click **Open**.

2 **Open the Logistic Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Logistic Regression window, select the **Variables tab**.
- Select **VERDICT(NP)** in the **Y: Group Variable** box. The *NP* value specifies that this category is to be used as the reference value.
- Enter the **RACE(A) COUNTY(DURHAM) OFFENSE(DRUNK)** in the **X's: Categorical Independent Variables** box. Note that the values in parentheses specify the reference value for each variable. These are specified so that the output will match that found in ***BMDP***.
- Double-click in the **Frequency Variable** box. This will bring up the variable selection window.
- Select the variable **Count** from the list of variables and then click **Ok**. "Count" will appear in the Frequency Variable box.

4 **Specify the Model with Interaction and Prior Probabilities.**

- Select the **Model tab**.
- Set **Which Model Terms** to **Up to 2-Way**. This will include the two-way interactions in the model.
- Set the **Prior Probabilities** to **Ni/N**. This indicates that the outcome frequencies found in the data will be used as the prior probabilities of group membership.

5 **Specify the reports.**

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **None**.
- Check the **Parameter Significance Tests**, **Analysis of Deviance**, and **Log-Likelihood and R-Squared** reports. All other reports should be unchecked.

6 **Run the procedure.**

- From the Run menu, select **Run Procedure**.

Selected portions of the output reports are shown below.

Logistic Regression Output

```
Parameter Significance Tests Section (Reference Value: Verdict = NP)
                     Regression                         Wald         Wald         Odds
                    Coefficient     Standard         Z-Value         Prob        Ratio
Parameter           (B or Beta)        Error        (Beta=0)        Level       Exp(B)
B0: Intercept
  G                     2.00583      0.50400           3.980      0.00007      7.43225
  NG                    0.72258      0.57465           1.257      0.20860      2.05975
B1: (County="ORANGE")
  G                     0.14731      1.15368           0.128      0.89840      1.15871
  NG                    1.83395      1.18755           1.544      0.12251      6.25854
B2: (Offense="MJTRAFFIC")
  G                    -0.30745      1.10221          -0.279      0.78029      0.73532
  NG                   -0.25450      1.23436          -0.206      0.83665      0.77531
B3: (Offense="PROPERTY")
  G                    -0.72178      0.83542          -0.864      0.38760      0.48589
  NG                    0.35757      0.89267           0.401      0.68874      1.42985
B4: (Offense="SPEED")
  G                     1.93682      1.08041           1.793      0.07303      6.93666
  NG                    0.87254      1.19650           0.729      0.46586      2.39297
B5: (Offense="VIOLENCE")
  G                    -0.15836      0.87409          -0.181      0.85624      0.85354
  NG                    1.07460      0.91294           1.177      0.23916      2.92882
B6: (Race="B")
  G                     1.44835      0.86924           1.666      0.09567      4.25608
  NG                   -1.10628      1.08369          -1.021      0.30733      0.33079
B7: (County="ORANGE")*(Offense="MJTRAFFIC")
  G                     0.45137      1.52019           0.297      0.76653      1.57046
  NG                   -0.53668      1.61710          -0.332      0.73998      0.58469
B8: (County="ORANGE")*(Offense="PROPERTY")
  G                     0.04871      1.41697           0.034      0.97258      1.04992
  NG                   -2.10279      1.47544          -1.425      0.15410      0.12212
B9: (County="ORANGE")*(Offense="SPEED")
  G                    -1.39431      1.37573          -1.014      0.31082      0.24800
  NG                   -2.66093      1.48387          -1.793      0.07294      0.06988
B10: (County="ORANGE")*(Offense="VIOLENCE")
  G                    -2.42314      1.36627          -1.774      0.07614      0.08864
  NG                   -3.93664      1.38198          -2.849      0.00439      0.01951
B11: (County="ORANGE")*(Race="B")
  G                     0.19528      0.81517           0.240      0.81067      1.21566
  NG                    0.83286      0.85899           0.970      0.33225      2.29990
B12: (Offense="MJTRAFFIC")*(Race="B")
  G                    -1.17876      1.35078          -0.873      0.38285      0.30766
  NG                    1.16592      1.50638           0.774      0.43894      3.20886
B13: (Offense="PROPERTY")*(Race="B")
  G                    -0.83367      1.27452          -0.654      0.51305      0.43445
  NG                    1.35214      1.42888           0.946      0.34400      3.86569

B14: (Offense="SPEED")*(Race="B")
  G                    -1.78987      1.25551          -1.426      0.15398      0.16698
  NG                    0.24862      1.45010           0.171      0.86387      1.28225
B15: (Offense="VIOLENCE")*(Race="B")
  G                    -2.31322      1.19041          -1.943      0.05199      0.09894
  NG                    0.51640      1.30133           0.397      0.69150      1.67598
```

```
Analysis of Deviance Section                        Increase
                                                  From Model
Term                                                Deviance         Prob
Omitted                 DF        Deviance      (Chi Square)        Level
All                      2       925.59805         146.82239      0.00000
COUNTY                   2       788.31126           9.53560      0.00850
OFFENSE                  8       802.98614          24.21048      0.00211
RACE                     2       797.83870          19.06304      0.00007
COUNTY*OFFENSE           8       798.81172          20.03607      0.01020
COUNTY*RACE              2       780.53878           1.76312      0.41414
OFFENSE*RACE             8       795.98619          17.21053      0.02799
None(Model)             30       778.77566

Log Likelihood & R-Squared Section
                                                   R-Squared    Reduction       Reduction
Term(s)                                Log    Of Remaining   From Model  From Saturated
Omitted                 DF      Likelihood         Term(s)    R-Squared       R-Squared
All                     30      -462.79903         0.00000
COUNTY                   2      -394.15563         0.87877      0.06104         0.12123
OFFENSE                  8      -401.49307         0.78483      0.15497         0.21517
RACE                     2      -398.91935         0.81778      0.12202         0.18222
COUNTY*OFFENSE           8      -399.40586         0.81155      0.12825         0.18845
COUNTY*RACE              2      -390.26939         0.92852      0.01129         0.07148
OFFENSE*RACE             8      -397.99309         0.82964      0.11016         0.17036
None(Model)             30      -389.38783         0.93980      0.00000         0.06020
None(Saturated)        120      -384.68554         1.00000                      0.00000
```

Notice how the interactions are labeled. For example, B15 is labeled (OFFENSE=VIOLENCE)*(RACE=B). This interaction variable is generated by multiplying the binary variable defined by (OFFENSE=VIOLENCE) with the binary variable defined by (RACE=B). The resulting variable is one if both of these conditions are true and zero otherwise.

Note that the *R*-squared is now 0.93980, so this model is almost as good as the saturated model.

Looking at the analysis of deviance table, we note that all terms are significant except for the County*Race interaction.

Example 6 – Odds Ratios

Lachin (2000) pages 90, 91, and 257 presents an analysis of hypothetical data from an ulcer healing clinical trial conducted to study the effectiveness of a drug over a placebo. There were 100 patients assigned to the group receiving the drug and another 100 patients assigned to the group receiving the placebo. The ulcers were stratified into one of three types: 1. Acid-dependent, 2. Drug dependent, and 3. Intermediate. Each ulcer was followed for a period of time after which it was considered healed or not. The data for this experiment are given below. These data have been entered into a database named LACHIN91.

LACHIN91 dataset (subset)

Count	Ulcer	Drug	Healed
16	1	1	1
26	1	1	0
20	1	0	1
27	1	0	0
9	2	1	1
3	2	1	0
4	2	0	1
5	2	0	0
28	3	1	1
18	3	1	0
16	3	0	1
28	3	0	0

You may follow along here by making the appropriate entries or load the completed template **Example6** from the Template tab of the Logistic Regression window.

1 **Open the LACHIN91 dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LACHIN91.S0**.
- Click **Open**.

2 **Open the Logistic Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Logistic Regression window, select the **Variables tab**.
- Set the **Y: Group Variable** box to **HEALED(0).** The zero in parentheses indicates that the value “0” is to be the reference value.
- Set the **X’s Categorical Independent Variables** box to **ULCER(1) DRUG(0).** The numbers in parentheses indicate the reference values of the two variables.
- Set the **Frequency Variable** box to **Count.**

4 Specify the reports.

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **None**.
- Check the **Run Summary**, **Parameter Significance Tests**, **Analysis of Deviance**, **Odds Ratios**, **Write Estimated Model**, and **Log-Likelihood and R-Squared** reports. All other reports should be unchecked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**.

Selected portions of the output reports are shown below.

Logistic Regression Output

```
Parameter Significance Tests Section (Reference Value: Healed = 0)
                Regression                      Wald        Wald         Odds
               Coefficient    Standard       Z-Value        Prob        Ratio
Parameter      (B or Beta)       Error      (Beta=0)       Level       Exp(B)
B0: Intercept     -0.48951     0.21833        -2.242     0.02496      0.61293
B1: (Drug=1)       0.50234     0.28845         1.742     0.08159      1.65259
B2: (Ulcer=2)      0.83527     0.50247         1.662     0.09645      2.30543
B3: (Ulcer=3)      0.32777     0.30424         1.077     0.28132      1.38787

Odds Ratios Section (Reference Value: Healed = 0)
                Regression        Odds     Lower 95%        Upper 95%
               Coefficient       Ratio    Confidence       Confidence
Parameter      (B or Beta)      Exp(B)         Limit            Limit
B0: Intercept     -0.48951     0.61293       0.39955          0.94027
B1: (Drug=1)       0.50234     1.65259       0.93894          2.90864
B2: (Ulcer=2)      0.83527     2.30543       0.86109          6.17243
B3: (Ulcer=3)      0.32777     1.38787       0.76451          2.51949

Analysis of Deviance Section
                                           Increase
                                         From Model
Term                                       Deviance         Prob
Omitted         DF         Deviance    (Chi Square)        Level
All              3        276.27807         6.58746      0.08628
DRUG             1        272.74521         3.05460      0.08051
ULCER            2        272.87155         3.18094      0.20383
None(Model)      3        269.69061
```

We note that neither DRUG nor ULCER is statistically significant at the 0.05 level using either the deviance tests in the Analysis of Deviance table or the Wald tests in the Parameter Significance Tests section. From the Odds Ratios section, we see that the odds of healing are increased 1.65259 when the drug is administered.

Example 7 – Matched Case-Control Study

Matched case-control studies should be analyzed using *conditional logistic regression*, a technique not currently available in ***NCSS***. However, 1:1 matched case-control studies may be analyzed using ***NCSS***. This type of design occurs when only one control is matched with each case. Collett (1991) describes the steps needed to analyze a 1:1 match case-control study using a regular logistic regression program. We will describe these steps using the same dataset as Collett (1991).

A matched case-control study was conducted to look at the impact of driving habits and place of residence on lower-back pain. A total of 217 matched pairs were recruited. In each pair, one individual was diagnosed as having an acute herniated disc (the case) and the other did not (the control). Controls were matched with cases on the basis of age (within ten years) and sex. The results were tabulated into the first five columns of the following dataset. These data have been entered into a database named COLLETT266.

COLLETT266 dataset

Count	Case Driver	Cntl Driver	Case Sub	Cntl Sub	Case DS	Cntl DS	Driver	Sub	DS
9	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	1	0
14	1	0	0	0	0	0	1	0	0
22	1	0	1	0	1	0	1	1	1
2	0	0	1	1	0	0	0	0	0
1	1	0	0	1	0	0	1	-1	0
4	1	0	1	1	1	0	1	0	1
10	0	1	0	0	0	0	-1	0	0
1	0	1	1	0	0	0	-1	1	0
20	1	1	0	0	0	0	0	0	0
32	1	1	1	0	1	0	0	1	1
7	0	1	0	1	0	1	-1	-1	-1
1	0	1	1	1	0	1	-1	0	-1
29	1	1	0	1	0	1	0	-1	-1
63	1	1	1	1	1	1	0	0	0
63	1	1	1	1	1	1	1	0	0

The columns in this table are defined as follows.

Count is the number of pairs with the indicated characteristics.
Case Driver is 1 if the case individual was a driver and 0 if not.
Cntl Driver is 1 if the control individual was a driver and 0 if not.
Case Sub is 1 if the case individual was a suburban resident and 0 if they lived in the city.
Cntl Sub is 1 if the control individual was a suburban resident and 0 if they lived in the city.
Case DS is the product of *CaseDrv* and *CaseSub*. This measures the case interaction.
Cntl DS is the product of *CntlDrv* and *CntlSub*. This measures the control interaction.
Driver is the difference between *CaseDrv* and *CntlDrv*.
Sub is the difference between *CaseSub* and *CntlSub*.
DS is the difference between *CaseDS* and *CntlDS*.

Only the last three columns are used in the analysis. A column of 1’s is added at the end of the dataset and labeled *Y*. This is the dependent variable. ***NCSS*** automatically adds a second group with a group value of zero. This group will be empty, but it is necessary to complete the analysis.

The method given by Collett (1991) is to use the differences between the case and control independent variable values as the regressor variables in a logistic regression. Also, the intercept term is not included in the model. We will do this in the following example.

You may follow along here by making the appropriate entries or load the completed template **Example7** from the Template tab of the Logistic Regression window.

1 Open the COLLETT266 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **COLLETT266.S0**.
- Click **Open**.

2 Open the Logistic Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Logistic Regression**. The Logistic Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Logistic Regression window, select the **Variables tab**.
- Set the **Y: Group Variable** box to **Y**. Note that the reference value will be an imaginary group zero.
- Set the **X’s Numeric Independent Variables** box to **Driver-DS**.
- Set the **Frequency Variable** box to **Count.**

4 Specify the model.

- On the Logistic Regression window, select the **Model tab**.
- Make sure the **Include Intercept** box is **not checked**.

5 Specify the reports.

- Select the **Reports tab**.
- Set the options **Row Classification Report**, **Row Classification Probabilities Report**, and **Simple Residuals Report** to **None**.
- Check the **Run Summary**, **Parameter Significance Tests**, **Analysis of Deviance**, **Odds Ratios**, **Write Estimated Model**, and **Log-Likelihood and R-Squared** reports. All other reports should be unchecked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**.

Selected portions of the output reports are shown below.

Logistic Regression Output

```
Analysis of Deviance Section
                                              Increase
                                            From Model
Term                                          Deviance         Prob
Omitted           DF          Deviance    (Chi Square)        Level
DRIVER             1         296.14391         4.93901      0.02626
DS                 1         291.28031         0.07542      0.78361
SUB                1         291.57933         0.37444      0.54060
None(Model)        3         291.20489

Parameter Significance Tests Section (Reference Value: Y = 0)
               Regression                       Wald          Wald          Odds
               Coefficient      Standard     Z-Value          Prob         Ratio
Parameter      (B or Beta)         Error    (Beta=0)         Level        Exp(B)
B1: Driver         0.69131       0.31893       2.168       0.03019       1.99633
B2: DS            -0.20579       0.66418      -0.310       0.75668       0.81400
B3: Sub            0.44385       0.72034       0.616       0.53778       1.55869

Odds Ratios Section (Reference Value: Y = 0)
               Regression          Odds      Lower 95%        Upper 95%
               Coefficient        Ratio     Confidence       Confidence
Parameter      (B or Beta)       Exp(B)          Limit            Limit
B1: Driver         0.69131       1.99633       1.06846          3.72998
B2: DS            -0.20579       0.81400       0.22145          2.99212
B3: Sub            0.44385       1.55869       0.37985          6.39607
```

The first step is to test the significance of the interaction term, DS. The deviance value, 0.07542, is not significant, so we decide to make another run without the interaction to enable us to more directly study the main effects: DRIVER and SUB.

Rerunning without the interaction produces the following report.

```
Parameter Significance Tests Section (Reference Value: Y = 0)
               Regression                       Wald          Wald          Odds
               Coefficient      Standard     Z-Value          Prob         Ratio
Parameter      (B or Beta)         Error    (Beta=0)         Level        Exp(B)
B1: Driver         0.65787       0.29398       2.238       0.02523       1.93068
B2: Sub            0.25546       0.22583       1.131       0.25797       1.29106

Odds Ratios Section (Reference Value: Y = 0)
               Regression          Odds      Lower 95%        Upper 95%
               Coefficient        Ratio     Confidence       Confidence
Parameter      (B or Beta)       Exp(B)          Limit            Limit
B1: Driver         0.65787       1.93068       1.08512          3.43514
B2: Sub            0.25546       1.29106       0.82931          2.00990

Analysis of Deviance Section
                                              Increase
                                            From Model
Term                                          Deviance         Prob
Omitted           DF          Deviance    (Chi Square)        Level
DRIVER             1         296.53786         5.25755      0.02185
SUB                1         292.56797         1.28766      0.25648
None(Model)        2         291.28031
```

The deviance tests indicate that DRIVER is significant, but Sub (suburban residence) is not. The point estimate for the odds ratio associated with driver is 1.93068. The 95% confidence interval for the odds ratio of DRIVER is 1.085 to 3.435. We conclude that the risk of a herniated disc is about twice as much for drivers as for non-drivers.

Chapter 325

Poisson Regression

Introduction

Poisson regression is similar to regular multiple regression except that the dependent (*Y*) variable is an observed count that follows the Poisson distribution. Thus, the possible values of *Y* are the nonnegative integers: 0, 1, 2, 3, and so on. It is assumed that large counts are rare. Hence, Poisson regression is similar to logistic regression, which also has a discrete response variable. However, the response is not limited to specific values as it is in logistic regression.

One example of an appropriate application of Poisson regression is a study of how the colony counts of bacteria are related to various environmental conditions and dilutions. Another example is the number of failures for a certain machine at various operating conditions. Still another example is vital statistics concerning infant mortality or cancer incidence among groups with different demographics.

Most books on regression analysis briefly discuss Poisson regression. We are aware of only one book that is completely dedicated to the discussion of the topic. This is the book by Cameron and Trivedi (1998). Most of the methods presented here were obtained from their book.

This program computes Poisson regression on both numeric and categorical variables. It reports on the regression equation as well as the goodness of fit, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform a subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values.

The Poisson Distribution

The Poisson distribution models the probability of *y* events (i.e. failure, death, or existence) with the formula

$$\Pr(Y = y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!} \qquad (y = 0,1,2,\ldots)$$

Notice that the Poisson distribution is specified with a single parameter μ. This is the mean incidence rate of a rare event per unit of *exposure*. Exposure may be time, space, distance, area, volume, or population size. Because exposure is often a period of time, we use the symbol *t* to represent the exposure. When no exposure value is given, it is assumed to be one.

The parameter μ may be interpreted as the risk of a new occurrence of the event during a specified exposure period, *t*. The probability of *y* events is then given by

$$\Pr(Y = y \mid \mu, t) = \frac{e^{-\mu t}(\mu t)^y}{y!} \qquad (y = 0,1,2,\ldots)$$

The Poisson distribution has the property that its mean and variance are equal.

The Poisson Regression Model

In Poisson regression, we suppose that the Poisson incidence rate μ is determined by a set of *k* regressor variables (the *X's*). The expression relating these quantities is

$$\mu = t\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Note that often, $X_1 \equiv 1$ and β_1 is called the *intercept*. The regression coefficients $\beta_1, \beta_2, \cdots, \beta_k$ are unknown parameters that are estimated from a set of data. Their estimates are labeled $b_1, b_2, \cdots, b_k$.

Using this notation, the fundamental Poisson regression model for an observation *i* is written as

$$\Pr(Y_i = y_i \mid \mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!}$$

where

$$\begin{aligned}\mu_i &= t_i \mu(\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}) \\ &= t_i \exp(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\end{aligned}$$

That is, for a given set of values of the regressor variables, the outcome follows the Poisson distribution.

Solution by Maximum Likelihood Estimation

The regression coefficients are estimated using the method of maximum likelihood. The logarithm of the likelihood function is

$$\ln[L(\mathbf{y}, \boldsymbol{\beta})] = \sum_{i=1}^{n} y_i \ln[t_i \mu(\mathbf{x}_\mathbf{i}'\boldsymbol{\beta})] - \sum_{i=1}^{n} t_i \mu(\mathbf{x}_\mathbf{i}'\boldsymbol{\beta}) - \sum_{i=1}^{n} \ln(y_i!)$$

Note that some statistical packages ignore the last term since it does not involve the regression parameters. This will make their calculated log-likelihoods different from ours.

The likelihood equations may be formed by taking the derivatives with respect to each regression coefficient and setting the result equal to zero. Doing this leads to a set of nonlinear equations that admits no closed-form solution. Thus, an iterative algorithm must be used to find the set of regression coefficients that maximum the log-likelihood. Using the method of iteratively reweighted least squares, a solution may be found in five or six iterations. However, the algorithm requires a complete pass through the data at each iteration, so it is relatively slow for problems with a large number of rows. With today's computers, this is becoming less and less of an issue.

Distribution of the MLE's

Applying the usual maximum likelihood theory, the asymptotic distribution of the maximum likelihood estimates (MLE's) is multivariate normal. That is,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\beta} V_{\hat{\boldsymbol{\beta}}})$$

where

$$V_{\hat{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1}$$

Remember that in the Poisson model the mean and the variance are equal. In practice, the data almost always reject this restriction. Usually, the variance is greater than the mean—a situation called *overdispersion*. The increase in variance is represented in the model by a constant multiple of the variance-covariance matrix. That is, we use

$$V_{\hat{\boldsymbol{\beta}}} = \phi \left(\sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1}$$

where ϕ is estimated using

$$\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

NCSS provides the option of using ϕ (phi) in the calculation of the variances of the regression coefficients.

Goodness of Fit Tests

Overall performance of the model is measured by two chi-square tests. These are the Pearson statistic

$$P_{\mathrm{P}} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

and the deviance, or *G*, statistic

$$D_{\mathrm{P}} = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right\}$$

Both of these statistics are approximately chi-square distributed with *n* - *k* degrees of freedom. When a test is rejected, there is a significant lack of fit. When a test is not rejected, there is no evidence of lack of fit.

The Pearson statistic is only chi-square distributed when you are analyzing grouped data, so if you are not using a frequency variable, you should not use the Pearson statistic as a goodness of fit test. The Pearson statistic is often used as a test of overdispersion.

Deviance

The deviance is twice the difference between the maximum achievable log-likelihood and the log-likelihood of the fitted model. In multiple regression under normality, the deviance is the residual sum of squares. In the case of Poisson regression, the deviance is a generalization of the sum of squares. The formula for the deviance is

$$D(\mathbf{y},\hat{\boldsymbol{\mu}}) = 2\left\{LL_{\mathbf{y}} - LL_{\hat{\boldsymbol{\mu}}}\right\}$$

Pseudo R-Squared Measures

The *R*-squared statistic does not extend to Poisson regression models. Various pseudo *R*-squared tests have been proposed. These pseudo measures have the property that, when applied to the linear model, they match the interpretation of the linear model *R*-squared. In Poisson regression, the most popular pseudo *R*-squared measure is function of the log-likelihoods of three models:

$$R^2 = \frac{LL_{fit} - LL_0}{LL_{\max} - LL_0}$$

where

$$LL_0 = \sum_{i=1}^{n} y_i \ln[t_i \hat{\mu}] - \hat{\mu}\sum_{i=1}^{n} t_i - \sum_{i=1}^{n} \ln(y_i!) \quad \text{where} \quad \hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} t_i}$$

$$LL_{\max} = \sum_{i=1}^{n} y_i \ln[t_i y_i] - \sum_{i=1}^{n} t_i y_i - \sum_{i=1}^{n} \ln(y_i!)$$

$$LL_{fit} = \sum_{i=1}^{n} y_i \ln[t_i \hat{\mu}(\mathbf{x}_i'\boldsymbol{\beta})] - \sum_{i=1}^{n} t_i \hat{\mu}(\mathbf{x}_i'\boldsymbol{\beta}) - \sum_{i=1}^{n} \ln(y_i!)$$

Note that LL_0 is the log-likelihood of the intercept-only model, LL_{fit} is the log-likelihood of the current model, and $LL_{\max}$ is the maximum log-likelihood possible. The maximum log-likelihood occurs when the actual responses (the y_i 's) exactly equal the predicted responses (the μ_i 's).

Notice that this value of *R*-squared varies between zero and one, with a perfect fit occurring at one. Also note that it assumes that there is an intercept in the model. This may be an actual explicit intercept or an implicit intercept (as when you use a complete set of indicator variables to represent a categorical variable).

Residuals

As in any regression analysis, a complete residual analysis should be employed. This involves plotting the residuals against various other quantities such as the regressor variables (to check for outliers and curvature) and the response variable. Various residuals may be of interest. These will be presented next.

Raw Residual

The raw residual is the difference between the actual response and the estimated value from the model. Because in the Poisson case, the variance is equal to the mean, we expect that the variances of the residuals are unequal. This can lead to difficulties in the interpretation of the raw residuals. However, it is still popular. The formula for the raw residual is

$$r_i = y_i - \hat{\mu}_i$$

Pearson Residual

The Pearson residual corrects for the unequal variance in the residuals by dividing by the standard deviation. The formula for the Pearson residual is

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i}}$$

Deviance Residual

The deviance residual is another popular residual. It is popular because the sum of squares of these residuals is the deviance statistic. The formula for the deviance residual is

$$d_i = sign(y_i - \hat{\mu}_i)\sqrt{2\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}}$$

Hat Values

The Hat matrix is used in residual diagnostics to measure the influence of each observation. The hat values, h_{ii} , are the diagonal entries of the Hat matrix which is calculated using

$$H = W^{1/2}X\left(X'WX\right)^{-1}X'W^{1/2}$$

where W is a diagonal matrix made up of $\hat{\mu}_i$.

The hat values should be studied themselves, to understand which observations have a large influence on the fitted regression coefficients. Large hat values are those that are larger than 2*k*/*n*. They are also used to further standardize residuals as is shown next.

Studentized Pearson Residual

The formula for the studentized Pearson residual is

$$sp_i = \frac{p_i}{\sqrt{1 - h_{ii}}}$$

Studentized Deviance Residual

The formula for the studentized deviance residual is

$$sd_i = \frac{d_i}{\sqrt{1 - h_{ii}}}$$

Subset Selection

Subset selection refers to the task of finding a small subset of the available regressor variables that does a good job of predicting the dependent variable. Because Poisson regression must be solved iteratively, the task of finding the best subset can be time consuming. Hence, techniques which look at all possible combinations of the regressor variables are not feasible. Instead, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Before discussing the details of these two algorithms, it is important to comment on a couple of issues that can come up. The first issue is what to do about the binary variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms used here search on model terms rather than on the individual variables. Thus, the whole set of binary variables associated with a given term are considered together for inclusion in, or deletion from, the model. Its all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and designate them as Numeric Variables.

Hierarchical Models

A second issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to only consider hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of *R*-squared. Enter this term into the model.
3. Continue adding terms until a preset limit on the maximum number of terms in the model is reached.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations so that other, more time consuming methods, are not feasible, or when you have far too many possible regressor variables and you want to reduce the number of terms in the selection pool.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of *R*-squared. If a switch can be found, it is made and the candidate terms are again searched to determine if another switch can be made.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some F-to-enter and F-to-remove tests whose properties are not well understood to begin with.

Data Structure

At a minimum, datasets to be analyzed by Poisson regression must contain a dependent variable and one or more independent variables. For each categorical variable, the program generates a set of binary (0 and 1) variables that express the same information. For example, in the table below, the discrete variable AgeGroup will be replaced by the variables Ag2 through Ag6 (Ag1 is not needed).

Koch et. al. (1986) present the following data taken from the Third National Cancer Survey. This dataset contains the number of new melanoma cases in 1969-1971 among white males in two areas for various age groups. The size of the estimated population at risk is given in the variable Population.

KOCH36 dataset

Melanoma	Area	AgeGroup	Population	AG1	AG2	AG3	AG4	AG5	AG6
61	0	<35	2880262	1	0	0	0	0	0
76	0	35-44	564535	0	1	0	0	0	0
98	0	45-54	592983	0	0	1	0	0	0
104	0	54-64	450740	0	0	0	1	0	0
63	0	65-74	270908	0	0	0	0	1	0
80	0	>74	161850	0	0	0	0	0	1
64	1	<35	1074246	1	0	0	0	0	0
75	1	35-44	220407	0	1	0	0	0	0
68	1	45-54	198119	0	0	1	0	0	0
63	1	54-64	134084	0	0	0	1	0	0
45	1	65-74	70708	0	0	0	0	1	0
27	1	>74	34233	0	0	0	0	0	1

Missing Values

If missing values are found in any of the independent variables being used, the row is omitted. If only the value of the dependent variable is missing, that row will not be used during the estimation process, but its predicted value will be generated and reported on.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable

Specify the dependent (response) variable. This is the variable to be predicted by the independent variables. The values in this variable should be non-negative integers (zero is okay).

Frequency Variable

Frequency Variable

This is an optional variable containing the frequency (observation count) for each row. Usually, you would leave this option blank and let each row receive the default frequency of one.

If your data have already been summarized, this option lets you specify how many actual rows each physical row represents.

Numeric Independent Variables

X's: Numeric Independent Variables

Specify the numeric (continuous) independent variables. By numeric, we mean that the values are numeric and at least ordinal. Nominal variables, even when coded with numbers, should be specified as Categorical Independent Variables. Although you may specify binary (0-1) variables here, they are better analyzed when you specify them as Categorical Independent Variables.

If you want to create powers and cross-products of these variables, specify an appropriate model in the 'Custom Model' field under the Model tab.

If you want to create predicted values of *Y* for values of *X* not in your database, add the *X* values to the bottom of the database. They will not be used during estimation, but predicted values will be generated for them.

Categorical Independent Variables

X's: Categorical Independent Variable(s)

Specify categorical (nominal) independent variables in this box. By categorical we mean that the variable has only a few unique, numeric or text, values like 1, 2, 3 or Yes, No, Maybe. The values are used to identify categories.

The values in a categorical variable are not used directly in the regression analysis. Instead, a set of numeric variables is substituted for them. Suppose a categorical variable has *G* categories. ***NCSS*** automatically generates the *G*-1 indicator variables that are needed for the analysis. The type of indicator variable created is determined by the selection for the *Default Reference Value* and the *Default Contrast Type*. The type of indicator created can also be controlled by entering the reference value and contrast type directly according to the syntax below. See the Default Reference Value and Default Contrast Type sections below for a discussion of the reference value and contrast type options.

You can create the interactions among these variables automatically using the *Custom Model* field under the Model tab.

Syntax

The syntax for specifying a categorical variable is *VarName*(*RefValue*;*CType*) where *VarName* is the name of the variable, *RefValue* is the reference value, and *CType* is the type of numeric variables generated: B for binary, P for polynomial, R for contrast with the reference value, and S for a standard set of contrasts.

For example, suppose a categorical variable, STATE, has four values: Texas, California, Florida, and New York. To process this variable, the values are arranged in sorted order: California, Florida, New York, and Texas. Next, the reference value is selected. If a reference value is not specified, the default value specified in the *Default Reference Value* window is used. Finally, the method of generating numeric variables is selected. If such a method is not specified, the contrast type selected in the *Default Contrast Type* window is used. Possible ways of specifying this variable are

STATE	**RefValue = Default, CType = Default**
STATE(New York)	**RefValue = New York, CType = Default**
STATE(California;R)	**RefValue = California, CType = Contrast with Reference**
STATE(Texas;S)	**RefValue = Texas, CType = Standard Set**

More than one category variable may be designated using a list. Examples of specifying three variables with various options are shown next.

```
STATE  BLOODTYPE  GENDER
STATE(California;R)  BLOODTYPE(O)  GENDER(F)
STATE(Texas;S)  BLOODTYPE(O;R)  GENDER(F;B)
```

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting**

  Use the first value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**

  Use the last value in alpha-numeric sorted order as the reference value.

The reference value may also be designated within parentheses after the name of the categorical independent variable, in which case the default reference value is ignored. For example, suppose that the categorical independent variable, STATE, has four values: 1, 3, 4, and 5.

1. If this option is set to 'First Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 1.
2. If this option is set to 'Last Value after Sorting' and the categorical independent variable is entered as 'STATE', the reference value would be 5.
3. If the categorical independent variable is entered as 'STATE(4)', the choice for this setting would be ignored, and the reference value would be 4.

Default Contrast Type

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to something other than 'Binary'.

- **Binary (This is the default)**

  Categories are converted to numbers using a set of binary indicator variables by assigning a '1' to the active category and a '0' to all other values. For example, suppose a categorical variable has G categories. ***NCSS*** automatically generates the G-1 binary (indicator) variables that are used in the regression. These indicator variables are set to 1 for those rows in which the value of this variable is equal to a certain value. They are set to 0 otherwise. The G-1 occurs because the Gth indicator variable is redundant (when all G-1 indicators are 0, wIfe know that the Gth indicator variable would be a 1). The value that is skipped is called the Reference Value.

  If your model includes interactions, using the binary indicator type may cause strange results.

  For the STATE variable, three binary variables would be generated. Suppose that the *Default Contrast Type* was 'Binary' and the statement used was 'STATE(Florida)'. The categories would be converted to numbers as follows:

STATE	B1	B2	B3
California	1	0	0
Florida	0	0	0
New York	0	1	0
Texas	0	0	1

- **Contrast with Reference**

  Categories are converted to numbers using a set of contrast variables by assigning a '1' to the active category, a '-1' to the reference value, and a '0' to all other values. A separate contrast is generated for each value other than the reference value.

  For the STATE variable, three numeric variables would be generated. Suppose the *Default Contrast Type* was 'Contrast with Reference', the *Default Reference Type* was 'Last Value after Sorting', and the variable was entered as 'STATE'. The categories would be converted to numbers as follows:

STATE	R1	R2	R3
California	1	0	0
Florida	0	1	0
New York	0	0	1
Texas	-1	-1	-1

- **Polynomial**

  If a variable has five or fewer categories, it can be converted to a set of polynomial contrast variables that account for the linear, quadratic, cubic, quartic, and quintic relationships. Note that these assignments are made after the values are sorted. Usually, the polynomial method is used on a variable for which the categories represent the actual values. That is, the values themselves are ordinal, not just category identifiers. Also, it is assumed that these values are equally spaced. Note that with this method, the reference value is ignored.

  For the STATE variable, linear, quadratic, and cubic variables are generated. Suppose that the *Default Contrast Type* was 'Polynomial' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	Linear	Quadratic	Cubic
California	-3	1	-1
Florida	-1	-1	3
New York	1	-1	-3
Texas	3	1	1

- **Standard Set**

  A variable can be converted to a set of contrast variables using a standard set of contrasts. This set is formed by comparing each value with those below it. Those above it are ignored. Note that these assignments are made after the values are sorted. The reference value is ignored.

  For the STATE variable, three numeric variables are generated. Suppose that the *Default Contrast Type* was 'Standard Set' and the statement used was 'STATE'. The categories would be converted to numbers as follows:

STATE	S1	S2	S3
California	-3	0	0
Florida	1	-2	0
New York	1	1	-1
Texas	1	1	1

Exposure Variable

T: Exposure Variable

Specify an optional variable containing exposure values. If this option is left blank, all exposures will be set to 1.0. This variable is specified when the exposures are different for each row.

The exposure is the amount of time, space, distance, volume, or population size from which the dependent variable is counted. For example, exposure may be the time in days, months, or years during which the values on that row were obtained. It may be the number of individuals at risk or the number of man-years from which the dependent variable is measured.

Each exposure must be a positive (non-zero) number or the row is ignored during the estimation phase.

Options

Alpha Level

Alpha is the significance level used in the hypothesis tests. One minus alpha is the confidence level of the confidence intervals. A value of 0.05 is most commonly used. This corresponds to a chance of error of 1 in 20. You should not be afraid to use other values since 0.05 became popular in pre-computer days when it was the only value available.

Typical values range from 0.001 to 0.20.

Use Dispersion Phi in SE's

Indicate whether to use the Phi multiplier in the calculation of the standard errors of the regression coefficients.

The Poisson model assumes that the mean and variance are identical. Usually, the variance is larger than the mean (called *overdispersion*). A correction can be applied to the standard errors by multiplying them by the Phi coefficient.

Note that this correction will not change the estimated regression coefficients.

Model Tab

These options control the regression model.

Subset Selection

Subset Selection

This option specifies the subset selection algorithm used to reduce the number of independent variables that used in the regression model. Note that since the solution algorithm is iterative, the selection process can be very time consuming. The Forward algorithm is much quicker than the Forward with Switching algorithm, but the Forward algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the generated individual binary variables. That is, either all binary variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **None**

  No subset selection is attempted. All specified independent variables are used in the regression equation.

- **(Hierarchical) Forward**

  With this algorithm, the term with the largest log likelihood is entered into the model. Next, the term that increases the log likelihood the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

  When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the log likelihood does not change significantly.

- **(Hierarchical) Forward with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest log likelihood is entered into the regression model. The term which increases the log likelihood the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, the likelihood function is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in the log likelihood. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Max Terms in Subset

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of the log likelihood.

Note that the intercept is counted in this number.

Estimation Options

The following options are used during the likelihood maximization process.

Maximum Iterations

Specifies the maximum number of iterations allowed during the iteration procedure. If this number is reached, the procedure is terminated prematurely. Typically, the maximum likelihood procedure converges in five or six iterations, so a value of twenty here should be ample.

Convergence Zero

This option specifies the convergence target for the maximum likelihood estimation procedure. When all of the maximum likelihood equations are less than this amount, the algorithm is assumed to have converged. In theory, all of the equations should be zero. However, about the best that can be achieved is 1E-13, so you should set this value to a number a little larger than this such as the default of 1E-9.

The actual value can be found by looking at the Maximum Convergence value on the Run Summary report.

Model Specification

Which Model Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a straight-forward regression model, select *Up to 1-Way*.

The options are:

- **Full Model**

  The complete, saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables).

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

  Note that the discussion of the Custom Model option discusses the interpretation of this model.

- **Up to 1-Way**

  This option generates a model in which each variable is represented by a single model term. No cross-products or interaction terms are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

  This is the option to select when you want to analyze the independent variables specified without adding any other terms.

  For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C

- **Up to 2-Way**

  This option specifies that all main effects and two-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C

- **Up to 3-Way**

  All main effects, two-way interactions, and three-way interactions are included in the model. For example, if you have three independent variables A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

- **Up to 4-Way**

  All main effects, two-way interactions, three-way interactions, and four-way interactions are included in the model. For example, if you have four independent variables A, B, C, and D, this would generate the model:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D.

- **Custom Model**

  The model specified in the *Custom Model* box is used.

Include Intercept

Check this box to include an intercept (constant term) in your model.

Under most circumstances, you will want to include an intercept. The only time you may not need an intercept is when you have generated a set of indicator variables for a discrete variable and you want to include all of them instead of omitting one of them.

Write Model in Custom Model Field

When this option is checked, no data analysis is performed when the procedure is run. Instead, a copy of the full model is stored in the Custom Model box. You can then edit the model as desired. This option is useful when you want to be selective about which terms to keep and you have several variables.

Note that the program will not do any calculations while this option is checked.

Model Specification – Custom Model

Max Term Order

This option specifies that maximum number of variables that can occur in an interaction term in a custom model. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would be excluded from the model.

This option is particularly useful when used with the bar notation of a custom model to allow a simple way to remove unwanted high-order interactions.

Custom Model

This options specifies a custom model. It is only used when the *Which Model Terms* option is set to *Custom Model*. A custom model specifies the terms (single variables and interactions) that are to be kept in the model.

Interactions

An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of indicator variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the indicator variables generated from the categorical variable.

Syntax

A model is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A + B + C + A*B + A*C + B*C + A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B + A*C.

Some examples will help to indicate how the model syntax works:

A|B = A + B + A*B

A|B A*A B*B = A + B + A*B + A*A + B*B

Note that you should only repeat numeric variable. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A + B + A*A + A*B + B*B

A|B|C = A + B + C + A*B + A*C + B*C + A*B*C

(A + B)*(C + D) = A*C + A*D + B*C + B*D

(A + B)|C = (A + B) + C + (A + B)*C = A + B + C + A*C + B*C

Reports Tab

The following options control which reports are displayed.

Select Reports – Summaries

Run Summary ... Means

Each of these options specifies whether the indicated report is calculated and displayed.

Select Reports – Subset Selection

Subset Selection - Summary and Subset Selection - Detail

Indicate whether to display these subset selection reports.

Select Reports – Estimation

Regression Coefficients ... Rate Coefficients

Indicate whether to display these estimation reports.

Select Reports – Goodness-of-Fit

Lack-of-Fit Tests ... Log-Likelihood and R-Squared

Indicate whether to display these model goodness-of-fit reports.

Select Reports – Row-by-Row Lists

Residuals ... Incidence

Indicate whether to display these list reports. Note that since these reports provide results for each row, they may be too long for normal use when requested on large databases.

Incidence Counts

Up to five incidence counts may be entered. The probabilities of these counts under the Poisson regression model will be displayed on the Incidence Report.

These values must be non-negative integers.

Exposure Value

Specify the exposure (time, space, distance, volume, etc.) value to be used as a multiplier on the Incidence Report. All items on that report are scaled to this amount. For example, if your data was scaled in terms of events per month but you want the Incidence report scaled to events per year, you would enter ‘12’ here.

Select Plots

Incidence (Y/T) vs X Plot ... Resid vs X Plot

Indicate whether to display these plots.

Plot Options

Residual Plotted

This option specifies which of the five types of residuals are shown on the residual plots.

Format Tab

These options control format of the reports.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Skip Line After

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Report Options – Decimal Places

Y ... Chi-Square Decimals

These options specify the number of decimal places shown on the reports for the indicated values.

Incidence vs X Plot to Resid vs X Plot Tabs

These options control the attributes of the various plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Storage Tab

These options let you specify if, and where on the database, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify variables that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the database when the procedure is run.

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**

  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Variable In

The first item is stored in this variable. Each additional item that is checked is stored in the variables immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these variables is automatically replaced, so be careful..

Data Storage Options – Select Items to Store

Expanded X Values ... Covariance Matrix

Indicated whether to store these row-by-row values, beginning at the variable indicated by the *Store First Variable In* option. Note that several of these values include a different value for each group and so they require several columns when they are stored.

Expanded X Values

This option refers to the experimental design matrix. They include all binary and interaction variables generated.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Poisson Regression using a Dataset with Indicator Variables

This section presents several examples. In the first example, the data shown earlier in the Data Structure section and found in the KOCH36 database will be analyzed. Koch et. al. (1986) presented this dataset. It contains the number of new melanoma cases in 1969-1971 among white males in two areas for various age groups. The size of the estimated population at risk is given in the variable Population.

This dataset is instructive because it shows how easily categorical variables are dealt with. In this example, two categorical variables, AREA and AGEGROUP, will be included in the regression model. The dataset can also be used to validate the program since the results are given in Koch (1986).

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Poisson Regression window.

1 **Open the KOCH36 database.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **KOCH36.S0**.
- Click **Open**.

2 **Open the Poisson Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines,** then **Poisson Regression**. The Poisson Regression procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Poisson Regression window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable** box. This will bring up the variable selection window.
- Select **Melanoma** from the list of variables and click **Ok**. “Melanoma” will appear in the Y: Dependent Variable box.
- Double-click in the **X’s: Categorical Independent Variables** box.
- Enter **Area(0) AgeGroup(<35)** in the **X’s: Categorical Independent Variables** box. The values in parentheses give the reference value for each variable.
- Double-click in the **T: Exposure Variable** box.
- Select **Population** from the list of variables and click **Ok**.
- The rest of this panel can be left at the default values.

4 **Specify the model.**

- Select the **Model tab**.
- Set the **Subset Selection** option to **None**.
- Set the **Which Model** option to **Up to 1-Way**.

5 Specify the reports.

- Select the **Reports tab**.
- Check all of the reports and plots. Normally, you would not want all of them, but we need them now so we can document them.
- Set the **Incidence Counts** to **5 10 15 20 25**.
- Set the **Exposure Value** to **100000**.

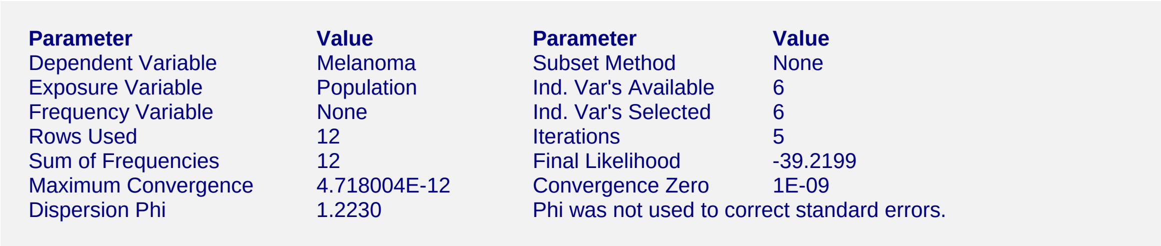
6 Specify the decimals.

- Select the **Format tab**.
- Set the number of **decimal places for Probability** to **6**.

7 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top) or press the F9 function key.

Run Summary Report



```
Parameter                Value              Parameter            Value
Dependent Variable       Melanoma           Subset Method        None
Exposure Variable        Population         Ind. Var's Available 6
Frequency Variable       None               Ind. Var's Selected  6
Rows Used                12                 Iterations           5
Sum of Frequencies       12                 Final Likelihood     -39.2199
Maximum Convergence      4.718004E-12       Convergence Zero     1E-09
Dispersion Phi           1.2230             Phi was not used to correct standard errors.
```

This report provides several details about the data and the MLE algorithm.

Dependent, Exposure, and Frequency Variables

These variables are listed to provide a record of the variables that were analyzed.

Rows Used

This is the number of rows used by the estimation algorithm. Rows with missing values and filtered rows are not included. Always check this value to make sure that you are analyzing all of the data you intended to.

Sum of Frequencies

This is the number of observations used by the estimation algorithm. If you specified a Frequency Variable, this will be greater than the number of rows. If not, they will be equal.

Subset Method

This is the type of subset selection that was run.

Ind. Var’s Available

This is the number of independent variables that you have selected.

No. of X’s in Model

This is the number of actual *X*-variables generated from the terms in the model that was used in the analysis.

Iterations

This is number of iterations used by the estimation algorithm. Usually, the algorithm will terminate in five or six iterations.

Maximum Convergence

The estimation algorithm continues until all of the likelihood equations are close to zero. This is largest value of all of these equations. It should be close to zero or the algorithm was terminated before it had converged.

Convergence Zero

The estimation algorithm continues until all of the likelihood equations are close to zero. This is *zero* to the algorithm. When the maximum convergence value is less than this amount, the algorithm has converged. Compare this value to the Maximum Convergence value.

Final Likelihood

This is the value of the log likelihood that was achieved for this run.

Dispersion Phi

This line gives the estimated value of the dispersion phi. It also indicates whether phi was used to adjust the standard errors of the regression coefficients and the predicted values.

Model Summary Section

```
                 Model    Error          Log                                  Pseudo
Model               DF       DF   Likelihood     Deviance          AIC     R-Squared
Intercept            1       11    -484.0223     895.8197     897.8197        0.0000
Model                7        5     -39.2199       6.2149      20.2149        0.9931
Maximum             12        0     -36.1125       0.0000      24.0000        1.0000
```

This report is analogous to the analysis of variance table. It summarizes the goodness of fit of the model.

Model

This is the term(s) that are reported about on this row of the report. Note that the model line includes the intercept.

Model DF

This is the number of variables in the model.

Error DF

This is the number of observations minus the number of variables.

Log Likelihood

This is the value of the log-likelihood function for the intercept only model, the chosen model, and the saturated model that fits the data perfectly. By comparing these values, you obtain an understanding of how well you model fits the data.

Deviance

The deviance is the generalization of the sum of squares in regular multiple regression. It measures the discrepancy between the fitted values and the data.

AIC

This is Akaike’s information criterion (AIC). It is equal to the deviance plus twice the number of parameters in the model. It combines a measure of the discrepancy between the fitted values and the data (the deviance) with a measure of the simplicity of the model (twice the number of parameters). It has been shown that using AIC to compare competing models with different numbers of parameters amounts to selecting the model with the minimum estimate of the mean squared error of prediction.

Pseudo R-Squared

This is the generalization of regular *R*-squared in multiple regression. This value is discussed in detail in the Technical Details section of the chapter. Its formula is

$$R^2 = \frac{LL_{fit} - LL_0}{LL_{\max} - LL_0}$$

Lack-of-Fit Tests Section

```
                            Chi^2               Prob
Test                DF      Value              Level
Pearson              5       6.12           0.295180
G Statistic          5       6.21           0.285867
These tests indicate whether there is a significant lack of fit to the data by the model.
```

This report provides the results of two goodness-of-fit tests. They indicate whether the current model adequately fits the data. The tests themselves are described in the Technical Details section of this chapter.

Test

Indicates which of the two tests is shown on this line. Note that the *G* Statistic test is more accurate in small samples. The Pearson test is often used as a test for overdispersion.

DF

Both of these tests are chi-square tests. This is the value of the degrees of freedom. It is equal to the number of observations minus the number of parameters in the regression model.

Chi^2 Value

This is the value of the chi-square test statistic.

Prob Level

This is the probability level of the test. The null hypothesis is that the model fits the data adequately. The alternative hypothesis is that the model is an inadequate representation of the data. If this probability level is less than some cutoff value such as 0.10 or 0.05, there is a significant lack of fit.

Means Report

```
Variable            Mean       Minimum      Maximum
Melanoma          68.667        27.000      104.000
Population    554422.917     34233.000  2880262.000
```

This report gives the mean, minimum, and maximum for each of the numeric variables in the analysis. Use it to check for obvious data errors.

Regression Coefficients Section

```
                  Regression                  Wald's                   Lower 95.0%   Upper 95.0%
Independent      Coefficient   Standard        Chi^2          Prob      Confidence    Confidence
Variable                 (B)      Error     (Ho:B=0)         Level           Limit         Limit
Intercept          -10.65831    0.09518     12538.43      0.000000       -10.84487     -10.47175
(AgeGroup="35-44")
                     1.79737    0.12093       220.92      0.000000         1.56036       2.03439
(AgeGroup="45-54")
                     1.91309    0.11844       260.90      0.000000         1.68095       2.14522
(AgeGroup="54-64")
                     2.24180    0.11834       358.89      0.000000         2.00987       2.47374
(AgeGroup="65-74")
                     2.36572    0.13152       323.56      0.000000         2.10795       2.62349
(AgeGroup=">74")
                     2.94468    0.13205       497.30      0.000000         2.68587       3.20349
(Area=1)             0.81948    0.07103       133.11      0.000000         0.68027       0.95870
Dispersion Phi                   1.2230

Estimated Poisson Regression Model
Exp( -10.6583092620666 + 1.79737495802663*(AgeGroup="35-44") + 1.91308772800916*(AgeGroup="45-54") +
2.24180245796945*(AgeGroup="54-64") + 2.36572417048965*(AgeGroup="65-74") +
2.94467922306083*(AgeGroup=">74") + .819484586814038*(Area=1) )
```

This report provides the estimated regression model and associated statistics. It provides the main results of the analysis.

Validation

Koch (1986) gives the following estimates and standard errors.

Independent Variable	ML Estimate	Standard Error
Intercept	-10.66	0.01
Area	0.82	0.07
AG2	1.80	0.12
AG3	1.91	0.12
AG4	2.24	0.12
AG5	2.37	0.13
AG6	2.94	0.13

As you can see, these results match those provided by ***NCSS*** exactly—validating our algorithms. These results were also validated using SAS.

Independent Variable

This item provides the name of the independent variable shown on this line of the report. The *Intercept* refers to the optional constant term. The *Dispersion Phi* is the estimated value of the phi coefficient.

Note that whether a line is skipped after the name of the independent variable is displayed is controlled by the Skip Lines After option in the Format tab.

Regression Coefficient

These are the maximum-likelihood estimates of the regression coefficients, $b_1, b_2, \cdots, b_k$. Their direct interpretation is difficult because the formula for the predicted value involves the exponential function.

Standard Error

These are the asymptotic standard errors of the regression coefficients, the s_{b_i}. The estimate the precision of the regression coefficient. The standard errors are the square roots of the diagonal elements of this covariance matrix. The covariance matrix is obtained by inverting the observed information matrix evaluated at the maximum likelihood estimates.

If you Use Dispersion Phi option, the corrected standard error is shown. This is found by multiplying the simple standard error by the square root of phi. That is, the value displayed is s'_{b_i} where

$$s'_{b_i} = s_{b_i}\sqrt{\phi}$$

Wald's Chi^2 (Ho:b=0)

This is the one degree of freedom chi-square statistic for testing the null hypothesis that $\beta_i = 0$ against the alternative that $\beta_i \neq 0$. The chi-square value is called a *Wald statistic*. This test has been found to follow the chi-square distribution only in large samples.

The test is calculated using

$$\chi_1^2 = \left(\frac{b_i}{s'_{b_i}}\right)^2$$

Prob Level

The probability of obtaining a chi-square value greater than the above. This is the significance level of the test. If this value is less than some predefined alpha level, say 0.05, the variable is said to be statistically significant.

Lower and Upper Confidence Limits

These provide a large-sample confidence interval for the values of the coefficients. The width of the confidence interval provides you with a sense of how precise the regression coefficients are. Also, if the confidence interval includes zero, the variable is not *statistically significant.* The formula for the calculation of the confidence interval is

$$b_i \pm z_{1-\alpha/2}s'_{b_i}$$

where $1-\alpha$ is the confidence coefficient of the confidence interval and *z* is the appropriate value from the standard normal distribution.

Dispersion Phi

This is the estimate of the overdispersion correction multiplier, phi. Remember that in the Poisson model the mean and the variance are equal. In practice, the data almost always reject this restriction. Usually, the variance is greater than the mean—a situation called *overdispersion*. The increase in variance is represented in the model by a constant multiple of the variance-covariance matrix. That is, we use

$$V_{\hat{\boldsymbol{\beta}}} = \phi\left(\sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1}$$

where ϕ is estimated using

$$\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Estimated Poisson Regression Model

This expression displays the estimated regression model in written format. It may be copied to the clipboard and used elsewhere. For example, you could copy it and paste it as a Variable Transformation.

Note that transformation must be less than 255 characters. Since this formula is often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file, including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

Analysis of Deviance Section

```
                                          Increase
                                        From Model
Term                                      Deviance         Prob
Omitted           DF       Deviance   (Chi Square)        Level
All                1       968.0446
AGEGROUP           5       875.1835         796.74     0.000000
AREA               1       202.6602         124.22     0.000000
None(Model)        7        78.4398
```

This report is the Poisson regression analog of the analysis of variance table. It displays the results of a chi-square test used to test whether each of the individual terms in the regression are statistically significant after adjusting for all other terms in the model.

This report is not produced during a subset selection run.

Note that this report requires that a separate regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.

The “All” line refers to the intercept-only model. The “None(Model)” refers to the complete model with no terms removed.

Note that it is usually not advisable to include an interaction term in a model when one of the associated main effects is missing—which is what happens here. However, in this case, we believe this to be a useful test.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the chi-square test displayed on this line.

Deviance

The deviance is equal to minus two times the log likelihood achieved by the model being described on this line of the report. See the discussion given earlier in this chapter for a technical discussion of the deviance. A useful way to interpret the deviance is as the analog of the residual sum of squares in multiple regression. This value is used to create the difference in deviance that is used in the chi-square test.

Increase From Model Deviance (Chi Square)

This is the difference between the deviance for the model described on this line and the deviance of the complete model. This value follows the chi-square distribution in medium to large samples. This value can the thought of as the analog of the residual sum of squares in multiple regression. Thus, you can think of this value as the increase in the residual sum of squares that occurs when this term is removed from the model.

Another way to interpret this test is as a redundancy test because it tests whether this term is redundant after considering all of the other terms in the model.

Prob Level

This is the significance level of the chi-square test. This is the probability that a chi-square value with degrees of freedom DF is equal to this value or greater. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

Log Likelihood & R-Squared Section

```
                                        R-Squared     Reduction      Reduction
Term(s)                          Log  Of Remaining    From Model  From Saturated
Omitted             DF    Likelihood      Term(s)     R-Squared      R-Squared
All                  1     -17.18588      0.00000
All                  1     -484.0223       0.0000
AGEGROUP             5     -437.5917       0.1037        0.8894         0.8963
AREA                 1     -101.3301       0.8544        0.1387         0.1456
None(Model)          7      -39.2199       0.9931        0.0000         0.0069
None(Saturated)     12      -36.1125       1.0000                       0.0000
```

This report provides the log likelihoods and *R*-squared values of various models. This report is not produced during a subset selection run.

Note that this report requires that a separate regression be run for each line. Thus, if the running time is too long, you might consider omitting this report.

Term Omitted

This is the term that is omitted from the model. The “All” line refers to the intercept-only model. The “None(Model)” refers to the complete model with no terms removed. The “None(Saturated)” line gives the results for the saturated model.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the term displayed on this line.

Log Likelihood

This is the log likelihood of the model displayed on this line. Note that this is the log likelihood of the regression without the term listed.

R-Squared of Remaining Term(s)

This is the *R*-squared of the model displayed on this line. Note that the model does not include the term listed at the beginning of the line.

Note that this is a pseudo *R*-squared as discussed earlier in this chapter.

Reduction From Model R-Squared

This is amount that *R*-squared is reduced when the term is omitted from the regression model. This reduction is calculated from the *R*-squared achieved by the full model.

This quantity is used to determine if removing a term causes a large reduction in *R*-squared. If it does not, then the term can be safely removed from the model.

Reduction From Saturated R-Squared

This is amount that *R*-squared is reduced when the term is omitted from the regression model. This reduction is calculated from the *R*-squared achieved by the saturated model. This item is included because it shows how removal of this term impacts the best *R*-squared that is possible.

Rate Section

```
                      Regression          Rate     Lower 95.0%     Upper 95.0%
Independent          Coefficient         Ratio      Confidence      Confidence
Variable                     (B)      [Exp(B)]           Limit           Limit
Intercept              -10.65831       0.00002         0.00002         0.00003
(AgeGroup="35-44")
                         1.79737       6.03379         4.76055         7.64756
(AgeGroup="45-54")
                         1.91309       6.77397         5.37066         8.54396
(AgeGroup="54-64")
                         2.24180       9.41028         7.46233        11.86672
(AgeGroup="65-74")
                         2.36572      10.65175         8.23138        13.78381
(AgeGroup=">74")
                         2.94468      19.00457        14.67098        24.61823
(Area=1)                 0.81948       2.26933         1.97442         2.60830
```

This report provides the rate ratio for each independent variable.

Independent Variable

This item provides the name of the independent variable shown on this line of the report. The *Intercept* refers to the optional constant term.

Regression Coefficient

These are the maximum-likelihood estimates of the regression coefficients, $b_1, b_2, \cdots, b_k$. Their direct interpretation is difficult because the formula for the predicted value involves the exponential function.

Rate Ratio

These are the exponentiated values of the regression coefficients. The formula used to calculate these is

$$RR_i = e^{b_i}$$

The rate ratio is mainly useful for interpretation of the regression coefficients of indicator variables. In this case, they estimate the incidence of the response variable (melanoma in this example) in the given category relative to the category whose indicator variable was omitted (usually called the *control* group).

Lower and Upper Confidence Limits

These provide a large-sample confidence interval for the rate ratios. The formula for the calculation of the confidence interval is

$$\exp\left(b_i \pm z_{1-\alpha/2} s'_{b_i}\right)$$

where $1-\alpha$ is the confidence coefficient of the confidence interval and *z* is the appropriate value from the standard normal distribution.

Covariances of Regression Coefficients Section

The covariance matrix of the regression coefficients is not displayed as a report. However, it may be stored on the database for further investigation and use.

The covariance matrix is obtained by inverting the observed information matrix evaluated at the maximum likelihood estimates. If the Use Dispersion Phi option was checked, the original values are multiplied by phi.

Residuals Section

```
           Melanoma     Predicted         Raw     Pearson    Deviance    Population
Row             (Y)         Value    Residual    Residual    Residual           (T)
1                61       67.6998     -6.6998     -0.8143     -0.8283       2880262
2                76       80.0638     -4.0638     -0.4542     -0.4581        564535
3                98       94.4150      3.5850      0.3690      0.3667        592983
4               104       99.6974      4.3026      0.4309      0.4279        450740
5                63       67.8263     -4.8263     -0.5860     -0.5932        270908
6                80       72.2979      7.7021      0.9058      0.8904        161850
7                64       57.3002      6.6998      0.8851      0.8686       1074246
8                75       70.9362      4.0638      0.4825      0.4780        220407
9                68       71.5850     -3.5850     -0.4237     -0.4273        198119
10               63       67.3026     -4.3026     -0.5245     -0.5302        134084
11               45       40.1737      4.8263      0.7614      0.7469         70708
12               27       34.7021     -7.7021     -1.3075     -1.3609         34233
```

This report provides the predicted values and various types of residuals. Large residuals indicate data points that were not fit well by the regression model. You may consider removing rows with large residuals and refitting, but you must be certain that you have a good reason for doing so. You cannot remove them simply because they have large residuals.

Row

The row number of the item. If you have excluded some rows by using a filter or if some of the rows had missing values, the row number identifies the original row on the database.

Y

This is the value of the dependent variable.

Predicted Value

This is the predicted value of *Y*. It is the Poisson incidence rate, $\hat{\mu}_i$, estimated by

$$\begin{aligned}\hat{\mu}_i &= t_i \hat{\mu}\left(\mathbf{x}_i'\mathbf{b}\right) \\ &= t_i \exp\left(b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}\right)\end{aligned}$$

Raw Residual

The raw residual is the different between the actual response and the estimated value from the model. The formula for the raw residual is

$$r_i = y_i - \hat{\mu}_i$$

Pearson Residual

The Pearson residual corrects for the unequal variance in the residuals by dividing by the standard deviation. The formula for the Pearson residual is

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i}}$$

Deviance Residual

The deviance residual is another popular residual. It is popular because the sum of squares of these residuals is the deviance statistic. The formula for the deviance residual is

$$d_i = sign\left(y_i - \hat{\mu}_i\right)\sqrt{2\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i - \hat{\mu}_i\right)\right\}}$$

where *sign(x)* is 1 if *x* is greater than or equal to 0 and -1 otherwise.

T

The value of the exposure variable (if active) is provided for your reference.

Predicted Values Section

Row	Melanoma (Y)	Predicted Value	Standard Error	Lower 95.0% Confidence Limit	Upper 95.0% Confidence Limit	Population (T)
1	61	67.6998	6.4440	55.0698	80.3297	2880262
2	76	80.0638	7.0419	66.2619	93.8657	564535
3	98	94.4150	7.8780	78.9743	109.8556	592983
4	104	99.6974	8.2257	83.5752	115.8195	450740
5	63	67.8263	6.7681	54.5610	81.0916	270908
6	80	72.2979	7.1850	58.2156	86.3802	161850
7	64	57.3002	5.5790	46.3656	68.2349	1074246
8	75	70.9362	6.3609	58.4691	83.4034	220407
9	68	71.5850	6.2636	59.3085	83.8615	198119
10	63	67.3026	5.9387	55.6630	78.9423	134084
11	45	40.1737	4.2609	31.8226	48.5249	70708
12	27	34.7021	3.7454	27.3612	42.0430	34233

This report provides the predicted values along with their standard errors and confidence limits.

If you want to generate predicted values and confidence limits for *X* values not on your database, you should add them to the bottom of the database, leaving *Y* blank (if you are using an exposure variable, set the value of *T* to a desired value). These rows will not be included in the estimation algorithm, but they will appear on this report with estimated *Y*'s.

Row

The row number of the item. If you have excluded some rows by using a filter or if some of the rows had missing values, the row number identifies the original row on the database.

Y

This is the value of the dependent variable.

Predicted Value

This is the predicted value of *Y*. It is the predicted mean of the Poisson distribution, $\hat{\mu}_i$, estimated by

$$\begin{aligned}\hat{\mu}_i &= t_i \hat{\mu}\left(\mathbf{x}_\mathbf{i}'\mathbf{b}\right) \\ &= t_i \exp\left(b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}\right)\end{aligned}$$

Standard Error

The standard error of the predicted value is a measure of the precision of the estimated value. The formula for the standard error is

$$se_{\hat{\mu}_i} = \hat{\mu}_i \sqrt{\mathbf{x}_i' \mathbf{V}_{\hat{\boldsymbol{\beta}}} \mathbf{x}_i'}$$

where

$$V_{\hat{\boldsymbol{\beta}}} = \phi \left(\sum_{i=1}^{n} \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1}$$

Note that if ϕ is not used, it is set to one in the above formulas.

Confidence Limits

These limits define a large-sample confidence interval for μ_i. The formula is

$$\hat{\mu}_i \pm (z_{1-\alpha/2})(se_{\hat{\mu}_i})$$

T

The value of the exposure variable (if active) is provided for you reference.

Residual Diagnostics Section

```
                                               Studentized  Studentized
         Melanoma    Predicted         Raw         Pearson     Deviance           Hat
Row           (Y)        Value    Residual        Residual     Residual      Diagonal
1              61      67.6998     -6.6998         -1.3095      -1.3321        0.6134
2              76      80.0638     -4.0638         -0.7361      -0.7425        0.6194
3              98      94.4150      3.5850          0.6303       0.6264        0.6573
4             104      99.6974      4.3026          0.7602       0.7548        0.6787
5              63      67.8263     -4.8263         -1.0285      -1.0411        0.6754
6              80      72.2979      7.7021          1.6939       1.6651        0.7140
7              64      57.3002      6.6998          1.3095       1.2852        0.5432
8              75      70.9362      4.0638          0.7361       0.7293        0.5704
9              68      71.5850     -3.5850         -0.6303      -0.6357        0.5481
10             63      67.3026     -4.3026         -0.7602      -0.7685        0.5240
11             45      40.1737      4.8263          1.0285       1.0089        0.4519
12             27      34.7021     -7.7021         -1.6939      -1.7632        0.4042
High Leverage Cutoff                                                         1.166667
```

This report provides the hat diagonals and studentized residuals. It allows you to study the leverage (influence) of each observation.

Row

The row number of the item. If you have excluded some rows by using a filter or if some of the rows had missing values, the row number identifies the original row on the database.

Y

This is the value of the dependent variable.

Predicted Value

This is the predicted value of Y. It is the Poisson incidence rate, $\hat{\mu}_i$, estimated by

$$\begin{aligned}\hat{\mu}_i &= t_i \hat{\mu}(\mathbf{x}_i'\mathbf{b}) \\ &= t_i \exp(b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki})\end{aligned}$$

Raw Residual

The raw residual is the difference between the actual response and the estimated value from the model. The formula for the raw residual is

$$r_i = y_i - \hat{\mu}_i$$

Studentized Pearson Residual

The studentized Pearson residual is found be dividing the regular Pearson residual by the square root of one minus the hat diagonal. The formula is

$$(sp)_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i(1-h_{ii})}}$$

Studentized Deviance Residual

The studentized deviance residual is found be dividing the regular deviance residual by the square root of one minus the hat diagonal. The formula is

$$(sd)_i = sign(y_i - \hat{\mu}_i)\sqrt{\frac{2\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}}{1-h_{ii}}}$$

Hat Diagonal

This is the value of the influence measure, h_{ii}. The Hat matrix is used in residual diagnostics to measure the influence of each observation. The hat values, h_{ii}, are the diagonal entries of the Hat matrix which is calculated using

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

where W is a diagonal matrix made up of $\hat{\mu}_i$.

The hat values should be studied to understand which observations have the greatest influence on the fitted regression coefficients. Large hat values are those that are larger than 2*k*/*n*.

Incidence Section when Exposure = 100000

	Average	Prob that	Prob that	Prob that	Prob that	Prob that
	Incidence	Count is	Count is	Count is	Count is	Count is
Row	Rate	5	10	15	20	25
1	2.3505	0.056990	0.000135	0.000000	0.000000	0.000000
2	14.1822	0.003313	0.062866	0.100093	0.030868	0.002778
3	15.9220	0.001037	0.035105	0.099684	0.054827	0.008800
4	22.1186	0.000011	0.001914	0.028111	0.079991	0.066422
5	25.0366	0.000001	0.000357	0.009747	0.051537	0.079521
6	44.6697	0.000000	0.000000	0.000000	0.000016	0.000457
7	5.3340	0.173603	0.024788	0.000297	0.000001	0.000000
8	32.1842	0.000000	0.000003	0.000332	0.006156	0.033343
9	36.1323	0.000000	0.000000	0.000036	0.001201	0.011606
10	50.1944	0.000000	0.000000	0.000000	0.000001	0.000034
11	56.8164	0.000000	0.000000	0.000000	0.000000	0.000001
12	101.3703	0.000000	0.000000	0.000000	0.000000	0.000000

This report gives the predicted incidence rate and Poisson probabilities for various counts.

Row

The row number of the item. If you have excluded some rows by using a filter or if some of the rows had missing values, the row number identifies the original row on the database.

Average Incidence Rate

This is the predicted incidence rate calculated using the formula

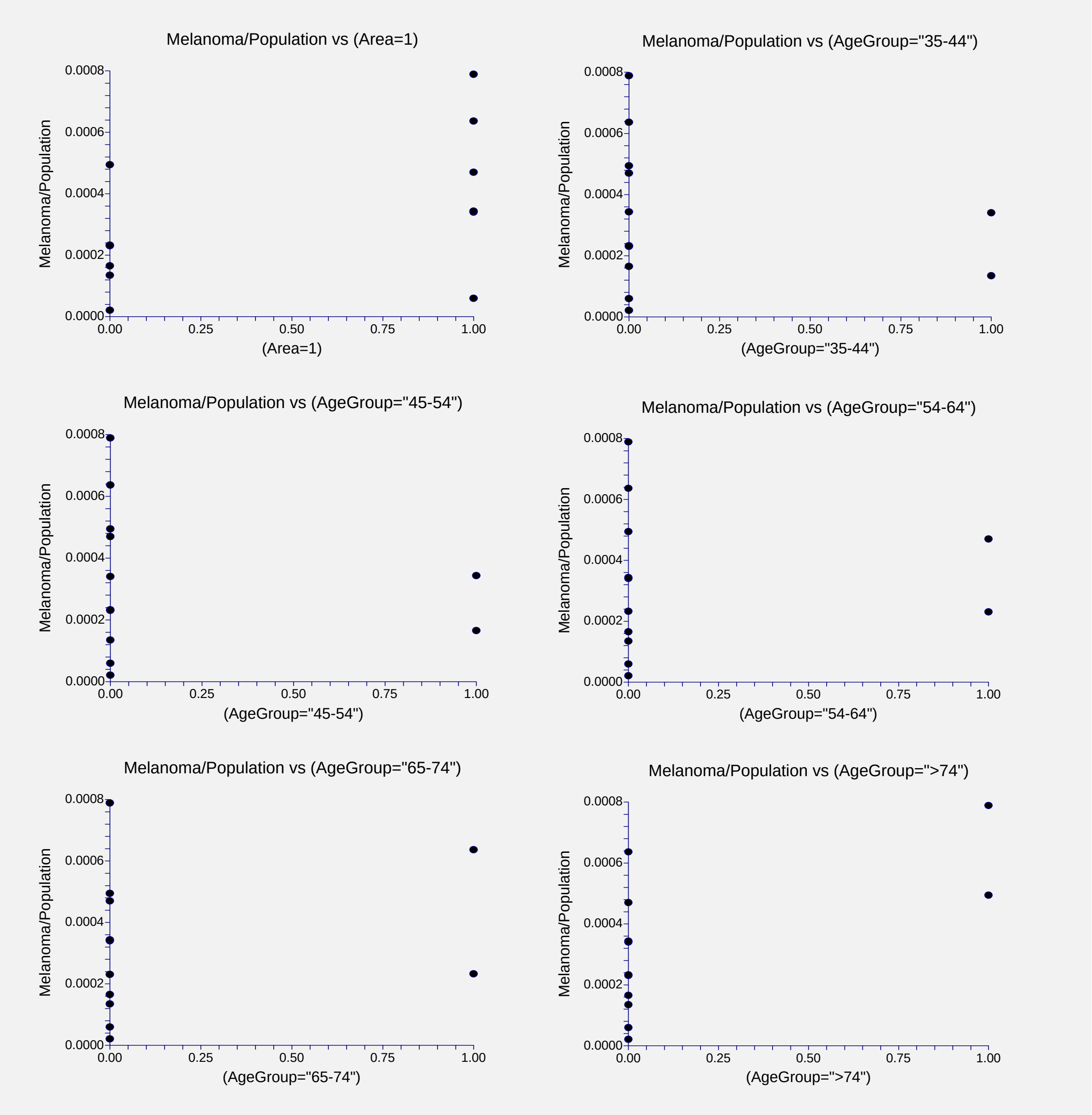
$$\hat{\mu}_i = T\hat{\mu}\left(\mathbf{x}_\mathbf{i}'\mathbf{b}\right)$$

Note that the calculation is made for a specific exposure value, not the value of *T* on the database. This allows you to make valid comparisons of the incidence rates.

Prob that Count is Y

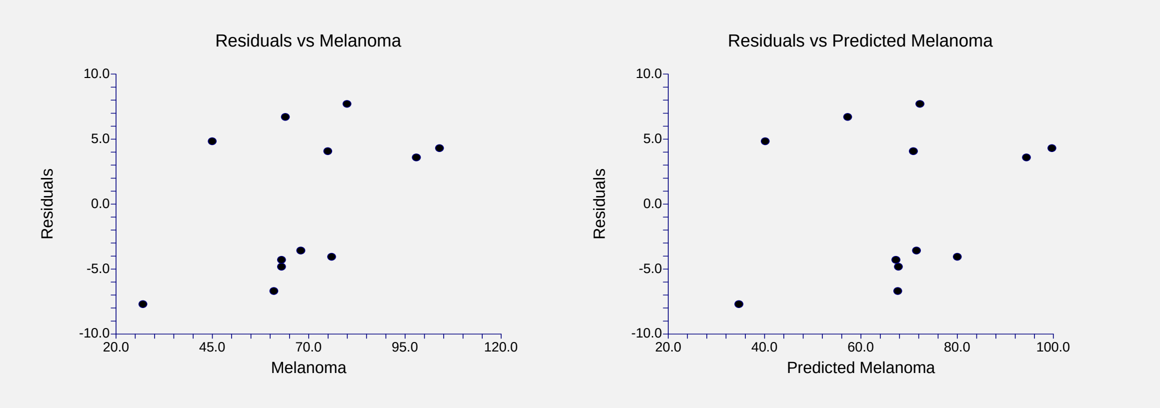
Using the Poisson probability distribution, the probability of obtaining exactly *Y* events during the exposure amount given in the Exposure Value box is calculated for the values of *Y* specified in the Incidence Counts box.

Plots of Y/T (Incidence) vs X



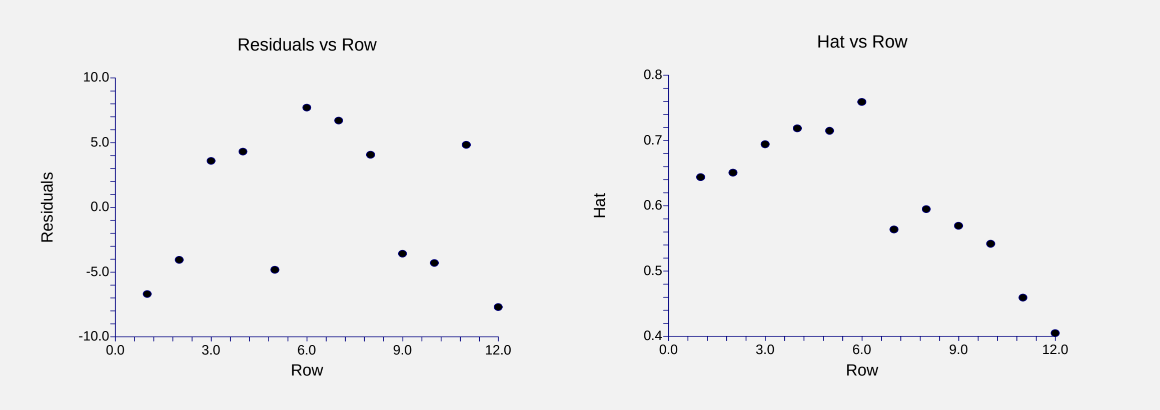
These plots show each of the independent variables plotted against the incidence as measured by Y/T. They should be scanned for outliers and curvilinear patterns.

Plots of Residuals vs. Y and Predicted Y



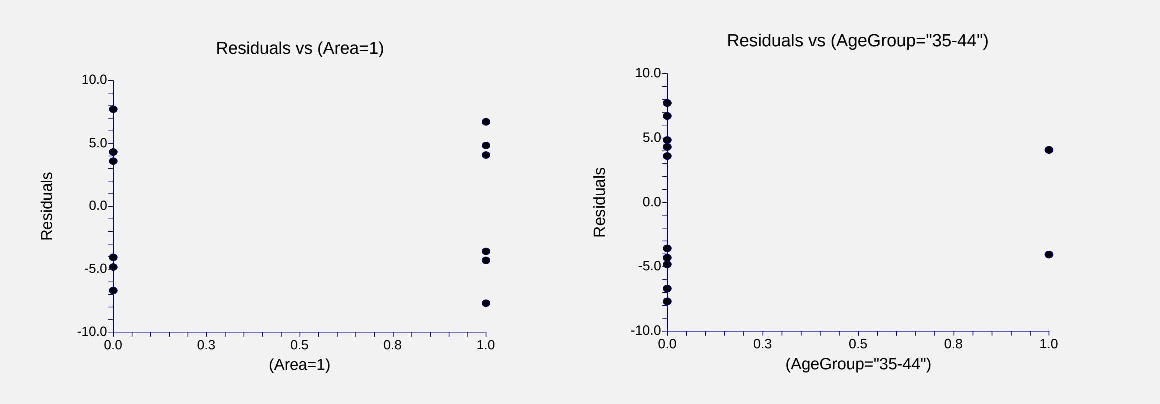
These plots show the residuals versus the dependent variable and the predicted value of the dependent variable. They are used to spot outliers.

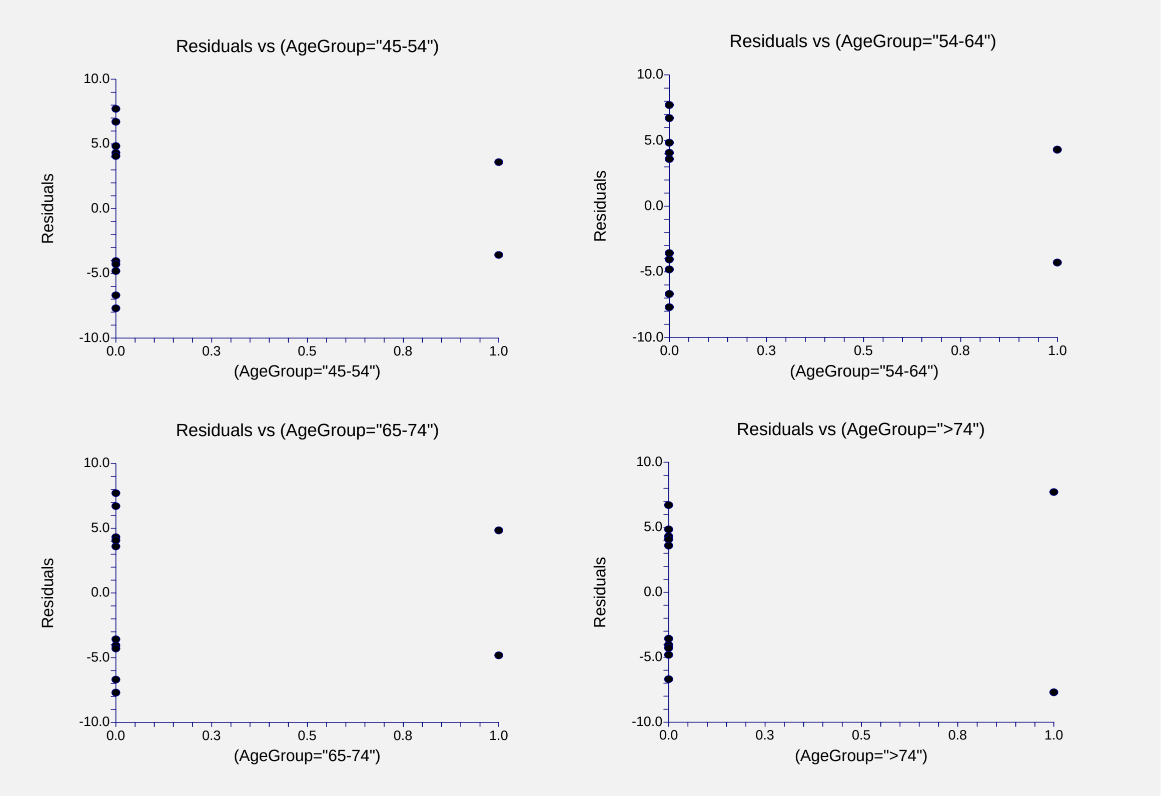
Plots of Residuals and Hats vs. Row



These plots show the residuals and the hat values versus the row numbers. They are used to quickly spot rows that have large residuals or large hat values.

Plots of Residuals and X's





These plots show the residuals plotted against the independent variables. They are used to spot outliers. They are also used to find curvilinear patterns that are not represented in the regression model.

Example 2a – Subset Selection

This example will demonstrate how to select an appropriate subset of the independent variables that are available. The dataset to be analyzed consists of ten independent variables, a dependent variable, a frequency variable, and an exposure variable. The dependent variable was generated using independent variables X1, X2, and X3 using the formula

$$Count = Int\left[TimeExp\left(0.6 + 0.1X1 + 0.2X2 + 0.3X3\right)\right]$$

Variables X4, X5, and X6 were copies of X1 plus a small random component. Similarly, X7 and X8 were near copies of X2 and X9 and X10 were near copies of X3. These near copies of the original variables were added to cause confusion to the selection algorithm. The forty rows of data are stored in the POISREG database.

Now we assume that we do not know how the data were generated. Our task is to find a subset of the ten independent variables that does a good job of fitting the data. We plan to make two runs. The goal of the first run will be to find an appropriate subset size. Then, in the second run, we will identify the variables in this subset and estimate the various regression statistics.

You may follow along here by making the appropriate entries or load the completed template **Example2a** from the Template tab of the Poisson Regression window.

1 **Open the POISREG database.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **POISREG.S0**.
- Click **Open**.

2 **Open the Poisson Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines,** then **Poisson Regression**. The Poisson Regression procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Poisson Regression window, select the **Variables tab**.
- Set the **Y: Dependent Variable** to **Count**.
- Set the **Frequency Variable** to **Cases**.
- Set the **X's: Numeric Independent Variables** to **X1-X10**.
- Set the **T: Exposure Variable** to **Time**.

4 **Specify the model.**

- On the Poisson Regression window, select the **Model tab**.
- Set the **Subset Selection** to **Hierarchical Forward with Switching**.
- Set the **Max Terms in Subset** to **6**.
- Set the **Which Model** to **Up to 1-Way**.
- The rest of this panel can be left at the default values.

5 **Specify the reports.**

- Select the **Reports tab**.
- Uncheck all of the reports and plots except **Run Summary**, **Subset Selection - Summary**, and **Subset Selection - Detail** (these should be checked).

6 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top) or press the F9 function key.

Run Summary Report

```
Parameter              Value              Parameter           Value
Dependent Variable     Count              Subset Method       Forward/Switching
Exposure Variable      Time               Ind. Var's Available 10
Frequency Variable     Cases              No. of X's in Model 5
Rows Used              40                 Iterations          20
Sum of Frequencies     130                Final Likelihood    -288.8153
Maximum Convergence    3.420807E-06       Convergence Zero    1E-09
Dispersion Phi         0.0138             Phi was not used to correct standard errors.
```

This report provides several details about the data and the MLE algorithm as it fit the best model found during the search. We note that, as expected, there were 40 rows used. The fact that 20 iterations were needed to solve the likelihood equations is a source of concern because this shows that the algorithm may not have converged. This may have been due to our fitting of a model that had too many terms.

Subset Selection Summary Section

```
Number of          Log
Terms       Likelihood      R-Squared       Deviance           AIC
1            -730.6939         0.0000       885.5007      887.5007
2            -434.0619         0.6700       292.2366      296.2366
3            -348.4077         0.8634       120.9282      126.9282
4            -288.8552         0.9979         1.8233        9.8233
5            -288.8343         0.9980         1.7815       11.7815
6            -288.8153         0.9980         1.7434       13.7434
```

This report will help us determine an appropriate subset size. By scanning each column, we can see that three variables are needed. All of these measures are functions of each other. However, they each offer insight into the appropriate subset size.

In this example, the four measures unanimously point to three as the appropriate subset size.

Number of Variables

This is the number of terms in the model including the intercept. Each line presents the results for the best model found for that subset size. The first line presents the results for the intercept-only model.

Log Likelihood

This is the value of the log likelihood function. Since the goal of maximum likelihood is to maximize this value, we want to select a subset size after which the log likelihood is not increased significantly.

In this example, after three terms are added (in addition to the intercept) the log likelihood does not change a great deal. The log likelihood points to a subset size of three terms plus the intercept for a total of four.

R-Squared

This is the value of pseudo *R*-squared—a measure of the adequacy of the model. Since our goal is to maximize this value, we want to select a subset size after which the this value is not increased significantly.

In this example, after four terms are included, the *R*-squared is 0.9979 and it does not change a great deal. The *R*-squared values point to a subset size of four.

Deviance

Deviance is a measure of the lack of fit. Hence, we want to select a subset size after which the deviance is not significantly decreased.

In this example, after four terms are included, the Deviance is 1.8233 and it does not change a great deal. The Deviance values point to a subset size of four.

AIC

These are the Akaike information criterion values for each subset size. This criterion measures both the lack of fit and the size of the regression model. Our goal is to minimize this value.

In this example, the subset size of four gives the lowest value AIC and is thus the subset size implied by this statistic.

Subset Selection Detail Section

		No. of	No. of	Log		Term	Term
Step	Action	Terms	X's	Likelihood	R-Squared	Entered	Removed
1	Add	1	1	-730.6939	0.0000	Intercept	
2	Add	2	2	-434.0619	0.6700	X3	
3	Add	3	3	-348.4423	0.8634	X2	
4	Switch	3	3	-348.4077	0.8634	X9	X3
5	Add	4	4	-289.2634	0.9970	X6	
6	Switch	4	4	-289.0943	0.9974	X8	X2
7	Switch	4	4	-288.8552	0.9979	X3	X9
8	Add	5	5	-288.8343	0.9980	X5	
9	Add	6	6	-288.8201	0.9980	X7	
10	Switch	6	6	-288.8153	0.9980	X2	X5

This report shows the progress of the subset selection algorithm through its various steps. It shows the original term added at each step and any switching that was done.

Step

This is the number of the step in the subset selection process.

Action

Two actions are possible at each step: Add or Switch. *Add* means that the subset size was increased and the term entered as added to the set of active regressor variables. *Switch* means that the subset size remained the same while one active regressor was removed and another was activated.

No. of Terms

This is the number of active terms (including the intercept) at the end of this step.

No. of X's

This is the number of active variables (excluding the intercept) at the end of this step. This reminds you of how many *X* variables were generated for each term involving a categorical variable.

Log Likelihood

This is the value of the log likelihood after this step was completed.

R-Squared

This is the pseudo *R*-squared value after this step was completed.

Variable Entered

This is the name of the regressor that was added to the list of active regressor variables.

Variable Removed

In switching steps, this is the name of the variable that was removed from the list of active regressor variables.

Example 2b – Subset Selection Continued

Example 2a completed the first step in the subset selection process by indicating that a subset of four terms is appropriate. Now, a second run must be made to find those terms.

The instructions provide here assume that you have just completed Example 2a. If you have not, you must complete it first since we will only tell you want needs to be changed.

You may follow along here by making the appropriate entries or load the completed template **Example2b** from the Template tab of the Poisson Regression window.

1 **Specify the model.**

- On the Poisson Regression window, select the **Model tab**.
- Set the **Max Terms in Subset** to **4**.
- The rest of this panel can be left at the default values.

2 **Specify the reports.**

- Select the **Reports tab**.
- Uncheck all of the reports and plots except **Run Summary**, **Subset Selection - Summary**, **Subset Selection – Detail**, **Regression Coefficients**, and **Residuals** (these should be checked).

3 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top) or press the F9 function key.

Run Summary Report

```
Parameter                 Value              Parameter             Value
Dependent Variable        Count              Subset Method         Forward/Switching
Exposure Variable         Time               Ind. Var's Available  10
Frequency Variable        Cases              No. of X's in Model   3
Rows Used                 40                 Iterations            7
Sum of Frequencies        130                Final Likelihood      -288.8552
Maximum Convergence       6.193126E-10       Convergence Zero      1E-09
Dispersion Phi            0.0142             Phi was not used to correct standard errors.
```

We note that the final model converged in only five iterations and the Maximum Convergence is less than Convergence Zero. This means that the algorithm terminated normally.

Subset Selection Summary Section

```
Number of         Log
Terms      Likelihood     R-Squared      Deviance           AIC
1           -730.6939        0.0000      885.5007      887.5007
2           -434.0619        0.6700      292.2366      296.2366
3           -348.4077        0.8634      120.9282      126.9282
4           -288.8552        0.9979        1.8233        9.8233
```

This report again shows us that a subset size of four is a reasonable choice.

Subset Selection Detail Section

```
                 No. of  No. of        Log                Term        Term
Step   Action    Terms    X's   Likelihood  R-Squared    Entered     Removed
1         Add        1      1    -730.6939     0.0000    Intercept
2         Add        2      2    -434.0619     0.6700    X3
3         Add        3      3    -348.4423     0.8634    X2
4      Switch        3      3    -348.4077     0.8634    X9          X3
5         Add        4      4    -289.2634     0.9970    X6
6      Switch        4      4    -289.0943     0.9974    X8          X2
7      Switch        4      4    -288.8552     0.9979    X3          X9
```

This report shows the algorithm's journey through the maze of possible models. During the process, three variables were switched in order to achieve a better model.

Regression Coefficients Section

```
                Regression                 Wald's               Lower 95.0%   Upper 95.0%
Independent    Coefficient   Standard       Chi^2        Prob    Confidence    Confidence
Variable               (B)      Error   (Ho:B=0)        Level         Limit         Limit
Intercept         -0.12374    0.10638       1.35       0.2448      -0.33224       0.08476
X3                 0.01047    0.00041     656.32       0.0000       0.00967       0.01127
X8                 0.00677    0.00043     245.70       0.0000       0.00592       0.00761
X6                 0.00345    0.00031     121.68       0.0000       0.00283       0.00406
Dispersion Phi                 0.0142
```

This report provides the details of the model that was selected. We note the X3, X8, and X6 were included in the model. We assume that X8 is taking the place of X2 and X6 is taking the place of X1. In fact, we ran a Poisson regression with X1, X2, and X3 in the model. The log likelihood for this model was -288.9466, which is slightly less than the -288.8552 achieved by our best model. This concludes our discussion of this example. Usually, we would go on to study the residual plots and complete the analysis by making a third run with only the variables X3, X6, and X8 specified.

Chapter 330

Response Surface Regression

Introduction

This *Response Surface Analysis* (RSA) program fits a polynomial regression model with cross-product terms of variables that may be raised up to the third power. It calculates the minimum or maximum of the surface. The program also has a variable selection feature that helps you find the most parsimonious hierarchical model. **NCSS** automatically scans the data for duplicates so that a lack-of-fit test may be calculated using pure error.

One of the main goals of RSA is to find a polynomial approximation of the true nonlinear model, similar to the Taylor's series expansion used in calculus. Hence, you are searching for an approximation that works well in a specified region. As the region is reduced, the number of terms may also be reduced. In a very small region, a linear (first-order) approximation may be adequate. A larger region may require a quadratic (second-order) approximation.

Hierarchical Models

In the following discussion, the X's are independent variables with at least three distinct values (up to six X's may be specified). Y is the dependent variable. Z is a covariate (note that covariates do not have to have three or more levels). The β's are the regression coefficients or beta weights.

A polynomial model is one in which the X's occur as multiples of each other. Examples of polynomial models are:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \varepsilon_j$$

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{1j} X_{2j} + \varepsilon_j$$

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j}^2 + \beta_3 X_{1j}^2 X_{2j}^3 + \varepsilon_j$$

A hierarchical model obeys the following rule: all lower-order terms that can be constructed by reducing the exponents of the variables in a term are also in the model. For example, if the term X1X$_2^2$ is in the model, so are X1, X2, X1X2, and X$_2^2$.Notice that each of these terms can be created be decreasing the exponents of the two variables that form the original term (noting that X0j = 1). Note that the first two models are hierarchical, but the third is not.

Hierarchical models enjoy several useful properties, including stability, the ability to change the scale (coding) of a variable, and a general relationship with ANOVA modeling. However, they usually require the fitting of more parameters than nonhierarchical models. This **NCSS** procedure fits only hierarchical models. If nonhierarchical models are desired, they can be fit using the Multiple Regression module.

Model Selection

There are several strategies to variable selection and model building in regression analysis: forward selection, backward elimination, stepwise, all possible regressions, and more. However, none of these methods guarantee hierarchical models. We need a method that does. This **NCSS** program adopts a strategy that has been used for quite a while in dealing with hierarchical models. The strategy may be outlined as follows:

1. Begin with the most complicated model desired. **NCSS** allows terms of the form $X_1^iX_2^j$, where *i* and *j* are each less than or equal to three.
2. Search through all terms, marking those that are not necessary to maintain the hierarchical constraint on the model. This group of terms is available for removal.
3. Check each of the available terms to determine how much R-Squared is decreased if they are removed.
4. Remove the term that decreases R-Squared the least. Return to step 2. Note that this variable is never reconsidered for inclusion in the model.
5. If no available term can be identified that reduces R-Squared by an amount that is less than the specified cutoff value, the model selection procedure is terminated.

Assumptions and Limitations

The same assumptions and qualifications apply here as applied to multiple regression. We refer you to the Assumptions section in the Multiple Regression chapter for a discussion of these assumptions. We will here mention a couple of restrictions necessary for this algorithm to work.

Number of Observations

The number of observations must be at least one greater than the number of terms (including all cross products). A popular rule-of-thumb when using any variable selection procedure is that you have at least five observations for each term.

Unique Data Values

Since various powers of the variables are included, the structure of your data must allow for these powers to be fit. This means that if the maximum exponent on a variable is k, the number of unique values in that variable must be at least k+1. For example, suppose a variable consisted of two values: -1 and 1. You could not fit a model that included more than a linear (k=1) term in this variable. Again, suppose your data consisted of three values: -1,0,1. The maximum exponent that could be used with this variable is 2.

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown in the following table and is found in the ODOR database. This database relates a measurement of odor to three variables in a chemical process. Fifteen rows of data were obtained. The values of the three independent variables have been recoded so that they are -1, 0, and 1. A sixteenth row has been added. Notice that it does not contain a value in the *Odor* column. A predicted value will be generated for this row, but its values will not be used in the estimation process.

We suggest that you open this database now so that you can follow along with the example.

ODOR dataset

Odor	Temp	Ratio	Height
66	-1	-1	0
58	-1	0	-1
65	0	-1	-1
-31	0	0	0
39	1	-1	0
17	1	0	-1
7	0	1	-1
-35	0	0	0
43	-1	1	0
-5	-1	0	1
43	0	-1	1
-26	0	0	0
49	1	1	0
-40	1	0	1
-22	0	1	1
	1	1	0

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present for all but the dependent variable, a predicted value is generated for this row.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Dependent Variable

Specifies the dependent (Y) variable.

Minimum

This option lets you set a minimum for the depth (dependent variable) axis. If left blank, it is determined from the data.

Maximum

This option lets you set a maximum for the depth (dependent variable) axis. If left blank, it is determined from the data.

Decimals

This specifies the number of decimal places displayed in the reference numbers on the dependent (depth) axis.

Factor Variables

Factor Variable (A - F)

Specifies the variables to be used as independent variables. Each of these variables must be categorical—have only a few unique values. Think of *A* as X_1, *B* as X_2, etc. The terms of the hierarchical model will be generated from these variables. Note that each variable must have enough unique values to fit the highest exponent required of it in the model:

Number of Unique Values	Largest Exponent Possible
2	1 (linear)
3	2 (quadratic)
4	3 (cubic)

Constant

This option lets you specify a constant value to be used for this variable when it is not one of the pair of factors being displayed on the grid plot. If you leave this option blank, the factor average will be used.

Minimum

This option lets you set a minimum for the axis related to this factor. If left blank, it is determined from the data.

Maximum

This option lets you set a maximum for the axis related to this factor. If left blank, it is determined from the data.

Decimals

This specifies the number of decimal places displayed in the reference numbers on axis showing this factor.

Covariates

Covariate Variables

These are other independent variables included in the regression model, but their powers and cross-products will not be generated. They are not part of the "response surface."

Model Tab

These boxes define the hierarchical model in a shorthand notation.

Model Specification – Order

Order (A - F)

These boxes define the maximum exponent for each factor. Values from one to three are allowed. All terms of order less than or equal to this value are included in the model. For example, a '2' implies that X and X^2 are included.

Model Specification – Maximum Orders of Two-Way Terms

Maximum Orders of Two-Way Terms (AB - EF)

These boxes define the maximum exponent for each factor in the cross-product term of the corresponding variables. For example, "AC" represents the product of factors A and C.

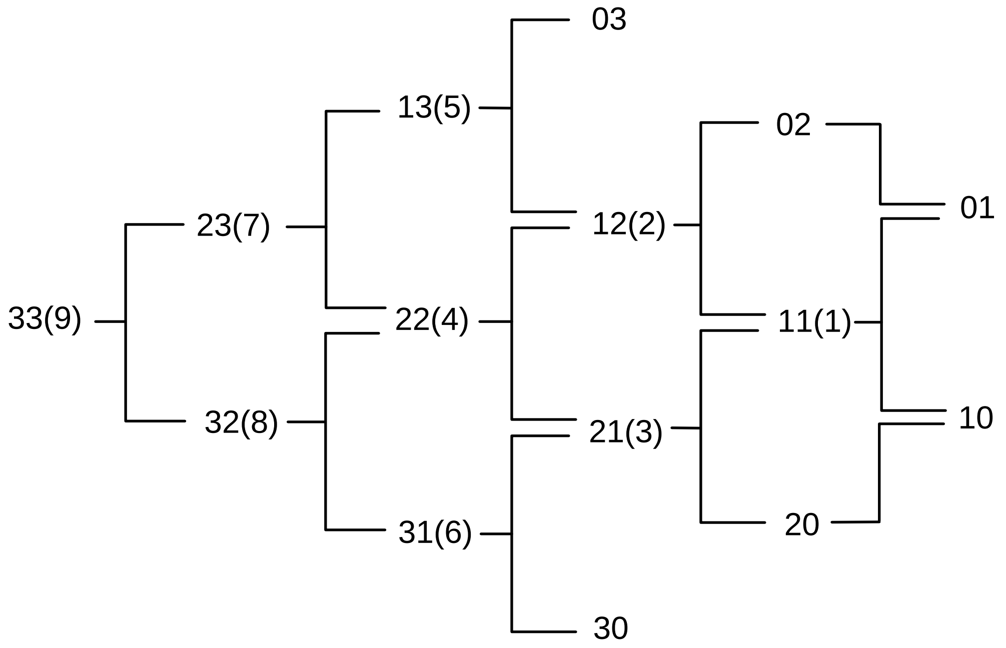
All subset terms (children) are also included in the model, so that the hierarchical nature of the model is maintained.

A code is used to specify the maximum exponents of each term. Up to three of these may be needed to specify the desired hierarchical model. The following table relates each coded value to the terms that it generates. This table will be generated for the *AB* term. The pattern extends in an obvious manner to the other cross-products. Note that a "10" is used to represent A, a "01" represents B, a "20" represents A^2, and so on.

Code	Term(s)	Cross Product Terms Actually Included
1	11	AB
2	12	AB, AB2
3	21	AB, A2B
23	12,21	AB, AB2, A2B
4	22	AB, A2B, AB2, A2B2
5	13	AB, AB2, AB3
35	21,13	AB, A2B, AB2, AB3

Code	Term(s)	Cross Product Terms Actually Included
45	22,13	AB, A2B, AB2, A2B2, AB3
6	31	AB, A2B, A3B
26	12,31	AB, AB2, A2B, A3B
46	22,31	AB, AB2, A2B, A2B2, A3B
56	13,31	AB, AB2, A2B, AB3, A3B
456	22,13,31	AB, AB2, A2B, AB3, A3B, A2B2
7	23	AB, A2B, AB2, A2B2, AB3, A2B3
67	31,23	AB, A2B, AB2, A2B2, AB3, A3B, A2B3
8	32	AB, AB2, A2B, A2B2, A3B, A3B2
58	13,32	AB, AB2, A2B, A2B2, AB3, A3B, A3B2
78	23,32	AB, AB2, A2B, A2B2, AB3, A3B, A2B3, A3B2
9	33	AB, AB2, A2B, A2B2, A3B, A3B2, AB3, A2B3, A3A3

The following tree diagram shows the hierarchical structure of this system. Each term generates all terms to the right of it. These terms are called *children.* A term that is a child of one term is not specified with that term. For example, terms 13, 22, and 31 could be selected together. However, the terms 23 and 21 could not be entered together since 21 is a child of 23 and will automatically be included when 23 is specified. Note the cross-product terms include their codes in parentheses. Also note that terms like 03 and 20 are specified in the One-Way Terms section.



The actual specification of the term is accomplished by selecting one or more codes from a list of possible models. For example, you might select “58.” This model represents the 13 and the 32 terms plus all their children.

The usual quadratic model is specified by selecting 2’s for the Order terms and 1’s for the Two-Way terms.

Model Selection

Conduct Model Selection

Specifies whether to search (checked) for the most parsimonious hierarchical model or simply fit the one that is specified (unchecked).

Minimum R-Squared Kept

Sets the minimum amount that an available term must add to the overall R-Squared to avoid being removed from the model. Note that this is the amount that the R-Squared is decreased if only this term is removed, leaving all other terms in the model.

Optimization Tab

These options control the Hooke and Jeeves optimization routine as described in Nash (1987).

Optimization Options

Optimization Goal

Specifies whether a minimum or maximum is sought.

Maximum Evaluations

Specifies the maximum number of function evaluations before the routine is aborted.

Set Linear Factors to Mean Values

Specifies whether linear-only variables should be set to their mean values or left to vary. Usually, you would set these to their mean values since a straight line has no minimum or maximum.

Set Cubic Factors to Mean Values

Specifies whether cubic variables should be set to their mean values or left to vary. Usually, you would set these to their mean values since a cubic function has no minimum or maximum.

Reports Tab

The following options control which reports are displayed.

Select Reports

Descriptive Statistics ... Residuals Reports

Specifies whether to output the various reports.

Select Plots

Probability Plot and Grid Plots

Specifies whether to output these plots.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimal Places

Response and Beta Decimals

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Grid Plots Tab

A grid plot of the response surface may be generated for each pair of factor variables. The following options control these plots.

Vertical Axis

Label

The text that will appear on the vertical axis of the plot.

Number of Slices

The number of divisions (blocks or symbols) along the vertical axis. This controls the coarseness of the grid.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on the vertical axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Horizontal Axis

Label

The text that will appear on the horizontal axis of the plot.

Number of Slices

The number of divisions (blocks or symbols) along the horizontal axis. This controls the coarseness of the grid.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on the horizontal axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Z Axis

Label

The text that will appear in the legend.

Number of Slices

The number of divisions along the dependent variable axis.

Plot Settings

Plot Style File

The style of the grid plot is set here by selecting a grid plot style file. If you want to change options on the grid plot that are not given below, you should change them in the Default Grid Plot procedure and load that style here.

Plot Style

This option lets you specify what plotting symbols to use in the plot. You can select blocks, regular symbols, or multicolored symbols. The actual symbols are specified in the Default Grid Plot template.

Show Legend

Indicate whether a legend of the dependent variable divisions should be generated.

Titles

Plot Title

The text that will be the title of the plot. Abbreviations of {Z}, {X}, and {Y} for the variables may also be used in the title to represent variable names.

Storage Tab

Various statistics calculated for each row may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is running.

The variables you specify must already have been defined on the current database. Remember that existing data will be replaced. Following is a description of the statistics that can be stored.

Data Storage Variables

Predicted Values

The predicted (Yhat) values.

Residuals

The residuals (Y-Yhat).

Expanded Factors

If you are going to analyze your data further using another regression module, you will need to generate the squares, cubes, and cross-product terms using Variable Transformations. This option lets you store these variables directly on the database. New variables containing all $X_i^I X_j^J$ for all values of i, I, j, and J in the current model will be created.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Response Surface Analysis

This section presents an example of how to run a response surface analysis of the data contained in the ODOR database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Response Surface Regression window.

1 Open the ODOR dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **ODOR.s0**.
- Click **Open**.

2 Open the Response Surface Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines**, then **Response Surface Regression**. The Response Surface Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Response Surface Regression window, select the **Variables tab**.
- Double-click in the **Dependent Variable** text box. This will bring up the variable selection window.
- Select **Odor** from the list of variables and then click **Ok**. "Odor" will appear in the Dependent Variable box.
- Double-click in the **Factor Variable - A** text box. This will bring up the variable selection window.
- Select **Temp** from the list of variables and then click **Ok**.
- Double-click in the **Factor Variable - B** text box. This will bring up the variable selection window.
- Select **Ratio** from the list of variables and then click **Ok**.
- Double-click in the **Factor Variable - C** text box. This will bring up the variable selection window.
- Select **Height** from the list of variables and then click **Ok**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Descriptive Statistics Section

```
Descriptive Statistics Section

Variable        Count    Mean       Minimum   Maximum
Temp            15       0          -1        1
Ratio           15       0          -1        1
Height          15       0          -1        1
Odor            15       15.2       -40       66
```

This report provides the count, mean, minimum, and maximum of each of the variables in the analysis. It allows you to determine if the data fall within reasonable limits.

Hierarchical Model Summary Section

```
Hierarchical Model Summary Section

Number of Terms Removed              0
Number of Terms Remaining            9
R-Squared Cutoff Value               0.010000
R-Squared of Final Model             0.881989

Coded Hierarchical Model
           A             B             C
A Temp     2             1(11)         1(11)
B Ratio                  2             1(11)
C Height                               2

Notes:
  For off-diagonal entries:
    1=u1w1, 2=u1w2, 3=u2w1, 4=u2w2, 5=u1w3, 6=u3w1, 7=u2w3, 8=u3w2, 9=u3w3.
  For diagonal entries:
    1=u1, 2=u2, 3=u3.
  Where u1=u, u2=u^2=u*u, and u3=u^3=u*u*u.
```

This report shows the hierarchical model that was specified. It also shows the final R-Squared value as well as the R-Squared cutoff that was used. It is mainly used to document the model used.

The specified model is determined by considering all nonzero entries. The *Notes* section at the bottom shows how the model is determined. For example, the first line is nonzero for the *terms AA*, *AB*, and *AC*. These codes represent the terms: Temp^2, (Temp)(Ratio), and (Temp)(Height). All child terms necessary to make this a hierarchical model are also generated.

Sequential ANOVA Section

```
Sequential ANOVA Section
                      Sequential        Mean                       Prob        Incremental
Source         DF     Sum-Squares       Square          F-Ratio    Level       R-Squared
Regression     9      18881.98          2097.998        4.15       0.065691    0.881989
 Linear        3      7143.25           2381.083        4.71       0.064071    0.333666
 Quadratic     3      11445.23          3815.078        7.55       0.026426    0.534614
 Lin x Lin     3      293.5             97.83334        0.19       0.896470    0.013710
Total Error    5      2526.417          505.2833                               0.118011
 Lack of Fit   3      2485.75           828.5833        40.75      0.024047    0.116111
 Pure Error    2      40.66667          20.33333                               0.001900

Sequential ANOVA Section Using Pure Error
                      Sequential        Mean                       Prob        Incremental
Source         DF     Sum-Squares       Square          F-Ratio    Level       R-Squared
Regression     9      18881.98          2097.998        103.18     0.009635    0.881989
 Linear        3      7143.25           2381.083        117.10     0.008479    0.333666
 Quadratic     3      11445.23          3815.078        187.63     0.005306    0.534614
 Lin x Lin     3      293.5             97.83334        4.81       0.176872    0.013710
Total Error    5      2526.417          505.2833                               0.118011
 Lack of Fit   3      2485.75           828.5833        40.75      0.024047    0.116111
 Pure Error    2      40.66667          20.33333                               0.001900
```

This display actually shows two reports. The top is the regular Sequential ANOVA Section defined below. Note that the denominator of the F-Ratios is the Total Error Mean Square. The bottom report is identical to the top, except that the denominator of the F-Ratios is now the Pure Error Mean Square.

This report is designed with two main goals:

1. Determine the sequential influence of the various power and cross-product terms.
2. Test for model lack of fit if repeated observations are available.

Source

The group of independent variables being tested.

Regression	Total of all terms in the model.
Linear	The total for X_i terms.
Quadratic	The total for X_i^2 terms.
Cubic	The total for X_i^3 terms.
Lin x Lin	The total for X_iX_j terms.
Lin x Quad	The total for $X_iX_j^2$ terms.
Quad x Quad	The total for $X_i^2X_j^2$ terms.
Lin x Cubic	The total for $X_iX_j^3$ terms.
Quad x Cubic	The total for $X_i^2X_j^3$ terms.
Cubic x Cubic	The total for $X_i^3X_j^3$ terms.

DF

The degrees of freedom associated with the group of terms.

Sequential Sum-Squares

The regression sum of squares added sequentially by each group of terms. Each group of terms adds this amount of sum of squares after accounting for the terms above it in the report.

Mean Square

The sum of squares divided by the degrees of freedom.

F-Ratio

The F-value formed by dividing the Mean Square by the Total Error Mean Square. Note that these tests are sequential in nature and should be considered from the bottom up. Note that in the second report, the Total Error Mean Square is replaced by the Pure Error Mean Square as the denominator of the F-ratio.

In the above example, the Lin x Lin F-ratio tests whether the linear-by-linear terms are significant in the regression model after considering the linear and quadratic terms. The Quadratic F-ratio tests whether the quadratic terms add significantly to a model consisting of the linear terms (ignoring the linear-by-linear terms).

In terms of the ODOR data, the tests are interpreted as follows:

Group	Terms	Hypothesis Tested
Lin x Lin	Temp x Ratio Temp x Height Ratio x Height	All coefficients of these variables are zero.
Quadratic	Temp x Temp Ratio x Ratio Height x Height	All coefficients of these variables are zero, ignoring the influence of the cross-product terms.
Linear	Temp Ratio Height	All coefficients of these variables are zero, ignoring the influence of the cross-product and quadratic terms

Prob Level

This is the right-tail probability or significance level of this test. Reject the hypothesis that the influence of the terms is zero when this value is less than a predetermined value of alpha, say 0.05.

Incremental R-Squared

The first line displays the total R-Squared for the complete model. The other lines display the amount of R-Squared that is added by each group of terms. Hence, the total of the rest of the lines equals the first.

Lack of Fit and Pure Error

These lines are only displayed if you have repeated observations from which the variability between identical observations may be estimated. The lack of fit tests the adequacy of the specified model. A significant F-test implies that a higher-order polynomial (such as cubic) or a different functional form would fit the data better.

If pure error is available, the F-tests are recalculated using the Pure Error Mean Square as the denominator rather than the Total Error Mean Square.

ANOVA Section

```
ANOVA Section
                    Last             Mean                        Prob         Term
Factor       DF     Sum-Squares      Square         F-Ratio      Level        R-Squared
Temp         4      5258.016         1314.504       2.60         0.161334     0.245605
Ratio        4      11044.6          2761.151       5.46         0.045377     0.515900
Height       4      3813.016         953.254        1.89         0.251025     0.178108
Total Error  5      2526.417         505.2833                                 0.118011
 Lack of Fit 3      2485.75          828.5833       40.75        0.024047     0.116111
 Pure Error  2      40.66667         20.33333                                 0.001900

ANOVA Section Using Pure Error
                    Last             Mean                        Prob         Term
Factor       DF     Sum-Squares      Square         F-Ratio      Level        R-Squared
Temp         4      5258.016         1314.504       64.65        0.015291     0.245605
Ratio        4      11044.6          2761.151       135.79       0.007324     0.515900
Height       4      3813.016         953.254        46.88        0.020994     0.178108
Total Error  5      2526.417         505.2833                                 0.118011
 Lack of Fit 3      2485.75          828.5833       40.75        0.024047     0.116111
 Pure Error  2      40.66667         20.33333                                 0.001900
```

This report tests the significance of each factor. This display actually shows two reports. The top is the regular ANOVA Section defined below. Note that the denominator of the F-Ratios is the Total Error Mean Square. The second report is identical to the top, except that the denominator of the F-Ratios is now the Pure Error Mean Square.

Factor

This line lists the factor being tested for deletion. All terms that include this factor are included in the test. In our example, the terms being tested are as follows:

Factor	Individual Terms Referred To
Temp	Temp, Temp x Ratio, Temp x Height, Temp x Temp.
Ratio	Ratio, Temp x Ratio, Ratio x Height, Ratio x Ratio.
Height	Height, Height x Ratio, Height x Temp, Height x Height.

Note that there is overlap in these terms (some cross-products occur twice).

DF

The degrees of freedom associated with the term(s).

Last Sum-Squares

The regression sum of squares that would be lost if this factor were omitted.

Mean Square

The sum of squares divided by the degrees of freedom.

F-Ratio

In the top report, the F-value is formed by dividing the Mean Square by the Total Error Mean Square. In the second report, the F-value is formed by dividing the Mean Square by the Pure Error Mean Square. Note that these tests are not sequential, but each tests the importance of the factor after considering all other factors.

Prob Level

This is the right-tail probability or significance level of this test. Reject the hypothesis that the influence of the terms is zero when this value is less than a predetermined value of alpha, say 0.05.

Term R-Squared

The amount that the R-Squared would decrease if this factor were removed from the model.

Lack of Fit / Pure Error

These lines are only displayed if you have repeated observations from which the variability between like observations may be estimated. The lack of fit tests the adequacy of the specified model. If this test is significant, conclude that a higher order polynomial (such as cubic), or a different functional form, would fit the data better.

Estimation Section

```
Estimation Section
                                 Regression   Standard                         Prob          Last
Parameter              DF        Coefficient  Error           T-Ratio          Level         R-Squared
Intercept              1         -30.66667
Temp                   1         -12.125      7.947353          -1.53          0.187613      0.054938
Ratio                  1         -17          7.947353          -2.14          0.085417      0.107995
Height                 1         -21.375      7.947353          -2.69          0.043321      0.170733
Temp^2                 1         32.08333     11.69819           2.74          0.040667      0.177530
Ratio^2                1         47.83333     11.69819           4.09          0.009457      0.394616
Height^2               1         6.083333     11.69819           0.52          0.625242      0.006383
Temp*Ratio             1         8.25         11.23925           0.73          0.495884      0.012717
Temp*Height            1         1.5          11.23925           0.13          0.899034      0.000420
Ratio*Height           1         -1.75        11.23925          -0.16          0.882357      0.000572

Estimation Section Using Pure Error
                                 Regression   Standard                         Prob          Last
Parameter              DF        Coefficient  Error           T-Ratio          Level         R-Squared
Intercept              1         -30.66667
Temp                   1         -12.125      1.594261          -7.61          0.016853      0.054938
Ratio                  1         -17          1.594261         -10.66          0.008680      0.107995
Height                 1         -21.375      1.594261         -13.41          0.005517      0.170733
Temp^2                 1         32.08333     2.346688          13.67          0.005307      0.177530
Ratio^2                1         47.83333     2.346688          20.38          0.002398      0.394616
Height^2               1         6.083333     2.346688           2.59          0.122137      0.006383
Temp*Ratio             1         8.25         2.254625           3.66          0.067241      0.012717
Temp*Height            1         1.5          2.254625           0.67          0.574315      0.000420
Ratio*Height           1         -1.75        2.254625          -0.78          0.518860      0.000572
```

This report shows the regression coefficient estimates of each term and their test of significance. This display actually shows two reports. The top is the regular Estimation Section defined below. Note that the Standard Errors are based on the Total Error Mean Square. The second report is identical to the top, except that the Standard Errors are now based on the Pure Error Mean Square.

Parameter

The particular term being displayed.

DF

The degrees of freedom associated with the term.

Regression Coefficient

The estimated value of the regression coefficient.

Standard Error

The standard error of the above regression coefficient. Note that the Total Error Mean Square is used for the top report, and the Pure Error Mean Square is used for the bottom report.

T-Ratio

The t-value for testing that this regression coefficient is zero after considering all other terms in the model. Note that the Total Error Mean Square is used for the top report, and the Pure Error Mean Square is used for the bottom report.

Prob Level

The probability or significance level of this test. If you were testing at the alpha equals 0.05 level of significance, this value would have to be less than 0.05 in order for the test to be deemed significant and the regression coefficient different from zero.

Last R-Squared

The amount that the R-Squared would decrease if this term were removed from the model.

Optimum Solution Section

```
Optimum Solution Section

                   Maximum              Optimum
Parameter          Exponent             Value
Temp               2                    0.1219125
Ratio              2                    0.1995746
Height             2                    1.770525

Function at optimum                     -52.02463
Number of Function Evaluations          359
Maximum Functions Evaluations           500
```

This report gives the results of the function minimization (or maximization) calculation.

Optimum Value

The value for each of the factors at the computed critical point. Covariates were evaluated at their means. Note that this solution is not constrained to fall within the design space. Note also that the values of some variables may be very large or small. This indicates that the function did not have a minimum (maximum) and that the search procedure was terminated by the maximum number of function evaluations. In this case, you might switch from finding a minimum to finding a maximum in the Optimization Goal box.

Function at Optimum

The value of the estimated function evaluated at the optimal values of each of the factors.

Residual Section

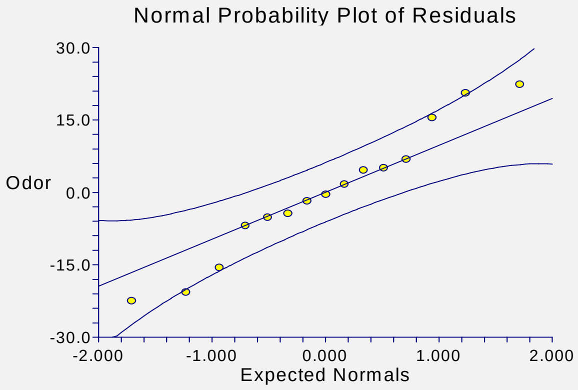
```
Residual Section

Row       Odor     Predicted      Residual
1           66        86.625       -20.625
2           58          42.5          15.5
3           65        59.875         5.125
4          -31     -30.66667     -.3333333
5           39        45.875        -6.875
6           17         15.25          1.75
7            7        29.375       -22.375
8          -35     -30.66667     -4.333333
9           43        36.125         6.875
10          -5         -3.25         -1.75
11          43        20.625        22.375
12         -26     -30.66667      4.666667
13          49        28.375        20.625
14         -40         -24.5         -15.5
15         -22       -16.875        -5.125
16                    28.375
```

This report shows the response variable, the predicted value based on the response surface equation, and the residual (the difference between the two).

Notice that a predicted value is given for row sixteen, but no residual or Odor value is given. If you look at row sixteen on the database, you will note that it has a missing value for Odor and thus was not used in estimating the regression equation. However, since there are values for the three independent variables, a predicted value can be generated. This shows how to automatically generate predicted values for a set of X's when the observed Y is not on your database.

Normal Probability Plot

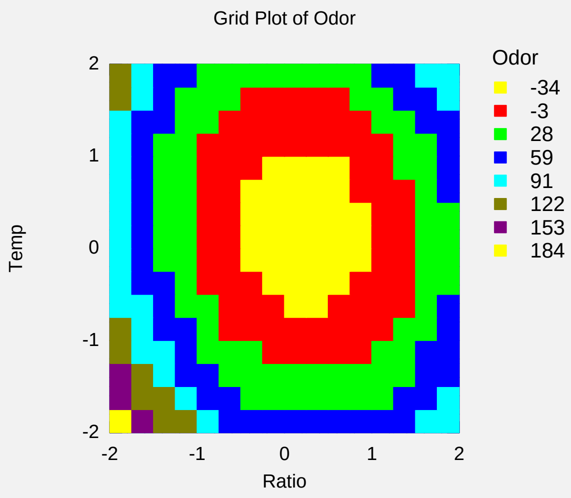
This plot displays a normal probability plot of the residuals for assessing the validity of the assumption of normality. Note that you should ignore this plot when you have less than about five observations per term in the model, since the assumption of independence of residuals cannot be demonstrated and thus the probability plot may give inaccurate results.



Contour Plot

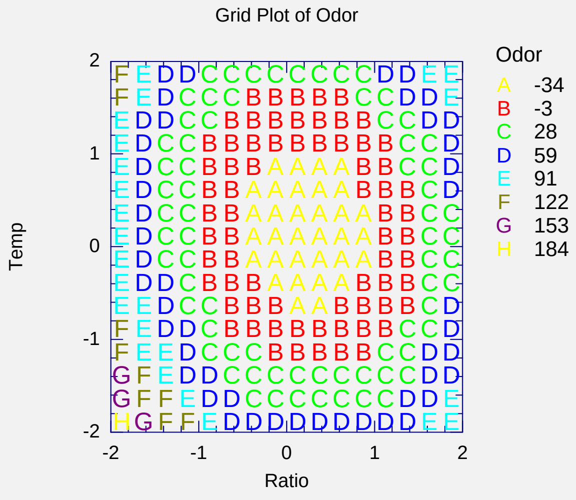
This contour (or grid) plot shows the value of the estimated equation at the center of each grid of rectangles. All factors not on either axis are evaluated at their mean value (unless a constant value was specified in the Factor Constant box).

The legend lists the lower end of each range. Hence, the first contour is -40 <= ***Odor*** < -20.



Contour Plot with Symbols

If you had set the Plot Style option to Symbol - Many Colors, you would have obtained the following plot. This plot is not as pretty, but is easier to view when printed in black and white.



Chapter 335

Ridge Regression

Introduction

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. Another biased regression technique, principal components regression, is also available in **NCSS**. Ridge regression is the more popular of the two methods.

Multicollinearity

Multicollinearity, or collinearity, is the existence of near-linear relationships among the independent variables. For example, suppose that the three ingredients of a mixture are studied by including their percentages of the total. These variables will have the (perfect) linear relationship: P1 + P2 + P3 = 100. During regression calculations, this relationship causes a division by zero which in turn causes the calculations to be aborted. When the relationship is not exact, the division by zero does not occur and the calculations are not aborted. However, the division by a very small quantity still distorts the results. Hence, one of the first steps in a regression analysis is to determine if multicollinearity is a problem.

Effects of Multicollinearity

Multicollinearity can create inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false, nonsignificant, p-values, and degrade the predictability of the model (and that’s just for starters).

Sources of Multicollinearity

To deal with multicollinearity, you must be able to identify its source. The source of the multicollinearity impacts the analysis, the corrections, and the interpretation of the linear model. There are five sources (see Montgomery [1982] for details):

1. *Data collection*. In this case, the data have been collected from a narrow subspace of the independent variables. The multicollinearity has been created by the sampling methodology—it does not exist in the population. Obtaining more data on an expanded range would cure this multicollinearity problem. The extreme example of this is when you try to fit a line to a single point.

2. *Physical constraints* of the linear model or population. This source of multicollinearity will exist no matter what sampling technique is used. Many manufacturing or service processes have constraints on independent variables (as to their range), either physically, politically, or legally, which will create multicollinearity.
3. *Over-defined model*. Here, there are more variables than observations. This situation should be avoided.
4. *Model choice or specification*. This source of multicollinearity comes from using independent variables that are powers or interactions of an original set of variables. It should be noted that if the sampling subspace of independent variables is narrow, then any combination of those variables will increase the multicollinearity problem even further.
5. *Outliers*. Extreme values or outliers in the *X*-space can cause multicollinearity as well as hide it. We call this outlier-induced multicollinearity. This should be corrected by removing the outliers before ridge regression is applied.

Detection of Multicollinearity

There are several methods of detecting multicollinearity. We mention a few.

1. Begin by studying pairwise scatter plots of pairs of independent variables, looking for near-perfect relationships. Also glance at the correlation matrix for high correlations. Unfortunately, multicollinearity does not always show up when considering the variables two at a time.
2. Consider the variance inflation factors (VIF). VIFs over 10 indicate collinear variables.
3. Eigenvalues of the correlation matrix of the independent variables near zero indicate multicollinearity. Instead of looking at the numerical size of the eigenvalue, use the condition number. Large condition numbers indicate multicollinearity.
4. Investigate the signs of the regression coefficients. Variables whose regression coefficients are opposite in sign from what you would expect may indicate multicollinearity.

Correction for Multicollinearity

Depending on what the source of multicollinearity is, the solutions will vary. If the multicollinearity has been created by the data collection, collect additional data over a wider *X*-subspace. If the choice of the linear model has increased the multicollinearity, simplify the model by using variable selection techniques. If an observation or two has induced the multicollinearity, remove those observations. Above all, use care in selecting the variables at the outset.

When these steps are not possible, you might try ridge regression.

Ridge Regression Models

Following the usual notation, suppose our regression equation is written in matrix form as

$$\underline{\mathbf{Y}} = \mathbf{X}\underline{\mathbf{B}} + \underline{\mathbf{e}}$$

where $\underline{\mathbf{Y}}$ is the dependent variable, **X** represents the independent variables, $\underline{\mathbf{B}}$ is the regression coefficients to be estimated, and $\underline{\mathbf{e}}$ represents the errors are residuals.

Standardization

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation, since we must somehow indicate whether the variables in a particular formula are standardized or not. To keep the presentation simple, we will make the following general statement and then forget about standardization and its confusing notation.

As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is in a standardized scale.

Ridge Regression Basics

In ordinary least squares, the regression coefficients are estimated using the formula

$$\hat{\underline{\mathbf{B}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

Note that since the variables are standardized, **X'X** = **R**, where **R** is the correlation matrix of independent variables. These estimates are unbiased so that the expected value of the estimates are the population values. That is,

$$E\left(\hat{\underline{\mathbf{B}}}\right) = \underline{\mathbf{B}}$$

The variance-covariance matrix of the estimates is

$$V\left(\hat{\underline{\mathbf{B}}}\right) = \sigma^2\mathbf{R}^{-1}$$

and since we are assuming that the y's are standardized, $\sigma^2 = 1$.

From the above, we find that

$$V\left(\hat{b}_j\right) = r^{jj} = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared value obtained from regression X_j on the other independent variables. In this case, this variance is the VIF. We see that as the R-squared in the denominator gets closer and closer to one, the variance (and thus VIF) will get larger and larger. The rule of thumb cut-off value for VIF is 10. Solving backwards, this translates into an R-squared value of 0.90. Hence, whenever the R-squared value between one independent variable and the rest is greater than or equal to 0.90, you will have to face multicollinearity.

Now, ridge regression proceeds by adding a small value, *k*, to the diagonal elements of the correlation matrix. (This is where ridge regression gets its name since the diagonal of ones in the correlation matrix may be thought of as a ridge.) That is,

$$\underline{\tilde{\mathbf{B}}} = \left(\mathbf{R} + k\mathbf{I}\right)^{-1}\mathbf{X'}\,\underline{\mathbf{Y}}$$

k is a positive quantity less than one (usually less than 0.3).

The amount of bias in this estimator is given by

$$E\left(\underline{\tilde{\mathbf{B}}} - \underline{\mathbf{B}}\right) = \left[\left(\mathbf{X'}\,\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X'}\,\mathbf{X} - \mathbf{I}\right]\underline{\mathbf{B}}$$

and the covariance matrix is given by

$$V\left(\underline{\tilde{\mathbf{B}}}\right) = \left(\mathbf{X'}\,\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X'}\,\mathbf{X}\left(\mathbf{X'}\,\mathbf{X} + k\mathbf{I}\right)^{-1}$$

It can be shown that there exists a value of *k* for which the mean squared error (the variance plus the bias squared) of the ridge estimator is less than that of the least squares estimator. Unfortunately, the appropriate value of *k* depends on knowing the true regression coefficients (which are being estimated) and an analytic solution has not been found that guarantees the optimality of the ridge solution. We will discuss more about determining *k* later.

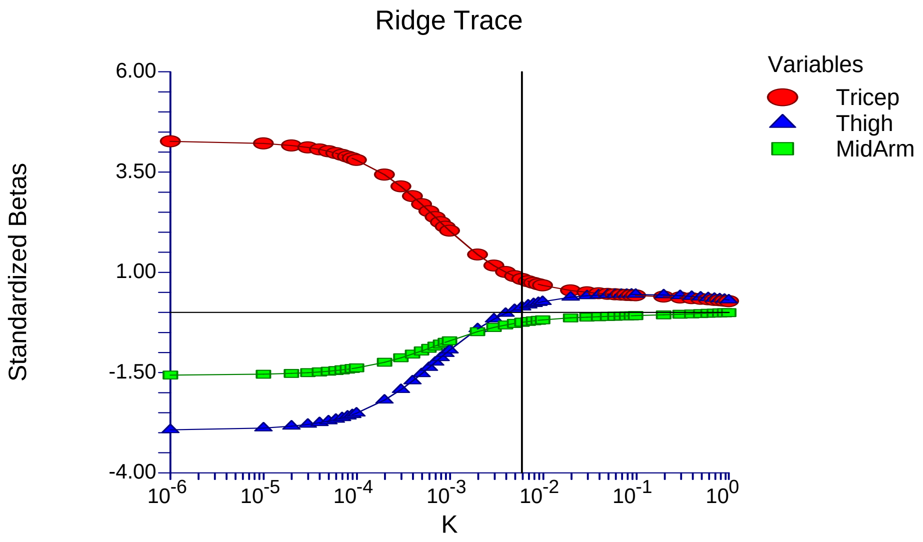
Alternative Interpretations of Ridge Regression

1. Ridge regression may be given a Bayesian interpretation. If we assume that each regression coefficient has expectation zero and variance *1/k*, then ridge regression can be shown to be the Bayesian solution.
2. Another viewpoint is referred to by detractors as the “phoney data” viewpoint. It can be shown that the ridge regression solution is achieved by adding rows of data to the original data matrix. These rows are constructed using 0 for the dependent variables and the square root of *k* or zero for the independent variables. One extra row is added for each independent variable. The idea that manufacturing data yields the ridge regression results has caused a lot of concern and has increased the controversy in its use and interpretation.

Choosing *k*

Ridge Trace

One of the main obstacles in using ridge regression is in choosing an appropriate value of *k*. Hoerl and Kennard (1970), the inventors of ridge regression, suggested using a graphic which they called the *ridge trace*. This plot shows the ridge regression coefficients as a function of *k*. When viewing the ridge trace, the analyst picks a value for *k* for which the regression coefficients have stabilized. Often, the regression coefficients will vary widely for small values of *k* and then stabilize. Choose the smallest value of *k* possible (which introduces the smallest bias) after which the regression coefficients have seem to remain constant. Note that increasing *k* will eventually drive the regression coefficients to zero. Following is an example of a ridge trace.



In this example, the values of *k* are shown on a logarithmic scale. We have drawn a vertical line at the selected value of *k* which is 0.006. A few notes are in order here.

First of all, the vertical axis contains the points for the least squares solution. These are labeled as 0.000001. This was done so that these coefficients may be seen. In actual fact, the logarithm of zero is minus infinity, so the least squares values cannot be displayed when the horizontal axis is put in a log scale.

We have displayed a large range of values. We see that adding *k* has little impact until *k* is about 0.0001. The action seems to stop somewhere near 0.006.

Analytic k

Hoerl and Kinnard (1976) proposed an iterative method for selecting *k*. This method is based on the formula

$$k = \frac{ps^2}{\underset{\sim}{\mathbf{B}}'\underset{\sim}{\mathbf{B}}}$$

To obtain the first value of *k*, we use the least squares coefficients. This produces a value of *k*. Using this new *k*, a new set of coefficients is found, and so on. Unfortunately, this procedure does not necessarily converge. In **NCSS**, we have modified this routine so that if the resulting *k* is greater than one, the new value of *k* is equal to the last value of *k* divided by two.

This calculated value of *k* is often used because humans tend to pick a *k* from the ridge trace that is too large.

Assumptions

The assumptions are the same as those used in regular multiple regression: linearity, constant variance (no outliers), and independence. Since ridge regression does not provide confidence limits, normality need not be assumed.

What the Professionals Say

Ridge regression remains controversial. In this section we will present the comments made in several books on regression analysis.

Neter, Wasserman, and Kutner (1983) state:

“Ridge regression estimates tend to be stable in the sense that they are usually little affected by small changes in the data on which the fitted regression is based. In contrast, ordinary least squares estimates may be highly unstable under these conditions when the independent variables are highly multicollinear. Also, the ridge estimated regression function at times will provide good estimates of mean responses or predictions of new observations for levels of the independent variables outside the region of the observations on which the regression function is based. In contrast, the estimated regression function based on ordinary least squares may perform quite poorly in such instances. Of course, any estimation or prediction well outside the region of the observations should always be made with great caution.

“A major limitation of ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known. Another limitation is that the choice of the biasing constant *k* is a judgmental one. While formal methods have been developed fo making this choice, these methods have their own limitations.”

John O. Rawlings (1988) states:

“While multicollinearity does not affect the precision of the estimated responses (and predictions) at the observed points in the X-space, it does cause variance inflation of estimated responses at other points. Park shows that the restrictions on the parameter estimates implicit in principal component regression are also optimal in MSE sense for estimation of responses over certain regions of the X-space. This suggests that biased regression methods may be beneficial in certain cases for estimation of responses also. The biased regression methods do not seem to have much to offer when the objective is to assign some measure of “relative importance” to the independent variables involved in a multicollinearity… Ridge regression attacks the multicollinearity by reducing the apparent magnitude of the correlations.”

Raymond H. Myers (1990) states:

“Ridge regression is one of the more popular, albeit controversial, estimation procedures for combating multicollinearity. The procedures discussed in this and subsequent sections fall into the category of biased estimation techniques. They are based on this notion: though ordinary least squares gives unbiased estimates and indeed enjoy the minimum variance of all linear unbiased estimators, there is no upper bound on the variance of the estimators and the presence of multicollinearity may produce large variances. As a result, one can visualize that, under the condition of multicollinearity, a huge price is paid for the unbiasedness property that one achieves by using ordinary least squares. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in stability of the regression coefficients. The coefficients become biased and, simply put, if one is successful, the reduction in variance is of greater magnitude than the bias induced in the estimators…

“Although, clearly, ridge regression should not be used to solve all model-fitting problems involving multicollinearity, enough positive evidence about ridge regression exists to suggest that it should be a part of any model builder’s arsenal of techniques.”

Draper and Smith (1981) state:

“From this discussion, we can see that the use of ridge regression is perfectly sensible in circumstances in which it is believed that large beta-values are unrealistic from a practical point of view. However, it must be realized that the choice of *k* is essentially equivalent to an expression of how big one believes those betas to be. In circumstances where one cannot accept the idea of restrictions on the betas, ridge regression would be completely inappropriate.”

“Overall, however, we would advise against the indiscriminate use of ridge regression unless its limitations are fully appreciated.”

Thomas Ryan (1997) states:

“The reader should note that, for all practical purposes, the ordinary least squares (OLS) estimator will also generally be biased because we can be certain that it is unbiased only when the model that is being used is the correct model. Since we cannot expect this to be true, we similarly cannot expect the OLS estimator to be unbiased. Therefore, although the choice between OLS and a ridge estimator is often portrayed as a choice between a biased estimator and an unbiased estimator, that really isn’t the case.”

“Ridge regression permits the use of a set of regressors that might be deemed inappropriate if least squares were used. Specifically, highly correlated variables can be used together, with ridge regression used to reduce the multicollinearity. If, however, the multicollinearity were extreme, such as when regressors are almost perfectly correlated, we would probably prefer to delete one or more regressors before using the ridge approach.”

Data Structure

The data are entered as three or more variables. One variable represents the dependent variable. The other variables represent the independent variables. An example of data appropriate for this procedure is shown below. These data were concocted to have a high degree of multicollinearity as follows. We put a sequence of numbers in X1. Next, we put another series of numbers in X3 that were selected to be unrelated to X1. We created X2 by adding X1 and X3. We made a few changes in X2 so that there was not perfect correlation. Finally, we added all three variables and some random error to form Y.

The data are contained in the RIDGEREG database. We suggest that you open this database now so that you can follow along with the example.

RIDGEREG dataset (subset)

X1	X2	X3	Y
1	2	1	3
2	4	2	9
3	6	4	11
4	7	3	15
5	7	2	13
6	7	1	13
7	8	1	17
8	10	2	21
9	12	4	25
10	13	3	27

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present on a row for all but the dependent variable, a predicted value is generated for that row.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variables

Specifies a dependent (*Y*) variable. If more than one variable is specified, a separate analysis is run for each.

Weight Variable

Weight Variable

Specifies a variable containing observation (row) weights for generating weighted-regression analysis. Rows which have zero or negative weights are dropped.

Independent Variables

X's: Independent Variables

Specifies the variable(s) to be used as independent (*X*) variables.

K Value Specification

Final K (On Reports)

This is the value of *k* that is used in the reports. You may specify a value or enter "Optimum," which will cause the value found in the analytic search for *k* to be used in the reports.

K Value Specification – K Trial Values

Values of K

Various trial values of *k* may be specified. The check boxes on the left select groups of values. For example, checking 0.001 to 0.009 indicates that you want to try the values 0.001, 0.002, 0.003, …, 0.009.

Minimum, Maximum, Increment

Use these options to enter a series of trial *k* values. For example, entering 0.1, 0.2, and 0.02 will cause the following *k* values to be used: 0.10, 0.12, 0.14, 0.16, 0.18, 0.20.

K Value Specification – K Search

K Search

This will cause a search to be made for the optimal value of *k* using the Hoerl’s (1976) algorithm (described above). If “Optimum” is entered for the Final K value, this value will be used in the final reports.

Max Iterations

This value limits the number of *k* values tried in the search procedure. It is provided since it is possible for the search algorithm to go into an infinite loop. A value between 20 and 50 should be ample.

Reports Tab

The following options control the reports and plots that are displayed.

Select Reports

Descriptive Statistics ... Predicted Values & Residuals

These options specify which reports are displayed.

Select Plots

Ridge Trace ... Residuals vs X's

These options specify which plots are displayed.

Report Options

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimal Places

Beta ... VIF Decimals

Each of these options specifies the number of decimal places to display for that particular item.

Plot Options

Show Legend

Indicate whether the legend is to be displayed.

Legend Text

Indicate the title text of the legend. Note that if two factors are being plotted, {G} is replaced by the word “Variables.”

Ridge Trace and VIF Plot Tabs

The options on this panel control the appearance of the ridge trace and the VIF plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Log Scale

This option lets you select logarithmic scale for the corresponding axis of the plot. In the case of the Ridge Trace plot, this is useful since the trial values of *k* will often span several orders of magnitude.

- **No**

  Use regular scaling.

- **Yes: Numbers**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as decimal numbers (e.g., 0.001, 0.01, 0.1).

- **Yes: Powers of Ten**

  Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (10^{-3}, 10^{-2}, 10^{-1}).

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Connect Line(s)

This option lets you specify whether you want to connect the points with a line.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Reference Lines

K Line

This option controls the characteristics of the vertical line at the specified value of *k* that may be displayed on the plot. This is the value of Final K that was specified in the Variables Tab.

0 Line

This option controls the characteristics of the horizontal line at zero that may be displayed on the plot.

Ridge & VIF Symbols Tab

These options specify the symbols used to represent the variables on the Ridge Trace and the VIF Plot.

Plotting Symbols

Variable (1-15)

The symbols used to represent the variables. Variable 1 represents the first variable, Variable 2 represents the second variable, and so on.

Histogram Tab

The options on this panel control the appearance of the histogram.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a histogram style file. This file sets all histogram options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Histogram procedure.

Number of Bars

Specify the number of intervals, bins, or bars used in the histogram.

Titles

Plot Title

This is the text of the title. The characters *{X}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Resid vs Yhat Plot and Resid vs X Plot Tabs

Various residual plots may be displayed to help you validate the assumptions of your regression analysis as well as investigate the fit of your estimated equation. The actual uses of these plots will be described later. The options on these panels control the appearance of the corresponding residual scatter plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Storage Tab

Predicted values and residuals may be calculated for each row and stored on the current database for further analysis. The selected statistics are automatically stored to the current database.

Note that existing data are replaced. Also, if you specify more than one dependent variable, you should specify a corresponding number of storage variables here. Following is a description of the statistics that can be stored.

Data Storage Variables

Predicted Values

The predicted (Yhat) values.

Residuals

The residuals (Y-Yhat).

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Ridge Regression Analysis

This section presents an example of how to run a ridge regression analysis of the data presented earlier in this chapter. The data are in the RIDGEREG database. In this example, we will run a regression of *Y* on *X1* - *X3*.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Ridge Regression window.

1 **Open the RIDGEREG dataset.**

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **RidgeReg.s0**.
- Click **Open**.

2 **Open the Ridge Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines**, then **Ridge Regression**. The Ridge Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Ridge Regression window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable(s)** text box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**. "Y" will appear in the Y: Dependent Variable(s) box.
- Double-click in the **X's: Independent Variables** text box. This will bring up the variable selection window.
- Select **X1 - X3** from the list of variables and then click **Ok**. "X1-X3" will appear in the X's: Independent Variables.
- Select **Optimum** in the **Final K (On Reports)** box. This will cause the optimum value found in the search procedure to be used in all of the reports.
- Check the box for **0.0001 to 0.0009** to include these values of *k*.
- Check the **K Search** box so that the optimal value of *k* will be found.

4 **Specify the reports.**

- Select the **Reports tab**.
- Check all reports and plots. We are selecting all of them so that we can document them.

5 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Descriptive Statistics Section

```
Descriptive Statistics Section
Standard
Variable     Count          Mean     Deviation    Minimum      Maximum
X1              18           9.5      5.338539          1           18
X2              18          11.5      5.404247          2           19
X3              18      2.166667      1.098127          1            4
Y               18      23.11111      10.87841          3           39
```

For each variable, the descriptive statistics of the nonmissing values are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

```
Correlation Matrix Section
                    X1            X2            X3            Y
X1            1.000000      0.987841     -0.015051     0.985544
X2            0.987841      1.000000      0.133813     0.995574
X3           -0.015051      0.133813      1.000000     0.116539
Y             0.985544      0.995574      0.116539     1.000000
```

Pearson correlations are given for all variables. Outliers, nonnormality, nonconstant variance, and nonlinearities can all impact these correlations. Note that these correlations may differ from pair-wise correlations generated by the correlation matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with one another may cause multicollinearity problems.

Least Squares Multicollinearity Section

```
Least Squares Multicollinearity Section
Independent        Variance          R-Squared
Variable           Inflation         Vs Other X's          Tolerance
X1                 477.2665               0.9979              0.0021
X2                 485.8581               0.9979              0.0021
X3                  11.7455               0.9149              0.0851
Since some VIF's are greater than 10, multicollinearity is a problem.
```

This report provides information useful in assessing the amount of multicollinearity in your data.

Variance Inflation

The variance inflation factor (VIF) is a measure of multicollinearity. It is the reciprocal of $1\text{-}R_x^2$, where R_x^2 is the R^2 obtained when this variable is regressed on the remaining independent variables. A VIF of 10 or more for large data sets indicates a multicollinearity problem since the R_x^2 with the remaining X's is 90 percent. For small data sets, even VIF's of 5 or more can signify multicollinearity.

$$VIF_j = \frac{1}{1 - R_j^2}$$

R-Squared vs Other X's

R_x^2 is the R^2 obtained when this variable is regressed on the remaining independent variables. A high R_x^2 indicates a lot of overlap in explaining the variation among the remaining independent variables.

Tolerance

Tolerance is 1- R_x^2, the denominator of the variance inflation factor.

Eigenvalues of Correlations

```
Eigenvalues of Correlations
                              Incremental      Cumulative       Condition
No.            Eigenvalue         Percent         Percent          Number
1                1.994969           66.50           66.50            1.00
2                1.004003           33.47           99.97            1.99
3                0.001027            0.03          100.00         1941.85
Some Condition Numbers greater than 1000. Multicollinearity is a SEVERE problem.
```

This section gives an eigenvalue analysis of the independent variables after they have been centered and scaled. Notice that in this example, the third eigenvalue is very small.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero indicate a multicollinearity problem in your data.

Incremental Percent

Incremental percent is the percent this eigenvalue is of the total. In an ideal situation, these percentages would be equal. Percents near zero indicate a multicollinearity problem in your data.

Cumulative Percent

This is the running total of the Incremental Percent.

Condition Number

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variances. Condition numbers greater than 1000 indicate a severe multicollinearity problem while condition numbers between 100 and 1000 indicate a mild multicollinearity problem.

Eigenvector of Correlations

```
Eigenvector of Correlations
No.            Eigenvalue              X1              X2              X3
1                1.994969        0.701391        0.707741        0.084573
2                1.004003       -0.134162        0.014553        0.990853
3                0.001027        0.700036       -0.706322        0.105159
```

This report displays the eigenvectors associated with each eigenvalue. The notion behind eigenvalue analysis is that the axes are rotated from the ones defined by the variables to a new set defined by the variances of the variables. Rotating is accomplished by taking weighted averages of the original variables. Thus, the first new axis could be the average of X1 and X2. The first new variable is constructed to account for the largest amount of variance possible from a single axis.

No.

The number of the eigenvalue.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is multicollinearity in your data. The eigenvalues represent the spread (variance) in the direction defined by this new axis. Hence, small eigenvalues indicate directions in which there is no spread. Since regression analysis seeks to find trends across values, when there is not a spread, the trends cannot be computed.

Table-Values

The table values give the eigenvectors. The eigenvectors give the weights that are used to create the new axis. By studying the weights, you can gain an understanding of what is happening in the data.

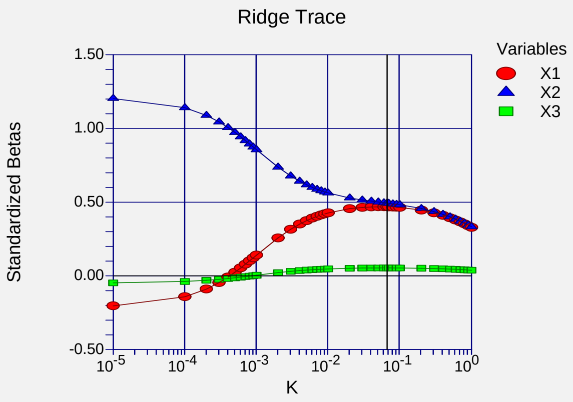
In the example above, we can see that the first factor (new variable associated with the first eigenvalue) is constructed by adding X1 and X2. Note that the weights are almost equal. X3 has a small weight, indicating that it does not play a role in this factor.

Factor 2 seems to be created complete from X3. X1 and X2 play only a small role in its construction.

Factor 3 seems to be the difference between X1 and X2. Again X3 plays only a small role. Hence, the interpretation of these eigenvectors leads to the following statements:

1. Most of the variation in X1, X2, and X3 can be accounted for by considering only two variables: Z = X1+X2 and X3.
2. The third dimension, calculated as X1-X2, is almost negligible and might be ignored.

Ridge Trace Section

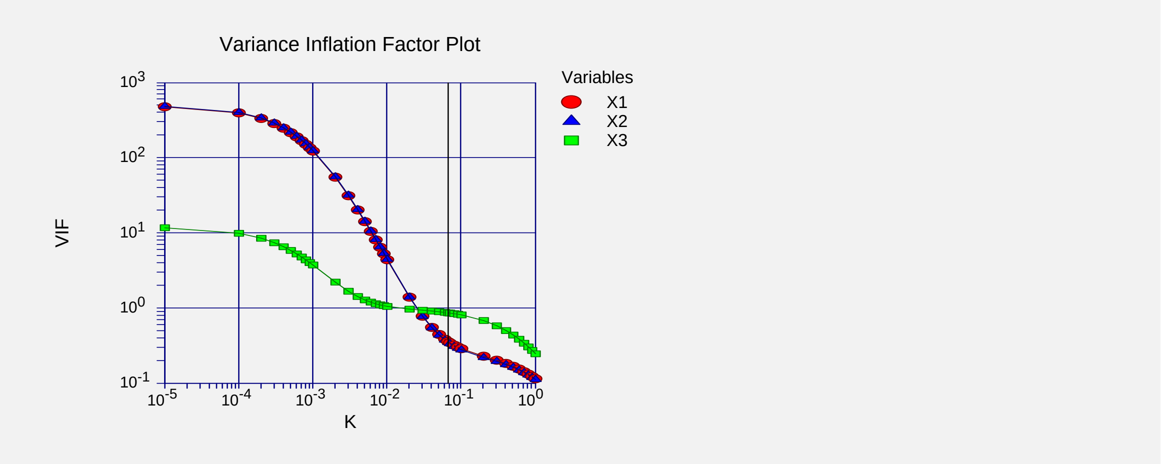


This is the famous *ridge trace* that is the signature of this technique. The plot is really very straight forward to read. It presents the standardized regression coefficients on the vertical axis and various values of *k* along the horizontal axis. Since the values of *k* span several orders of magnitude, we adopt a logarithmic scale along this axis.

The points on the left vertical axis (the left ends of the lines) are the ordinary least squares regression values. These occur for *k* equal zero. As *k* is increased, the values of the regression estimates change, often wildly at first. At some point, the coefficients seem to settle down and then gradually drift towards zero.

The task of the ridge regression analyst is to determine at what value of *k* these coefficients are at their stable values. A vertical line is drawn at the value selected for reporting purposes. It is anticipated that you would run the program several times until an appropriate value of *k* is determined. In this example, our search would be between 0.0001 and 0.1. The value selected on this graph happens to be 0.066237, the value obtained from the analytic search. We might be inclined to use an even smaller value of *k* such as 0.01. Remember, the smaller the value of *k*, the smaller the amount of bias that is included in the estimates.

Variance Inflation Factor Plot



This is a plot that we have added that shows the impact of *k* on the variance inflation factors. Since the major goal of ridge regression is to remove the impact of multicollinearity, it is important to know at what point multicollinearity has been dealt with. This plot shows this.

The currently selected value of *k* is shown by a vertical line.

Since the rule-of-thumb is that multicollinearity is not a problem once all VIFs are less than 10, we inspect the graph for this point. In this example, it appears that all VIFs are small enough once *k* is greater than 0.007. Hence, this is the value of *k* that this plot would indicate we use.

Since this plot indicates *k* = 0.007 and the ridge trace indicates a value near 0.01, we would select 0.007 as our final result. The rest of the reports are generated for this value of *k*.

Standardized Ridge Regression Coefficients Section

```
Standardized Ridge Regression Coefficients Section
k                      X1           X2           X3
0.000000          -0.2034       1.2029      -0.0475
0.000100          -0.1415       1.1404      -0.0382
0.000200          -0.0897       1.0881      -0.0304
0.000300          -0.0457       1.0436      -0.0238
0.000400          -0.0079       1.0054      -0.0181
0.000500           0.0249       0.9722      -0.0132
0.000600           0.0538       0.9431      -0.0088
0.000700           0.0793       0.9173      -0.0050
0.000800           0.1019       0.8944      -0.0016
0.000900           0.1223       0.8738       0.0015
0.001000           0.1406       0.8553       0.0042
0.002000           0.2572       0.7371       0.0217
0.003000           0.3157       0.6776       0.0305
0.004000           0.3509       0.6416       0.0358
0.005000           0.3743       0.6174       0.0394
0.006000           0.3910       0.6001       0.0419
0.007000           0.4035       0.5870       0.0438
0.008000           0.4131       0.5768       0.0452
0.009000           0.4208       0.5686       0.0464
0.010000           0.4270       0.5618       0.0473
0.020000           0.4555       0.5281       0.0517
0.030000           0.4641       0.5146       0.0531
0.040000           0.4673       0.5065       0.0537
0.050000           0.4684       0.5006       0.0539
0.060000           0.4683       0.4960       0.0540
0.066237           0.4680       0.4934       0.0540
0.070000           0.4677       0.4919       0.0540
0.080000           0.4666       0.4884       0.0539
0.090000           0.4653       0.4851       0.0538
```

This report gives the values that are plotted on the ridge trace. Note that the value found by the analytic search (0.066237) sticks out as you glance down the first column because it does not end in zeros.

Variance Inflation Factor Section

```
Variance Inflation Factor Section
k                     X1             X2             X3
0.000000        477.2665       485.8581        11.7455
0.000100        396.3965       403.5292         9.9204
0.000200        334.4756       340.4914         8.5229
0.000300        286.0151       291.1566         7.4291
0.000400        247.3784       251.8230         6.5570
0.000500        216.0793       219.9592         5.8505
0.000600        190.3708       193.7870         5.2702
0.000700        168.9966       172.0273         4.7877
0.000800        151.0345       153.7411         4.3822
0.000900        135.7951       138.2268         4.0381
0.001000        122.7546       124.9510         3.7436
0.002000         55.1972        56.1749         2.2172
0.003000         31.3037        31.8505         1.6761
0.004000         20.1831        20.5293         1.4232
0.005000         14.1214        14.3583         1.2845
0.006000         10.4576        10.6285         1.1999
0.007000          8.0755         8.2035         1.1442
0.008000          6.4402         6.5386         1.1054
0.009000          5.2691         5.3465         1.0771
0.010000          4.4019         4.4637         1.0557
0.020000          1.3976         1.4055         0.9693
0.030000          0.7792         0.7763         0.9372
0.040000          0.5527         0.5460         0.9146
0.050000          0.4443         0.4360         0.8951
0.060000          0.3835         0.3744         0.8771
0.066237          0.3581         0.3487         0.8664
0.070000          0.3456         0.3361         0.8602
0.080000          0.3200         0.3103         0.8439
0.090000          0.3016         0.2919         0.8283
0.100000          0.2878         0.2781         0.8131
```

This report gives the values that are plotted on the variance inflation factor plot. Note how easy it is to determine when all three VIFs are less than 10.

K Analysis Section

```
K Analysis Section
k                    R2          Sigma          B'B       Ave VIF       Max VIF
0.000000         0.9915         1.1028       1.4905      324.9567      485.8581
0.000100         0.9914         1.1119       1.3219      269.9487      403.5292
0.000200         0.9913         1.1199       1.1929      227.8300      340.4914
0.000300         0.9912         1.1272       1.0918      194.8669      291.1566
0.000400         0.9911         1.1339       1.0113      168.5862      251.8230
0.000500         0.9910         1.1401       0.9460      147.2964      219.9592
0.000600         0.9909         1.1459       0.8924      129.8093      193.7870
0.000700         0.9908         1.1513       0.8478      115.2705      172.0273
0.000800         0.9907         1.1565       0.8103      103.0526      153.7411
0.000900         0.9906         1.1614       0.7785       92.6867      138.2268
0.001000         0.9905         1.1661       0.7513       83.8164      124.9510
0.002000         0.9899         1.2065       0.6100       37.8631      56.1749
0.003000         0.9893         1.2406       0.5597       21.6101      31.8505
0.004000         0.9887         1.2719       0.5360       14.0452      20.5293
0.005000         0.9882         1.3014       0.5229        9.9214      14.3583
0.006000         0.9877         1.3297       0.5148        7.4287      10.6285
0.007000         0.9872         1.3571       0.5093        5.8077      8.2035
0.008000         0.9867         1.3837       0.5054        4.6947      6.5386
0.009000         0.9862         1.4096       0.5025        3.8976      5.3465
0.010000         0.9857         1.4349       0.5002        3.3071      4.4637
0.020000         0.9807         1.6639       0.4891        1.2575      1.4055
0.030000         0.9759         1.8619       0.4830        0.8309      0.9372
0.040000         0.9711         2.0389       0.4778        0.6711      0.9146
0.050000         0.9663         2.2000       0.4729        0.5918      0.8951
0.060000         0.9616         2.3487       0.4682        0.5450      0.8771
0.066237         0.9587         2.4361       0.4653        0.5244      0.8664
0.070000         0.9570         2.4871       0.4636        0.5140      0.8602
0.080000         0.9523         2.6170       0.4591        0.4914      0.8439
0.090000         0.9478         2.7396       0.4547        0.4739      0.8283
0.100000         0.9432         2.8558       0.4503        0.4597      0.8131
```

This report provides a quick summary of the various statistics that might go into the choice of *k*.

k

This is the actual value of *k*. Note that the value found by the analytic search (0.066237) sticks out as you glance down this column because it does not end in zeros.

R2

This is the value of R-squared. Since the least squares solution maximizes R-squared, the largest value of R-squared occurs when *k* is zero. We want to select a value of *k* that does not stray very much from this value.

Sigma

This is the square root of the mean squared error. Least squares minimizes this value, so we want to select a value of *k* that does not stray very much from the least squares value.

B'B

This is the sum of the squared standardized regression coefficients. Ridge regression assumes that this value is too large and so the method tries to reduce this. We want to find a value for *k* at which this value has stabilized.

Ave VIF

This is the average of the variance inflation factors.

Max VIF

This is the maximum variance inflation factor. Since we are looking for that value of *k* which results in all VIFs being less than 10, this value is very helpful in your selection of *k*.

Ridge vs. Least Squares Comparison Section

```
Ridge vs. Least Squares Comparison Section for k = 0.007000
                 Regular       Regular         Stand'zed   Stand'zed        Ridge          L.S.
Independent      Ridge         L.S.                Ridge        L.S.     Standard      Standard
Variable         Coeff's       Coeff's           Coeff's     Coeff's        Error         Error
Intercept        0.7721754     0.2230599
X1               0.8221589     -0.4144863         0.4035     -0.2034    0.1752076      1.094502
X2               1.181684      2.421286           0.5870      1.2029    0.1744428      1.090883
X3               0.4334136     -0.4703622         0.0438     -0.0475    0.3206249     0.8347205

R-Squared        0.9872        0.9915
Sigma            1.3571        1.1028
```

This report provides a detailed comparison between the ridge regression solution and the ordinary least squares solution to the estimation of the regression coefficients.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of b_0.

Regular Ridge (and L.S.) Coeff’s

These are the estimated values of the regression coefficients b_0, b_1, ..., b_p. The first column gives the values for ridge regression and the second column gives the values for regular least squares regression.

The value indicates how much change in *Y* occurs for a one-unit change in *x* when the remaining *X’s* are held constant. These coefficients are also called partial-regression coefficients since the effect of the other *X’s* is removed.

Stand’zed Ridge (and L.S.) Coeff’s

These are the estimated values of the standardized regression coefficients. The first column gives the values for ridge regression and the second column gives the values for regular least squares regression.

Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is:

$$b_{j,\,std} = b_j \left(\frac{s_y}{s_{x_j}} \right)$$

where s_y and s_{x_j} are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

Ridge (and L.S.) Standard Error

These are the estimated standard errors (precision) of the regression coefficients. The first column gives the values for ridge regression and the second column gives the values for regular least squares regression.

The standard error of the regression coefficient, s_{b_j} , is the standard deviation of the estimate.

Since one of the objects of ridge regression is to reduce this (make the estimates more precise), it is of interest to see how much reduction has taken place.

R-Squared

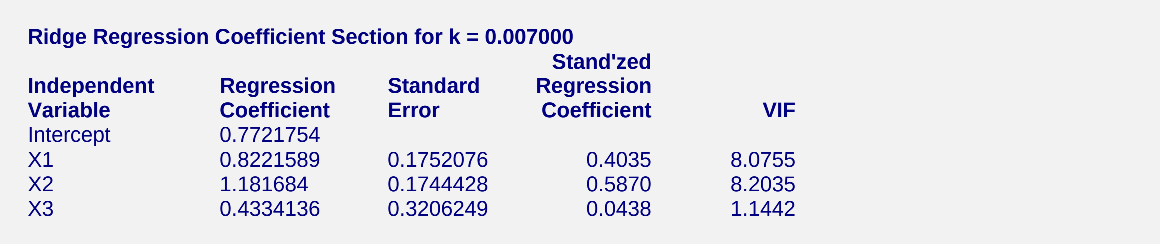
R-squared is the coefficient of determination. It represents the percent of variation in the dependent variable explained by the independent variables in the model. The R-squared values of both the ridge and regular regressions are shown.

Sigma

This is the square root of the mean square error. It provides a measure of the standard deviation of the residuals from the regression model.

It represents the percent of variation in the dependent variable explained by the independent variables in the model. The R-squared values of both the ridge and regular regressions are shown.

Ridge Regression Coefficient Section



```
Ridge Regression Coefficient Section for k = 0.007000
                                                         Stand'zed
Independent          Regression        Standard          Regression
Variable             Coefficient       Error             Coefficient          VIF
Intercept            0.7721754
X1                   0.8221589         0.1752076            0.4035         8.0755
X2                   1.181684          0.1744428            0.5870         8.2035
X3                   0.4334136         0.3206249            0.0438         1.1442
```

This report provides the details of the ridge regression solution.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of b_0.

Regression Coefficient

These are the estimated values of the regression coefficients b_0, b_1, ..., b_p. The value indicates how much change in *Y* occurs for a one-unit change in *x* when the remaining *X's* are held constant. These coefficients are also called partial-regression coefficients since the effect of the other *X's* is removed.

Standard Error

These are the estimated standard errors (precision) of the ridge regression coefficients. The standard error of the regression coefficient, s_{b_j} , is the standard deviation of the estimate. In regular regression, we divide the coefficient by the standard error to obtain a t statistic. However, this is not possible here because of the bias in the estimates.

Stand'zed Regression Coefficient

These are the estimated values of the standardized regression coefficients. Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is

$$b_{j,\,std} = b_j \left(\frac{s_y}{s_{x_j}} \right)$$

where s_y and s_{x_j} are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

VIF

These are the values of the variance inflation factors associated with the variables. When multicollinearity has been conquered, these values will all be less than 10. Details of what VIF is were given earlier.

Analysis of Variance Section

```
Analysis of Variance Section
                                               Sum of       Mean                          Prob
Source           DF            Squares         Square       F-Ratio        Level
Intercept        1             9614.223        9614.223
Model            3             1985.993        661.9978     359.4414       0.000000
Error            14            25.78437        1.841741
Total(Adjusted)  17            2011.778        118.3399

Mean of Dependent              23.11111
Root Mean Square Error         1.357108
R-Squared                      0.9872
Coefficient of Variation       0.058721
```

An analysis of variance (ANOVA) table summarizes the information related to the sources of variation in data.

Source

This represents the partitions of the variation in *y*. There are four sources of variation listed: intercept, model, error, and total (adjusted for the mean).

DF

The degrees of freedom are the number of dimensions associated with this term. Note that each observation can be interpreted as a dimension in *n*-dimensional space. The degrees of freedom for the intercept, model, error, and adjusted total are 1, *p*, *n*-*p*-1, and *n*-1, respectively.

Sum of Squares

These are the sums of squares associated with the corresponding sources of variation. Note that these values are in terms of the dependent variable, *y*.

Mean Square

The mean square is the sum of squares divided by the degrees of freedom. This mean square is an estimated variance. For example, the mean square error is the estimated variance of the residuals (the residuals are sometimes called the *errors*).

F-Ratio

This is the F statistic for testing the null hypothesis that all $\beta_j = 0$. This F-statistic has p degrees of freedom for the numerator variance and n-p-1 degrees of freedom for the denominator variance.

Since ridge regression produces biased estimates, this F-Ratio is not a valid test. It serves as an index, but it would not stand up under close scrutiny.

Prob Level

This is the p-value for the above F test. The p-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the p-value is less than α, say 0.05, the null hypothesis is rejected. If the p-value is greater than α, then the null hypothesis is accepted.

Root Mean Square Error

This is the square root of the mean square error. It is an estimate of σ, the standard deviation of the e_i*'s*.

Mean of Dependent Variable

This is the arithmetic mean of the dependent variable.

R-Squared

This is the coefficient of determination. It is defined in full in the Multiple Regression chapter.

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, computed by dividing root mean square error by the mean of the dependent variable. By itself, it has little value, but it can be useful in comparative studies.

$$CV = \frac{\sqrt{MSE}}{\bar{y}}$$

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section
Row      Actual         Predicted       Residual
1        3              4.391116        -1.391116
2        9              8.010056        0.9899445
3        11             12.06241        -1.062409
4        15             13.63284        1.367162
5        13             14.02158        -1.021584
6        13             14.41033        -1.410329
7        17             16.41417        0.5858285
8        21             20.03311        0.9668885
9        25             24.08546        0.9145348
10       27             25.65589        1.344106
11       25             26.04464        -1.044639
12       27             26.43338        0.5666152
13       29             28.43723        0.5627726
14       33             32.05617        0.9438325
15       35             36.10852        -1.108521
16       37             37.67895        -0.6789502
17       37             38.0677         -1.067695
18       39             38.45644        0.5435593
```

This section reports the predicted values and the sample residuals, or e_i's. When you want to generate predicted values for individuals not in your sample, add their values to the bottom of your database, leaving the dependent variable blank. Their predicted values will be shown on this report.

Actual

This is the actual value of *Y* for the i^{th} row.

Predicted

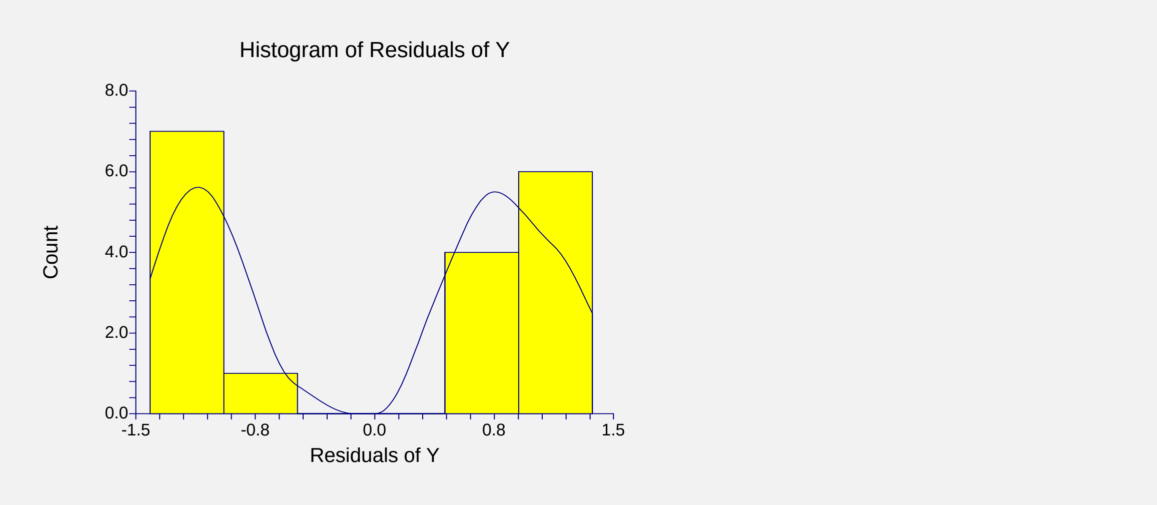
The predicted value of *Y* for the i^{th} row. It is predicted using the levels of the *X's* for this row.

Residual

This is the estimated value of e_i. This is equal to the *Actual* minus the *Predicted.*

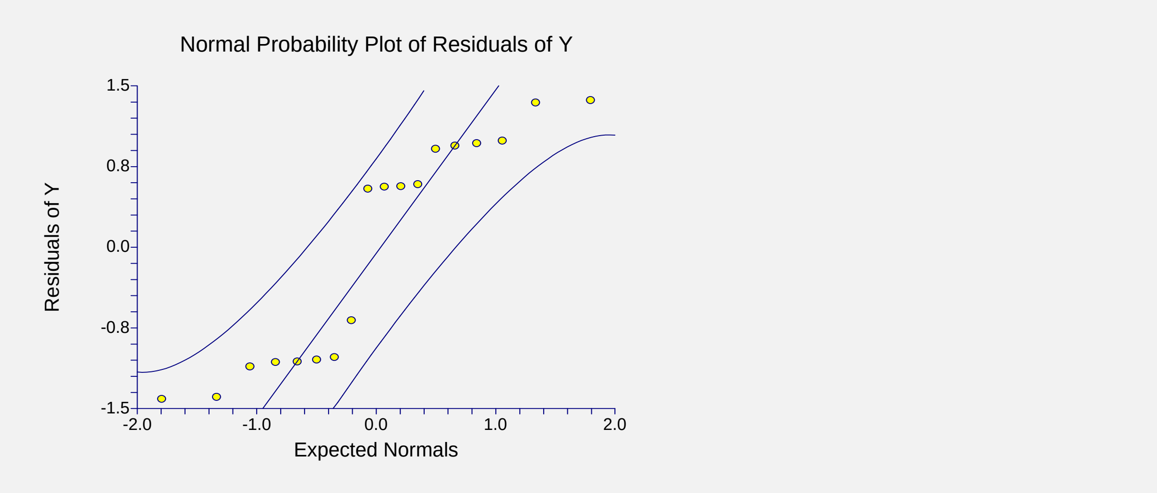
Histogram

The purpose of the histogram and density trace of the residuals is to display the distribution of the residuals.



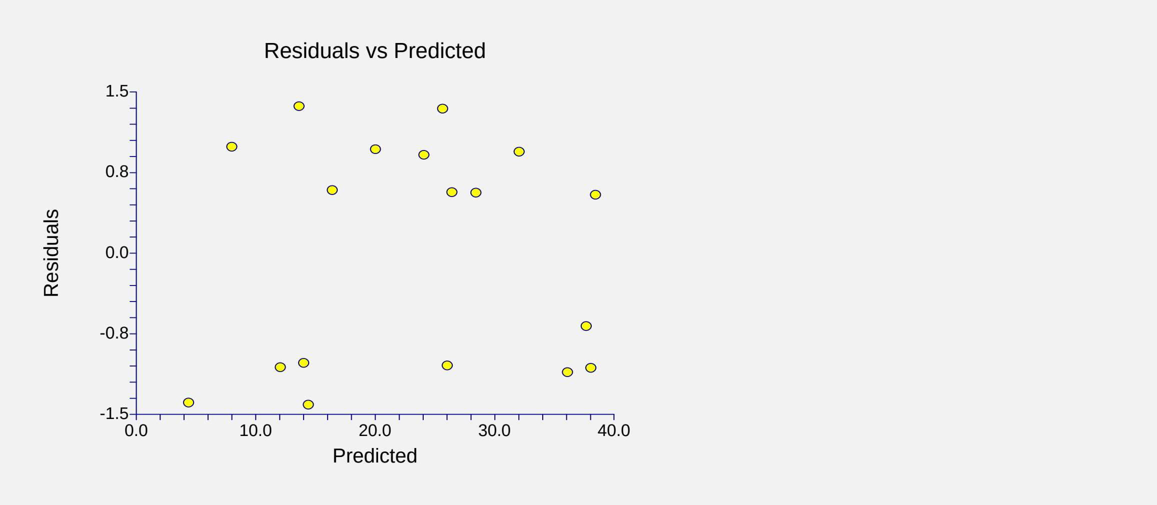
The odd shape of this histogram occurs because of the way in which these particular data were manufactured.

Probability Plot of Residuals



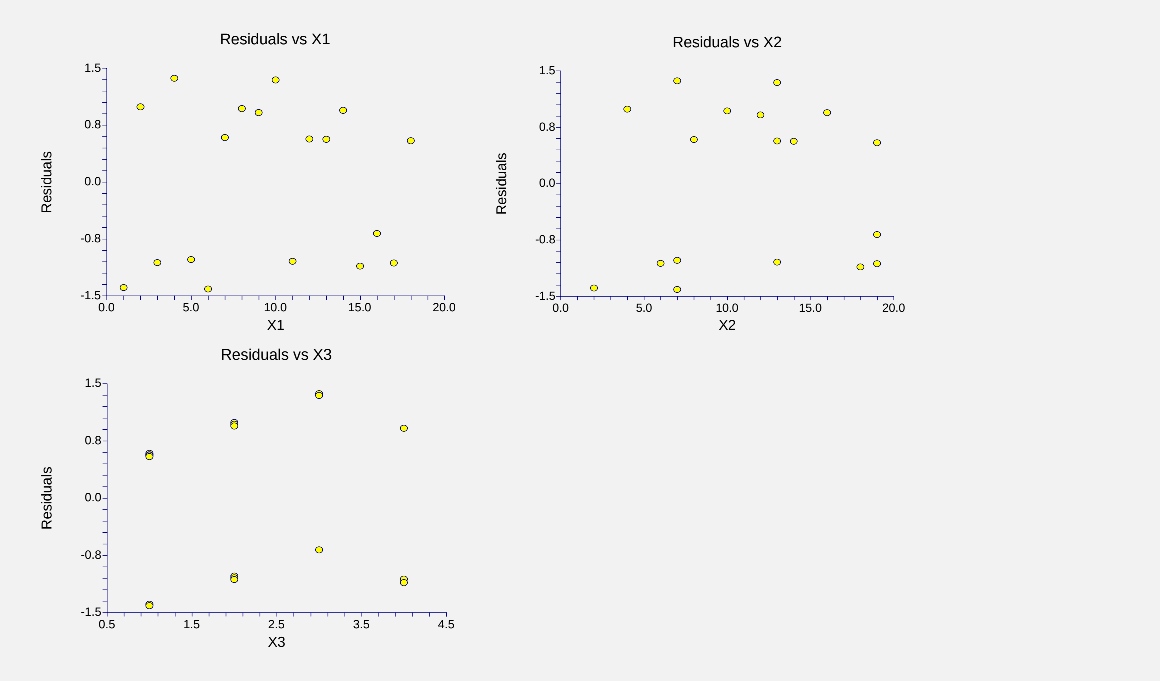
Residual vs Predicted Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance, a violation of a critical regression assumption. A sloping or curved band signifies inadequate specification of the model. A sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.



Residual vs Predictor(s) Plot

This is a scatter plot of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.



Chapter 340

Principal Components Regression

Introduction

Principal Components Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, principal components regression reduces the standard errors. It is hoped that the net effect will be to give more reliable estimates. Another biased regression technique, ridge regression, is also available in **NCSS**. Ridge regression is the more popular of the two methods.

Multicollinearity

Multicollinearity is discussed both in the Multiple Regression chapter and in the Ridge Regression chapter, so we will not repeat the discussion here. However, it is important to understand the impact of multicollinearity so that you can decide if some evasive action (like pc regression) would be beneficial.

Principal Components Regression Models

Following the usual notation, suppose our regression equation may be written in matrix form as

$$\underline{\mathbf{Y}} = \mathbf{X}\underline{\mathbf{B}} + \underline{\mathbf{e}}$$

where $\underline{\mathbf{Y}}$ is the dependent variable, $\mathbf{X}$ represents the independent variables, $\underline{\mathbf{B}}$ is the regression coefficients to be estimated, and $\underline{\mathbf{e}}$ represents the errors or residuals.

Standardization

The first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation, since we must somehow indicate whether the variables in a particular formula are standardized or not. To keep the presentation simple, we will make the following general statement and then forget about standardization and its confusing notation.

As far as standardization is concerned, all calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back to their original scale.

PC Regression Basics

In ordinary least squares, the regression coefficients are estimated using the formula

$$\hat{\underline{\mathbf{B}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

Note that since the variables are standardized, **X'X** = **R**, where **R** is the correlation matrix of independent variables.

To perform principal components (PC) regression, we transform the independent variables to their principal components. Mathematically, we write

$$\mathbf{X}'\mathbf{X} = \mathbf{PDP}' = \mathbf{Z}'\mathbf{Z}$$

where **D** is a diagonal matrix of the eigenvalues of **X'X**, **P** is the eigenvector matrix of **X'X**, and **Z** is a data matrix (similar in structure to **X**) made up of the principal components. **P** is orthogonal so that **P'P** = **I**.

We have created new variables **Z** as weighted averages of the original variables **X**. This is nothing new to us since we are used to using transformations such as the logarithm and the square root on our data values prior to performing the regression calculations. Since these new variables are principal components, their correlations with each other are all zero. If we begin with variables X1, X2, and X3, we will end up with Z1, Z2, and X3.

Severe multicollinearity will be detected as very small eigenvalues. To rid the data of the multicollinearity, we omit the components (the z's) associated with small eigenvalues. Usually, only one or two relatively small eigenvalues will be obtained. For example, if only one small eigenvalue were detected on a problem with three independent variables, we would omit Z3 (the third principal component).

When we regress **Y** on Z1 and Z2, multicollinearity is no longer a problem. We can then transform our results back to the **X** scale to obtain estimates of **B**. These estimates will be biased, but we hope that the size of this bias is more than compensated for by the decrease in variance. That is, we hope that the mean squared error of these estimates is less than that for least squares.

Mathematically, the estimation formula becomes

$$\hat{\underline{\mathbf{A}}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\underline{\mathbf{Y}} = \mathbf{D}^{-1}\mathbf{Z}'\underline{\mathbf{Y}}$$

because of the special nature of principal components. Notice that this is ordinary least squares regression applied to a different set of independent variables.

The two sets of regression coefficients, **A** and **B**, are related using the formulas

$$\underline{\mathbf{A}} = \mathbf{P}'\underline{\mathbf{B}}$$

and

$$\underline{\mathbf{B}} = \mathbf{P}\underline{\mathbf{A}}$$

Omitting a principal component may be accomplished by setting the corresponding element of $\underline{\mathbf{A}}$ equal to zero. Hence, the principal components regression may be outlined as follows:

1. Complete a principal components analysis of the **X** matrix and save the principal components in **Z**.
2. Fit the regression of **Y** on **Z** obtaining least squares estimates of **A**.
3. Set the last element of **A** equal to zero.
4. Transform back to the original coefficients using $\underline{\mathbf{B}} = \mathbf{P}\underline{\mathbf{A}}$.

Alternative Interpretation of PC Regression

It can be shown that omitting a principal component amounts to setting a linear constraint on the regression coefficients. That is, in the case of three independent variables, we add the constraint

$$p_{13}b_1 + p_{23}b_2 + p_{33}b_3 = 0$$

Note that this is a constraint on the coefficients, not a constraint on the dependent variable. Essentially, we have avoided the multicollinearity problem by avoiding the region of the solution space in which it occurs.

How Many PC's Should Be Omitted

Unlike the selection of *k* in ridge regression, the selection of the number of PC's to omit is relatively straight forward. We omit the PC's corresponding to small eigenvalues. Since the size of the typical eigenvalue of a correlation matrix is one, we omit those that are much smaller than one. Usually, the choice will be obvious.

Assumptions

The assumptions are the same as those used in regular multiple regression: linearity, constant variance (no outliers), and independence. Since PC regression does not provide confidence limits, normality need not be assumed.

Data Structure

The data are entered as three or more variables. One variable represents the dependent variable. The other variables represent the independent variables. An example of data appropriate for this procedure is shown below. These data were concocted to have a high degree of multicollinearity as follows. We put a sequence of numbers in X1. Next, we put another series of numbers in X3 that were selected to be unrelated to X1. We created X2 by adding X1 and X3. We made a few changes in X2 so that there was not perfect correlation. Finally, we added all three variables and some random error to form Y.

The data are contained in the RIDGEREG database. We suggest that you open this database now so that you can follow along with the example.

RIDGEREG dataset (subset)

X1	X2	X3	Y
1	2	1	3
2	4	2	9
3	6	4	11
4	7	3	15
5	7	2	13
6	7	1	13
7	8	1	17
8	10	2	21
9	12	4	25
10	13	3	27
11	13	2	25
12	13	1	27
13	14	1	29
14	16	2	33
15	18	4	35
16	19	3	37
17	19	2	37
18	19	1	39

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present on a row for all but the dependent variable, a predicted value is generated for that row.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variables

Y: Dependent Variable(s)

Specifies a dependent (*Y*) variable. If more than one variable is specified, a separate analysis is run for each.

Weight Variable

Weight Variable

Specifies a variable containing observation (row) weights for generating weighted-regression analysis. Rows which have zero or negative weights are dropped.

Independent Variables

X’s: Independent Variables

Specifies the variable(s) to be used as independent (*X*) variables.

Estimation Options

PC’s Omitted

This is the number of principal components that are omitted during the estimation procedure. PC regression is most useful in those cases where this value is set to one or two. If more than two eigenvalues are small, you probably should take some other evasive action such as completely removing the offending variable(s) from consideration.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Descriptive Statistics ... Predicted Values & Residuals

These options specify which reports are displayed.

Select Plots

Beta Trace ... Residuals vs X's

These options specify which plots are displayed.

Report Options

Precision

Specifies the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimal Places

Beta ... VIF Decimals

Each of these options specifies the number of decimal places to display for that particular item.

Plot Options

Show Legend

Indicate whether the legend is to be displayed.

Legend Text

Indicate the title text of the legend. Note that if two factors are being plotted, {G} is replaced by the word "Variables."

Beta Trace and VIF Plot Tabs

The options on this panel control the appearance of the beta trace and the VIF plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Log Scale (VIF Plot only)

This option lets you select logarithmic scale for the vertical axis of the VIF plot.

No

Use regular scaling.

Yes: Numbers

Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as decimal numbers (e.g., 0.001, 0.01, 0.1).

Yes: Powers of Ten

Use logarithmic scaling (base 10) in which the tick reference numbers are displayed as the exponents of ten (10^{-3}, 10^{-2}, 10^{-1}).

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Connect Line(s)

This option lets you specify whether you want to connect the points with a line.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Reference Lines

PC’s Line

This option controls the characteristics of the vertical line drawn at the specified value of the principal components.

0 Line

This option controls the characteristics of the horizontal line at zero that may be displayed on the plot.

Beta & VIF Symbols Tab

These options specify the symbols used to represent the variables on the Beta Trace and the VIF Plot.

Plotting Symbols

Variable (1-15)

The symbols used to represent the variables. Variable 1 represents the first variable, Variable 2 represents the second variable, and so on.

Histogram Tab

The options on this panel control the appearance of the histogram.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a histogram style file. This file sets all histogram options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Histogram procedure.

Number of Bars

Specify the number of intervals, bins, or bars used in the histogram.

Titles

Plot Title

This is the text of the title. The characters *{X}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Resid vs Yhat Plot and Resid vs X Plot Tabs

Various residual plots may be displayed to help you validate the assumptions of your regression analysis as well as investigate the fit of your estimated equation. The actual uses of these plots will be described later. The options on these panels control the appearance of the corresponding residual scatter plot.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Storage Tab

Predicted values and residuals may be calculated for each row and stored on the current database for further analysis. The selected statistics are automatically stored to the current database.

Note that existing data are replaced. Also, if you specify more than one dependent variable, you should specify a corresponding number of storage variables here. Following is a description of the statistics that can be stored.

Data Storage Variables

Predicted Values

The predicted (Yhat) values.

Residuals

The residuals (Y-Yhat).

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Principal Components Regression

This section presents an example of how to run a principal components regression analysis of the data presented above. The data are in the RIDGEREG database. In this example, we will run a regression of *Y* on *X1* - *X3*.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Principal Components Regression window.

1 Open the RIDGEREG dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **RidgeReg.s0**.
- Click **Open**.

2 Open the Principal Components Regression window.

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Other Regression Routines**, then **Principal Components Regression**. The Principal Components Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Principal Components Regression window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable(s)** text box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**. "Y" will appear in the Y: Dependent Variable(s) box.
- Double-click in the **X's: Independent Variables** text box. This will bring up the variable selection window.
- Select **X1 - X3** from the list of variables and then click **Ok**. "X1-X3" will appear in the X's: Independent Variables.

4 Specify the reports.

- Select the **Reports tab**.
- Check all reports and plots. We are selecting all of them so that we can document them.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Descriptive Statistics Section

```
Descriptive Statistics Section
Standard
Variable      Count          Mean     Deviation    Minimum       Maximum
X1               18           9.5      5.338539          1            18
X2               18          11.5      5.404247          2            19
X3               18      2.166667      1.098127          1             4
Y                18      23.11111      10.87841          3            39
```

For each variable, the descriptive statistics of the nonmissing values are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

```
Correlation Matrix Section
                     X1            X2            X3            Y
X1             1.000000      0.987841     -0.015051     0.985544
X2             0.987841      1.000000      0.133813     0.995574
X3            -0.015051      0.133813      1.000000     0.116539
Y              0.985544      0.995574      0.116539     1.000000
```

Pearson correlations are given for all variables. Outliers, nonnormality, nonconstant variance, and nonlinearities can all impact these correlations. Note that these correlations may differ from pair-wise correlations generated by the correlation matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with one another may cause multicollinearity problems.

Least Squares Multicollinearity Section

```
Least Squares Multicollinearity Section
Independent        Variance          R-Squared
Variable           Inflation        Vs Other X's          Tolerance
X1                  477.2665              0.9979             0.0021
X2                  485.8581              0.9979             0.0021
X3                   11.7455              0.9149             0.0851
Since some VIF's are greater than 10, multicollinearity is a problem.
```

This report provides information useful in assessing the amount of multicollinearity in your data.

Variance Inflation

The variance inflation factor (VIF) is a measure of multicollinearity. It is the reciprocal of 1-R_x^2, where R_x^2 is the R^2 obtained when this variable is regressed on the remaining independent variables. A VIF of 10 or more for large data sets indicates a multicollinearity problem since the R_x^2 with the remaining X's is 90 percent. For small data sets, even VIF's of 5 or more can signify multicollinearity.

$$VIF_j = \frac{1}{1 - R_j^2}$$

R-Squared vs Other X's

R_x^2 is the R-squared obtained when this variable is regressed on the remaining independent variables. A high R_x^2 indicates a lot of overlap in explaining the variation among the remaining independent variables.

Tolerance

Tolerance is just 1- R_x^2, the denominator of the variance inflation factor.

Eigenvalues of Correlations

```
Eigenvalues of Correlations
                                      Incremental        Cumulative         Condition
No.               Eigenvalue              Percent           Percent            Number
1                   1.994969                66.50             66.50              1.00
2                   1.004003                33.47             99.97              1.99
3                   0.001027                 0.03            100.00           1941.85
Some Condition Numbers greater than 1000. Multicollinearity is a SEVERE problem.
```

This section gives an eigenvalue analysis of the independent variables after they have been centered and scaled. Notice that in this example, the third eigenvalue is very small.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero indicate a multicollinearity problem in your data.

Incremental Percent

Incremental percent is the percent this eigenvalue is of the total. In an ideal situation, these percentages would be equal. Percents near zero indicate a multicollinearity problem in your data.

Cumulative Percent

This is the running total of the Incremental Percent.

Condition Number

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variances. Condition numbers greater than 1000 indicate a severe multicollinearity problem while condition numbers between 100 and 1000 indicate a mild multicollinearity problem.

Eigenvector of Correlations

```
Eigenvector of Correlations
No.               Eigenvalue                   X1                X2                X3
1                   1.994969             0.701391          0.707741          0.084573
2                   1.004003            -0.134162          0.014553          0.990853
3                   0.001027             0.700036         -0.706322          0.105159
```

This report displays the eigenvectors associated with each eigenvalue. The notion behind eigenvalue analysis is that the axes are rotated from those defined by the variables to a new set defined by the variances of the variables. Rotation is accomplished by taking weighted averages of the standardized original variables. The first new variable is constructed to account for the largest amount of variance possible from a single axis.

No.

The number of the eigenvalue.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero indicate multicollinearity in your data. The eigenvalues represent the spread (variance) in the direction defined by this new axis. Hence, small eigenvalues indicate directions in which there is no spread. Since regression analysis seeks to find trends across values, when there is not a spread, the trends cannot be computed accurately.

Table-Values

The table values give the eigenvectors. The eigenvectors give the weights that are used to create the new axis. By studying the weights, you can gain an understanding of what is happening in the data.

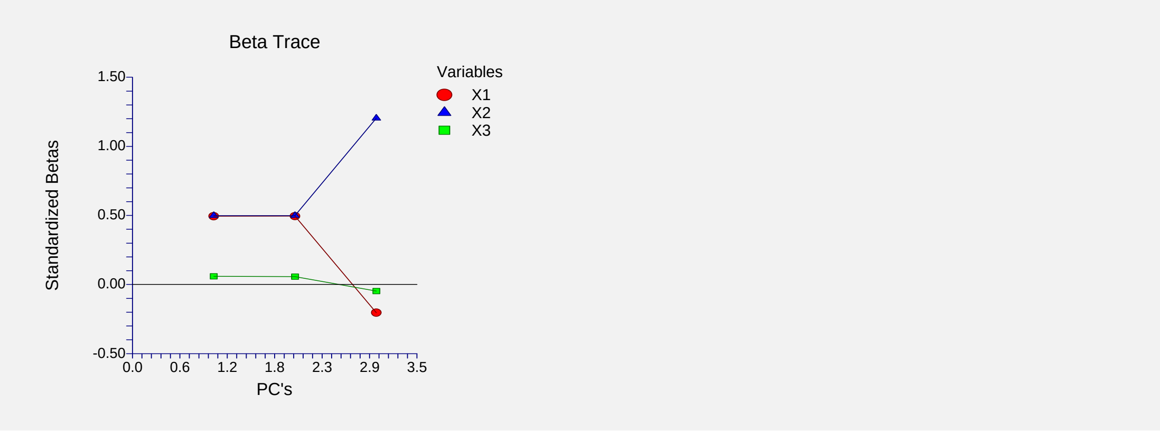
In the example above, we can see that the first factor (new variable associated with the first eigenvalue) is constructed by adding X1 and X2. Note that the weights are almost equal. X3 has a small weight, indicating that it does not play a role in this factor.

Factor 2 seems to be completely created from X3. X1 and X2 play only a small role in its construction.

Factor 3 seems to be the difference between X1 and X2. Again X3 plays only a small role. Hence, the interpretation of these eigenvectors leads to the following statements:

1. Most of the variation in X1, X2, and X3 can be accounted for by considering only two variables: Z = X1+X2 and X3.
2. The third dimension, calculated as X1-X2, is almost negligible and might be ignored.

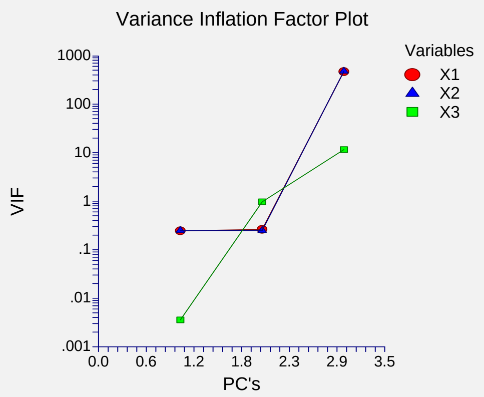
Beta Trace Section



This plot shows the standardized regression coefficients (often referred to as the betas) on the vertical axis and the number of principal components (PC's) included along the horizontal axis. Thus, the set on the right is the least squares set.

By studying this plot, you can determine what omitting a certain number of PC's has done to the estimated regression coefficients.

Variance Inflation Factor Plot



This is a plot that shows the effect of the omitted PC's on the variance inflation factors. Since the major goal of PC regression is to remove the impact of multicollinearity, it is important to know at what point multicollinearity has been dealt with. This plot shows this.

Since the rule-of-thumb is that multicollinearity is not a problem once all VIFs are less than 10, we inspect the graph for this point. In this example, it appears that all VIFs are less than 10 if only two of the three PC's are included.

Standardized Regression Coefficients Section

```
Standardized Regression Coefficients Section
PC's                X1            X2            X3
1               0.4942        0.4987        0.0596
2               0.4945        0.4987        0.0574
3              -0.2034        1.2029       -0.0475
```

This report gives the values that are plotted on the beta trace.

Variance Inflation Factor Section

```
Variance Inflation Factor Section
PC's                X1            X2            X3
1               0.2466        0.2511        0.0036
2               0.2645        0.2513        0.9815
3             477.2665      485.8581       11.7455
```

This report gives the values that are plotted on the variance inflation factor plot. Note how easy it is to determine when all three VIFs are less than 10.

Components Analysis Section

```
Components Analysis Section
PC's          R2          Sigma        B'B         Ave VIF      Max VIF
1             0.9905      1.1677       0.4965       0.1671      0.2511
2             0.9905      1.1674       0.4965       0.4991      0.9815
3             0.9915      1.1028       1.4905     324.9567      485.8581
```

This report provides a quick summary of the various statistics that might go into the choice of *k*.

PC's

This is the number of principal components included in the regression reported on this line.

R2

This is the value of R-squared. Since the least squares solution maximizes R-squared, the largest value of R-squared occurs at bottom of the report (when all PC's are included).

Sigma

This is the square root of the mean squared error. Least squares minimizes this value, so we want to select the number of PC's that does not stray very much from the least squares value.

B'B

This is the sum of the squared standardized regression coefficients. PC regression assumes that this value is too large and so the method tries to reduce this. We want to find the number of PC's at which this value has stabilized.

Ave VIF

This is the average of the variance inflation factors.

Max VIF

This is the maximum variance inflation factor. Since we are looking for the number of PC's which results in all VIFs being less than 10, this value is very helpful.

P.C. versus L.S. Comparison Section

```
P.C. vs. Least Squares Comparison Section with 1 Component Omitted
             Regular       Regular        Stand'zed   Stand'zed      Component        L.S.
Independent  Component     L.S.           Component        L.S.       Standard    Standard
Variable     Coeff's       Coeff's          Coeff's     Coeff's          Error       Error
Intercept    0.763326      0.2230599
X1           1.007698      -0.4144863        0.4945     -0.2034      0.0272776    1.094502
X2           1.003778      2.421286          0.4987      1.2029   2.626337E-02    1.090883
X3           0.568248      -0.4703622        0.0574     -0.0475      0.2554352   0.8347205

R-Squared    0.9905        0.9915
Sigma        1.1674        1.1028
```

This report provides a detailed comparison between the PC regression solution and the ordinary least squares solution to the estimation of the regression coefficients.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of b_0.

Regular Component (and L.S.) Coeff's

These are the estimated values of the regression coefficients b_0, b_1, ..., b_p. The first column gives the values for PC regression and the second column gives the values for regular least squares regression.

The value indicates how much change in *Y* occurs for a one-unit change in *X* when the remaining *X's* are held constant. These coefficients are also called partial-regression coefficients since the effect of the other *X's* is removed.

Stand'zed Component (and L.S.) Coeff's

These are the estimated values of the standardized regression coefficients. The first column gives the values for PC regression and the second column gives the values for regular least squares regression.

Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is:

$$b_{j,\,std} = b_j \left(\frac{s_y}{s_{x_j}} \right)$$

where s_y and s_{x_j} are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

Component (and L.S.) Standard Error

These are the estimated standard errors (precision) of the regression coefficients. The first column gives the values for PC regression and the second column gives the values for regular least squares regression.

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate.

Since one of the objects of PC regression is to reduce this (make the estimates more precise), it is of interest to see how much reduction has taken place.

R-Squared

R-squared is the coefficient of determination. It represents the percent of variation in the dependent variable explained by the independent variables in the model. The R-squared values of both the PC and regular regressions are shown.

Sigma

This is the square root of the mean squared error. It provides a measure of the standard deviation of the residuals from the regression model.

It represents the percent of variation in the dependent variable explained by the independent variables in the model. The R-squared values of both the PC and regular regressions are shown.

PC Coefficient Section

```
PC Coefficient Section
Principal            PC    Individual
Component   Coefficient     R-Squared       Eigenvalue
PC1              7.6653        0.9905         1.994969
PC2             -0.0245        0.0000         1.004003
PC3            -10.8457        0.0010         0.001027
```

This report provides the details of the regression based on the principal components (the Z's).

Principal Component

This is the number of the principal component being reported about on this line. The order here corresponds to the order of the eigenvalues. Thus, the first is associated with the largest eigenvalue and the last is associated with the smallest.

PC Coefficient

These are the estimated values of the regression coefficients a_1, ..., a_p. The value indicates how much change in Y occurs for a one-unit change in z when the remaining z's are held constant.

Individual R-Squared

This is the amount contributed to R-squared by this component.

Eigenvalue

This is the eigenvalue of this component.

PC Regression Coefficient Section

```
Regression Coefficient Section with 1 Component Omitted
                                                 Stand'zed
Independent        Regression    Standard       Regression
Variable           Coefficient   Error          Coefficient          VIF
Intercept          0.763326
X1                 1.007698      0.0272776           0.4945       0.2645
X2                 1.003778      2.626337E-02        0.4987       0.2513
X3                 0.568248      0.2554352           0.0574       0.9815
```

This report provides the details of the PC regression solution.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of b_0.

Regression Coefficient

These are the estimated values of the regression coefficients b_0, b_1, ..., b_p. The value indicates how much change in Y occurs for a one-unit change in x when the remaining X's are held constant. These coefficients are also called partial-regression coefficients since the effect of the other X's is removed.

Standard Error

These are the estimated standard errors (precision) of the PC regression coefficients. The standard error of the regression coefficient, s_{b_j} , is the standard deviation of the estimate. In regular regression, we divide the coefficient by the standard error to obtain a t statistic. However, this is not possible here because of the bias in the estimates.

Stand’zed Regression Coefficient

These are the estimated values of the standardized regression coefficients. Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When there are vastly different units involved for the variables, this is a way of making comparisons between variables. The formula for the standardized regression coefficient is:

$$b_{j,\,std} = b_j\left(\frac{s_y}{s_{x_j}}\right)$$

where s_y and s_{x_j} are the standard deviations for the dependent variable and the corresponding j^{th} independent variable.

VIF

These are the values of the variance inflation factors associated with the variables. When multicollinearity has been conquered, these values will all be less than 10. Details of what VIF is were given earlier.

Analysis of Variance Section

```
Analysis of Variance Section with 1 Component Omitted
                                      Sum of     Mean        Prob
Source           DF       Squares     Square     F-Ratio     Level
Intercept        1        9614.223    9614.223
Model            3        1992.698    664.2327   487.3907    0.000000
Error            14       19.07968    1.362834
Total(Adjusted)  17       2011.778    118.3399

Mean of Dependent         23.11111
Root Mean Square Error    1.167405
R-Squared                 0.9905
Coefficient of Variation  5.051271E-02
```

An analysis of variance (ANOVA) table summarizes the information related to the sources of variation in the data.

Source

This represents the partitions of the variation in *y*. There are four sources of variation listed: intercept, model, error, and total (adjusted for the mean).

DF

The degrees of freedom are the number of dimensions associated with this term. Note that each observation can be interpreted as a dimension in *n*-dimensional space. The degrees of freedom for the intercept, model, error, and adjusted total are 1, *p*, *n*-*p*-1, and *n*-1, respectively.

Sum of Squares

These are the sums of squares associated with the corresponding sources of variation. Note that these values are in terms of the dependent variable, *y*. The formulas for each are:

Mean Square

The mean square is the sum of squares divided by the degrees of freedom. This mean square is an estimated variance. For example, the mean square error is the estimated variance of the residuals (the residuals are sometimes called the *errors*).

F-Ratio

This is the F statistic for testing the null hypothesis that all $\beta_j = 0$. This F-statistic has p degrees of freedom for the numerator variance and n-p-1 degrees of freedom for the denominator variance.

Since PC regression produces biased estimates, this F-Ratio is not a valid test. It serves as an index, but it would not stand up under close scrutiny.

Prob Level

This is the p-value for the above F test. The p-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the p-value is less than α, say 0.05, the null hypothesis is rejected. If the p-value is greater than α, then the null hypothesis is accepted.

Root Mean Square Error

This is the square root of the mean square error. It is an estimate of σ, the standard deviation of the e_i*'s*.

Mean of Dependent Variable

This is the arithmetic mean of the dependent variable.

R-Squared

This is the coefficient of determination. It is defined in full in the Multiple Regression chapter.

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, computed by dividing root mean square error by the mean of the dependent variable. By itself, it has little value, but it can be useful in comparative studies.

$$CV = \frac{\sqrt{MSE}}{\bar{y}}$$

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section with 1 Component Omitted
Row     Actual        Predicted       Residual
1       3             4.346828        -1.346828
2       9             7.930331        1.069669
3       11            12.08208        -1.082081
4       15            13.52531        1.47469
5       13            13.96476        -0.9647598
6       13            14.40421        -1.40421
7       17            16.41569        0.5843138
8       21            19.99919        1.000811
9       25            24.15094        0.849061
10      27            25.59417        1.405833
11      25            26.03362        -1.033618
12      27            26.47307        0.5269322
13      29            28.48454        0.5154559
14      33            32.06805        0.9319535
15      35            36.2198         -1.219797
16      37            37.66302        -0.6630252
17      37            38.10247        -1.102475
18      39            38.54193        0.4580744
```

This section reports the predicted values and the sample residuals, or e_i’s. When you want to generate predicted values for individuals not in your sample, add their values to the bottom of your database, leaving the dependent variable blank. Their predicted values will be shown on this report.

Actual

This is the actual value of *Y* for the i^{th} row.

Predicted

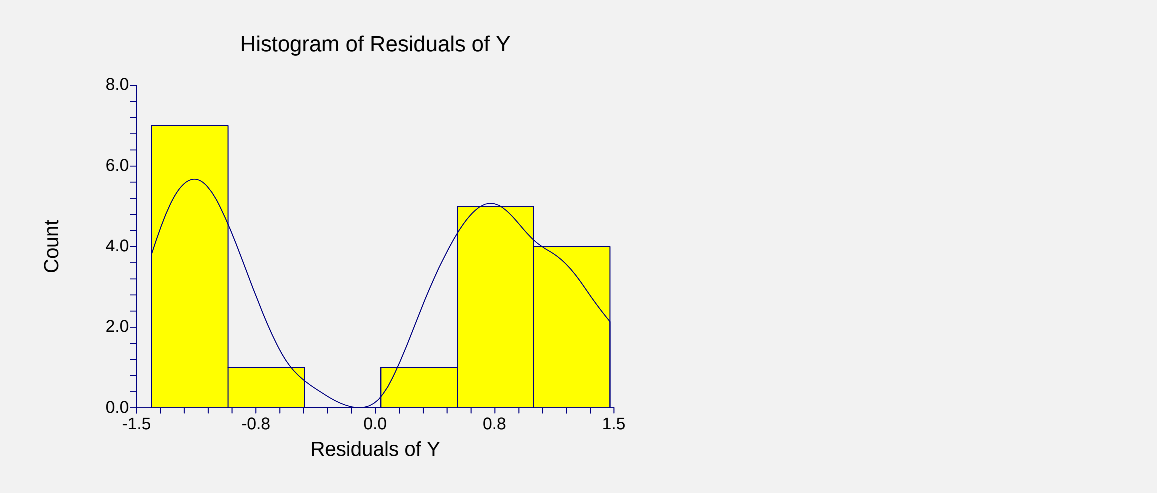
The predicted value of *Y* for the i^{th} row. It is predicted using the levels of the *X’s* for this row.

Residual

This is the estimated value of e_i. This is equal to the *Actual* minus the *Predicted.*

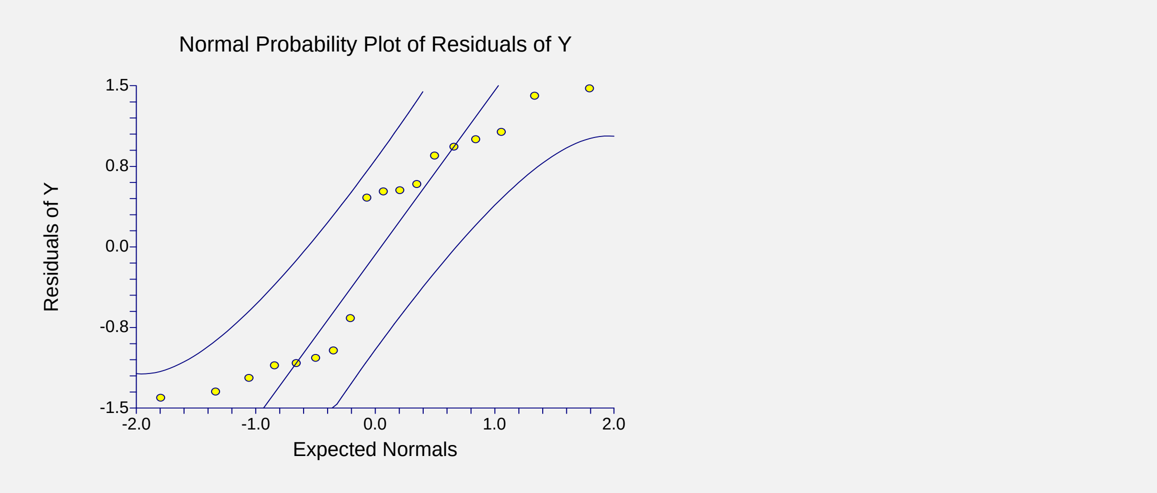
Histogram

The purpose of the histogram and density trace of the residuals is to display the distribution of the residuals.



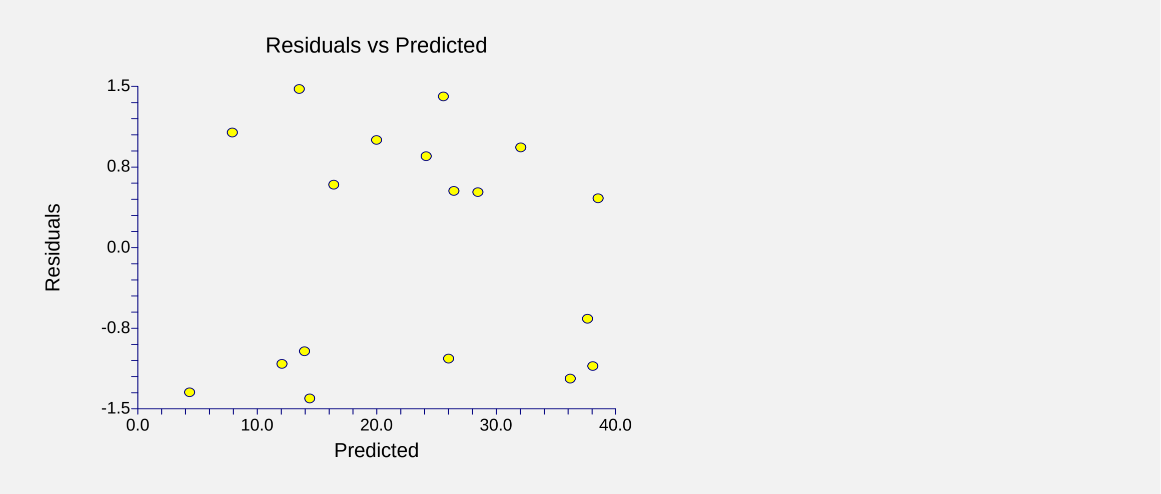
The odd shape of this histogram occurs because of the way in which these particular data were manufactured.

Probability Plot of Residuals



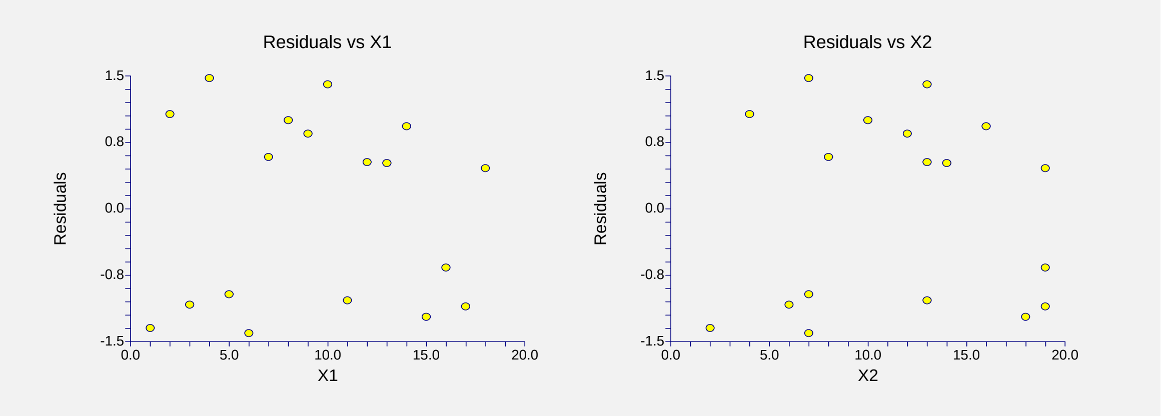
Residual vs Predicted Plot

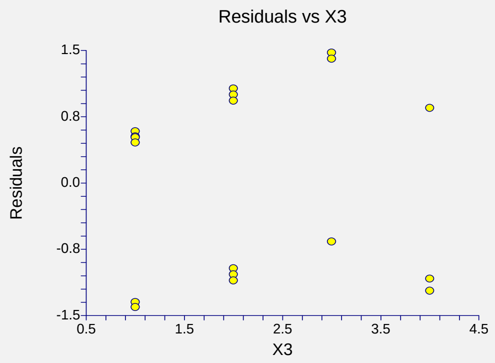
This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance, a violation of a critical regression assumption. A sloping or curved band signifies inadequate specification of the model. A sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.



Residual vs Predictor(s) Plot

This is a scatter plot of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.




Residuals vs X3
Residuals
1.5
0.8
0.0
-0.8
-1.5
0.5
1.5
2.5
3.5
4.5
X3

Chapter 345

Nondetects Regression

Introduction

This module fits the regression relationship between a positive-valued dependent variable (with, possibly, some nondetected responses) and one or more independent variables. The distribution of the residuals (errors) is assumed to follow the exponential, extreme value, logistic, log-logistic, lognormal, lognormal10, normal, or Weibull distribution. The Distribution Fitting module may be useful for determining a suitable distribution for use in Nondetects Regression.

Nondetects analysis is the analysis of data in which one or more of the values cannot be measured exactly because they fall below one or more detection limits. Detection limits often arise in environmental studies because of the inability of instruments to measure small concentrations. Some examples of sampling scenarios that lead to datasets with nondetects values are finding pesticide concentrations in water, determining chemical composition of soils, or establishing the number of particulates of a compound in the air.

A common practice for dealing with values which fall below the detection threshold is substitution. Often, each value which is below the detection limit is substituted with one half the detection limit. Evaluation of relationships among variables are then carried out using standard techniques (multiple regression) with the substituted data. Helsel (2005) warns of the potential data analysis biases that result if nondetects values are substituted. He particularly warns about the arbitrariness of substituting one half the detection limit (or zero, or the detection limit). Alternatively, if a proper distribution can be assumed for the variable with nondetects values, maximum likelihood distribution regression is a more appropriate analog to multiple regression with substituted values.

For a complete account of nondetects analysis, we suggest the book by Helsel (2005).

Technical Details

The linear regression equation is

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + Se$$

Here, *S* represents the value of a constant standard deviation, *Y* is the response or a transformation of the response (*ln()* or *log()*), the *X’s* are one or more independent variables, the *B*’s are the regression coefficients, and *e* is the residual (error) that is assumed to follow a particular probability distribution. The problem reduces to estimating the *B*’s and *S*. The density functions of the eight distributions that are fit by this module are given in the Distribution Fitting section and will not be repeated here.

As an example, we give detailed results for the lognormal distribution. The results for other distributions follow a similar pattern.

The lognormal probability density function may be written as

$$f(y|M,S)=\frac{1}{yS\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(y)-M}{S}\right)^2}$$

If we replace the location parameter, *M*, with the regression model, the density now becomes

$$f(y|B_0\cdots B_p,S)=\frac{1}{yS\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\ln(y)-\sum_{i=0}^{p}B_iX_i}{S}\right)^2\right\}$$

Maximum likelihood estimation consists of finding the values of the distribution parameters that maximize the log-likelihood of the data values. Loosely speaking, these are the values of the parameters, which maximize the probability that the current set of data values occur.

NCSS employs the Newton-Raphson algorithm with numerical differentiation to obtain the maximum likelihood estimates. These estimates have been shown to have optimality characteristics in large samples (number of responses greater than 20).

Data Structure

Nondetects responses are specified using up to three components: the response value (e.g., concentration or amount), an optional indicator of whether or not each observation was detected, and an optional frequency (count) specification. If no detection indicator is included, all response values represent detected responses. If the frequency (count) variable is omitted, all counts are assumed to be one.

Any number of independent variables may be specified as separate columns. In Nondetects Distribution Regression, all independent variables must be numeric. If categorical variables are to be used, corresponding zero-one variables must first be created.

Sample Dataset

The table below shows a dataset (fictitious) reporting 1,3-dichloropropene (1,3-DCP) concentrations (in μg/L) for 53 randomly chosen soil locations. Concentrations were determined following addition of one of two solutions to each sample: water or NaHSO4. Some of the soil samples resulted in concentrations below the laboratory minimum reporting limit of 0.13μg/L. The percent moisture in the soil sample is also reported. A value of zero in the DNondet column indicates 1,3-DCP was detected. A value of one in the DNondet column indicates 1,3-DCP was not detected. The Solution column is repeated with an appropriate zero-one variable column. These data are contained in the DCP dataset.

DCP dataset (subset)

DCP	DNondet	Moisture	Solution	Solution2
0.17	0	8.14	water	0
0.25	0	6.23	water	0
0.22	0	4.56	NaHSO4	1
0.28	0	7.39	water	0
0.13	1	11.91	water	0
0.18	0	6.43	NaHSO4	1
0.13	1	6.97	water	0
0.18	0	5.48	NaHSO4	1
0.26	0	6.12	NaHSO4	1
0.13	1	5.42	NaHSO4	1

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the probability distribution that is fit and the variables used in the analysis.

Response Variable

Response Variable

The values of this variable represent either the magnitude of a detected observations or detection limits, depending on the corresponding values of the Nondetection (Censor) Variable.

The values in this variable must be greater than zero. If the value is missing or non-positive, it is not used during the estimation phase.

Frequency Variable

Frequency Variable

This variable gives the count, or frequency, of the response displayed on that row. When omitted, each row receives a frequency of one. Frequency values should be positive integers. A frequency variable is often used to indicate the number of Nondetects.

Nondetection Variable

Nondetection (Censor) Variable

The values in this variable indicate whether the value of the Response Variable represents a nondetected (censored) observation or a detected observation. When a particular value of this variable indicates a Nondetect, the corresponding value of the Response Variable represents a lower detection limit.

These values may be text or numeric. The interpretation of these codes is specified by the 'Detected' and 'Not Detected' (Censored) options to the right of this option.

Only two values are used, the Detected value and the Not Detected value. The Unknown Censor option specifies what is to be done with values that do not match either the Detected value or the Not Detected value.

Rows with missing values (blanks) in this variable are omitted from the estimation phase, but results are shown in any reports that output predicted values.

Detected

When this value is encountered under the Nondetection (Censor) Variable it indicates that the value under the Response Variable was observed or detected. The value may be a number or a letter.

We suggest the letter 'D' or the number '0' when you are in doubt as to what to use.

A detected observation is one in which the value was measured exactly; for example, the concentration was such that the instrument was able to measure it.

Not Detected

When this value is encountered under the Nondetection (Censor) Variable it indicates that the value under the Response Variable was not actually observed (i.e., a nondetect) but represents a lower detection limit. That is, the observation is left-censored, and the actual value of the response is something below the detection limit.

The value may be a number or a letter. We suggest the letter 'N' or the number '1' when you are in doubt as to what to use.

A nondetect is a response in which the value was not measured exactly; for example, the concentration was such that the instrument was not able to measure it.

Probability Distribution

Distribution

This option specifies the probability distribution of the residuals (errors). All results are based on the probability distribution specified here.

Alpha Level

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Independent Variables

X’s: Independent Variables

Specify the independent variables. At least one independent variable must be specified here.

These variables may be thought of as additional variables for which statistical adjustment is desired. Discrete and/or continuous variables may be specified here. If discrete variables are to be specified, you should create and specify the appropriate number of indicator (dummy) variables. For example, if three groups are to be compared, two indicator variables will be needed to distinguish these groups.

Estimation Tab

The following options control the searching algorithms used during parameter estimation.

Estimation Options

Maximum Iterations

Many of the parameter estimation algorithms are iterative. This option assigns a maximum to the number of iterations used in any one algorithm. We suggest a value of at least 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained. If the number of iterations reaches this amount, you should re-run your analysis with a larger value.

Minimum Relative Change

This value is used to control the iterative algorithms used in maximum likelihood estimation. When the relative change in all of the parameters is less than this amount, the iterative procedure is terminated.

Parameter Adjustment

Newton's method calculates a change for each parameter value at each step. Instead of taking the whole parameter change, this option lets you take only a fraction of the indicated change. For datasets that diverge, taking only partial steps may allow the algorithm to converge. In essence, the algorithm tends to over correct the parameter values. This factor allows you to dampen this over correction. We suggest a value of about 0.2. This may increase the number of iterations (and you will have to increase the Maximum Iterations accordingly), but it provides a greater likelihood that the algorithm will converge.

Starting Sigma

Specify a starting value for *S*, the standard deviation of the residuals (errors*)*. Select '0 - Data' to calculate an appropriate value from the data. If convergence fails, try a different value.

Derivatives

This value specifies the machine precision value used in calculating numerical derivatives. Slight adjustments to this value can change the accuracy of the numerical derivatives (which impacts the variance/covariance matrix estimation).

Remember from calculus that the derivative is the slope calculated at a point along the function. It is the limit found by calculating the slope between two points on the function curve that are very close together. Numerical differentiation mimics this limit by calculating the slope between two function points that are very close together and then computing the slope. This value controls how close together these two function points are.

Numerical analysis suggests that this distance should be proportional to the machine precision of the computer. We have found that our algorithm achieves four-place accuracy in the variance-covariance matrix no matter what value is selected here (within reason). However, increasing or decreasing this value by two orders of magnitude may achieve six or seven place accuracy in the variance-covariance matrix. We have found no way to find the optimal value except trial and error.

Note that the parameter estimates do not seem to be influenced a great deal, only their standard errors.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report ... Residual Report

Each of these options specifies whether the indicated report is calculated and displayed.

Select Plots

X - Y Plots ... X - Residual Plots

Each of these options specifies whether the indicated plot is displayed.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also, note that all calculations are performed in double precision regardless of which option you select here. Single precision is for reporting purposes only.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, only value labels, or both for values of the group variable. Use this option if you want to automatically attach labels to the values of the group variable (such as 1=Male, 2=Female, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Decimal Places

Response and Probability Decimals

This option specifies the number of decimal places shown on reported response and probability values.

Plot Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. A {G} is replaced by the appropriate legend name.

X - Y Plots to X - Resid Plots Tabs

These options control the attributes of the plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Number Predicted

This options sets resolution of the plot along the horizontal axis. A value near 50 is usually adequate.

Titles

Plot Title

This option contains the text of the plot title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Symbols Tab

These options specify the attributes of the symbols used in the plots.

Plotting Symbols

Detected ... Predicted

This option specifies the symbol used for each type of data, censored, failed, and predicted. These symbols are provided to allow the various censoring types to be identified, even on black and white printers.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Nondetects Regression

This section presents an example of how to perform a nondetects normal distribution regression. The DCP dataset that will be used was described above. Suppose the researchers wish to establish the relationship between percent moisture in the soil sample and 1,3-DCP concentration. Further, they wish to determine if there are differences in the two solutions used for determining 1,3-DCP concentrations.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Nondetects Regression window.

1 Open the DCP dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **DCP.S0**.
- Click **Open**.

2 Open the Nondetects Regression window.

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines**, then **Nondetects Regression**. The Nondetects Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Nondetects Regression window, select the **Variables tab**.
- Double-click in the **Response Variable** box. This will bring up the variable selection window.
- Select **DCP** from the list of variables and then click **Ok**.
- Double-click in the **Nondetection (Censor) Variable** box. This will bring up the variable selection window.
- Select **DNondet** from the list of variables and then click **Ok**.
- Enter the values **0** and **1** for the **Detected** and **Not Detected** fields, respectively.
- Double-click in the **X’s: Independent Variables** box. This will bring up the variable selection window.
- Select **Moisture** and **Solution2** (you can use the control key) from the list of variables and then click **Ok**.
- Set the **Distribution** to **Normal**.

4 Specify the reports.

- On the Nondetects Regression window, select the **Reports tab**.
- Check the boxes of all the report options.

5 Adjust the axes.

- Select the **X - Y Plots tab**.
- Under Vertical (Y) Axis, click on **Tick Label Settings**.
- Change **Decimals** to **2**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Data Summary Section

```
Data Summary Section

Type of                                                    Hours            Hours
Observation                Rows          Count         Minimum          Maximum
Missing or Prediction      0
Detected                   39            39            0.140            0.350
Not Detected               13            13            0.130            0.130
Total (Nonmissing)         52            52            0.130            0.350


Means
Variable                Mean
DCP                     0.2071154
Moisture                7.520385
Solution2               0.5769231
```

This report displays a summary of the data that were analyzed. Scan this report to determine if there are any obvious data-entry errors by double-checking the counts and the minimum and maximum.

The means given for each variable are for detected and nondetected rows combined.

Parameter Estimation Section

```
Maximum Likelihood Parameter Estimation Section

Parameter           Parameter         Standard                        Prob            Lower           Upper
Name                Estimate          Error           Z Value         Level           95.0% C.L.      95.0% C.L.
Intercept           0.1821519         3.239192E-02    5.6234          0.0000          0.1186649       0.2456388
Moisture            -2.174437E-03     3.457449E-03    -0.6289         0.5294          -8.950911E-03   4.602039E-03
Solution2           5.185126E-02      2.305901E-02    2.2486          0.0245          6.656426E-03    9.704609E-02
Sigma               7.899636E-02      9.540465E-03    8.2801          0.0000          6.234572E-02    0.1000939

Approximate R-Squared
Log Likelihood      30.68732
Iterations          32
```

This report displays parameter estimates along with standard errors, significance tests, and confidence limits. Note that the significance levels and confidence limits all use large sample formulas. We suggest that you only use these results when the number of detected items is greater than twenty.

Parameter Estimates

These are the maximum likelihood estimates (MLE) of the parameters. They are the estimates that maximize the likelihood function. Details are found in Nelson (1990) pages 287 - 295.

Standard Error

The standard errors are the square roots of the diagonal elements of the estimated Variance Covariance matrix.

Z Value

The z value is equal to the parameter estimate divided by the estimated standard error. This ratio, for large samples, follows the normal distribution. It is used to test the hypothesis that the parameter value is zero. This value corresponds to the t value that is used in multiple regression.

Prob Level

This is the two-tailed p-value for testing the significance of the corresponding parameter. You would deem independent variables with small p-values (less than 0.05) important in the regression equation.

Upper and Lower 100(1-Alpha)% Confidence Limits

These are the lower and upper confidence limits for the corresponding parameters. They are large sample limits. They should be ignored when the number of detected items is less than thirty. For the regression coefficients *B*, the formulas are

$$CL_i = \hat{B}_i \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{B}_i} \qquad i = 0,\cdots,p$$

where $\hat{B}_i$ is the estimated regression coefficient, $\hat{\sigma}_{\hat{B}_i}$ is its standard error, and *z* is found from tables of the standard normal distribution.

For the estimate of sigma, the formula is

$$CL = \hat{S}\exp\left\{\frac{\pm z_{1-\alpha/2}\hat{\sigma}_{\hat{S}}}{\hat{S}}\right\}$$

Approximate R-Squared

R-Squared reflects the percent of variation in response explained by the independent variables in the model. A value near zero indicates a complete lack of fit, while a value near one indicates nearly a perfect fit.

This value is an 'approximate' R-squared because it is computed using the failed observations with regression coefficients which were based on all observations. The formula used is

$$R^2 = 1 - \frac{\sum_{k=1}^{n}\delta_k\left(y_k - \sum_{i=0}^{p}X_{ik}\hat{B}_i\right)^2}{\sum_{k=1}^{n}\delta_k\left(y_k - \bar{y}\right)^2}, \quad \bar{y} = \frac{\sum_{k=1}^{n}\delta_k y_k}{\sum_{k=1}^{n}\delta_k}$$

where δ_i is one if the observation was a failure, and zero otherwise. Approximate R-Squared values greater than one or less than zero are not reported.

Log Likelihood

This is the value of the log likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Iterations

This is the number of iterations that were required to solve the likelihood equations. If this is greater than the maximum you specified, you will receive a warning message. You should then increase the Maximum Iterations and rerun the analysis.

Variance Covariance Matrix

```
Variance Covariance Matrix

                Intercept        Moisture         Solution2        Sigma
Intercept       1.049237E-03     -9.388014E-05    -3.744217E-04    -9.745229E-06
Moisture        -9.388014E-05    1.195395E-05     8.386759E-06     -1.413499E-06
Solution2       -3.744217E-04    8.386759E-06     5.317181E-04     4.005026E-06
Sigma           -9.745229E-06    -1.413499E-06    4.005026E-06     9.102048E-05
```

This table gives an estimate of the asymptotic variance covariance matrix which is the inverse of the Fisher information matrix. The elements of the Fisher information matrix are calculated using numerical differentiation.

Residual Section

```
Residual Section
          (T)                   Predicted    Raw              Standardized     Cox-Snell
Row       DCP       T           T            Residual         Residual         Residual
1         0.170     0.17        0.1644519    5.548064E-03     7.023189E-02     0.7507667
2         0.250     0.25        0.1686051    8.139489E-02     1.030363         1.887698
3         0.220     0.22        0.2240877    -4.08768E-03     -5.174517E-02    0.6527078
4         0.280     0.28        0.1660828    0.1139172        1.442057         2.595036
5L        0.130     0.13        0.1562543    -2.625431E-02    -0.3323483       0.4617389
6         0.180     0.18        0.2200215    -4.002148E-02    -0.5066244       0.3655848
7L        0.130     0.13        0.166996     -3.699603E-02    -0.4683257       0.3853329
8         0.180     0.18        0.2220872    -0.0420872       -0.5327739       0.3525336
9         0.260     0.26        0.2206956    3.930444E-02     0.4975475        1.173115
10L       0.130     0.13        0.2222177    -9.221766E-02    -1.167366        0.129575
11        0.170     0.17        0.1652347    4.765267E-03     6.032261E-02     0.7424439
12        0.330     0.33        0.220065     0.109935         1.391647         2.500857
.         .         .           .            .                .                .
.         .         .           .            .                .                .
.         .         .           .            .                .                .
```

This report displays the predicted value and residual for each row. The report provides predicted values for all rows with values for the independent variables. Hence, you can add rows of data with missing time values to the bottom of your database and obtain the predicted values for them from this report. The report also allows you to obtain predicted values for nondetects observations.

You should ignore the residuals for nondetects observations, since the residual is calculated as if the response value were an actual response.

Row

This is the number of the observation being reported on. Nondetects observations have a letter (L for left-censored) appended to the row number.

(T) Response

This is the original value of the dependent variable.

Predicted T

This is the predicted transformed value of the dependent variable (usually time). Note that *y* depends on the distribution being fit. For the Weibull, exponential, lognormal, and log-logistic distributions, the *y* is *ln(t)*. For the lognormal10 distribution, *y* is *log(t)*. For the extreme value, normal, and logistic distributions, *y* is *t*. The formula for *y* is

$$\hat{y} = \sum_{i=0}^{p} x_i B_i$$

Raw Residual

This is the residual in the *y* scale. The formula is

$$r_k = y_k - \sum_{i=0}^{p} x_i B_i$$

Note that the residuals of censored observations are not directly interpretable, since there is no obvious value of y. The row is displayed so that you can see the predicted value for this censored observation.

Standardized Residual

This is the residual standardized by dividing by the standard deviation. The formula is

$$r_k' = \frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}$$

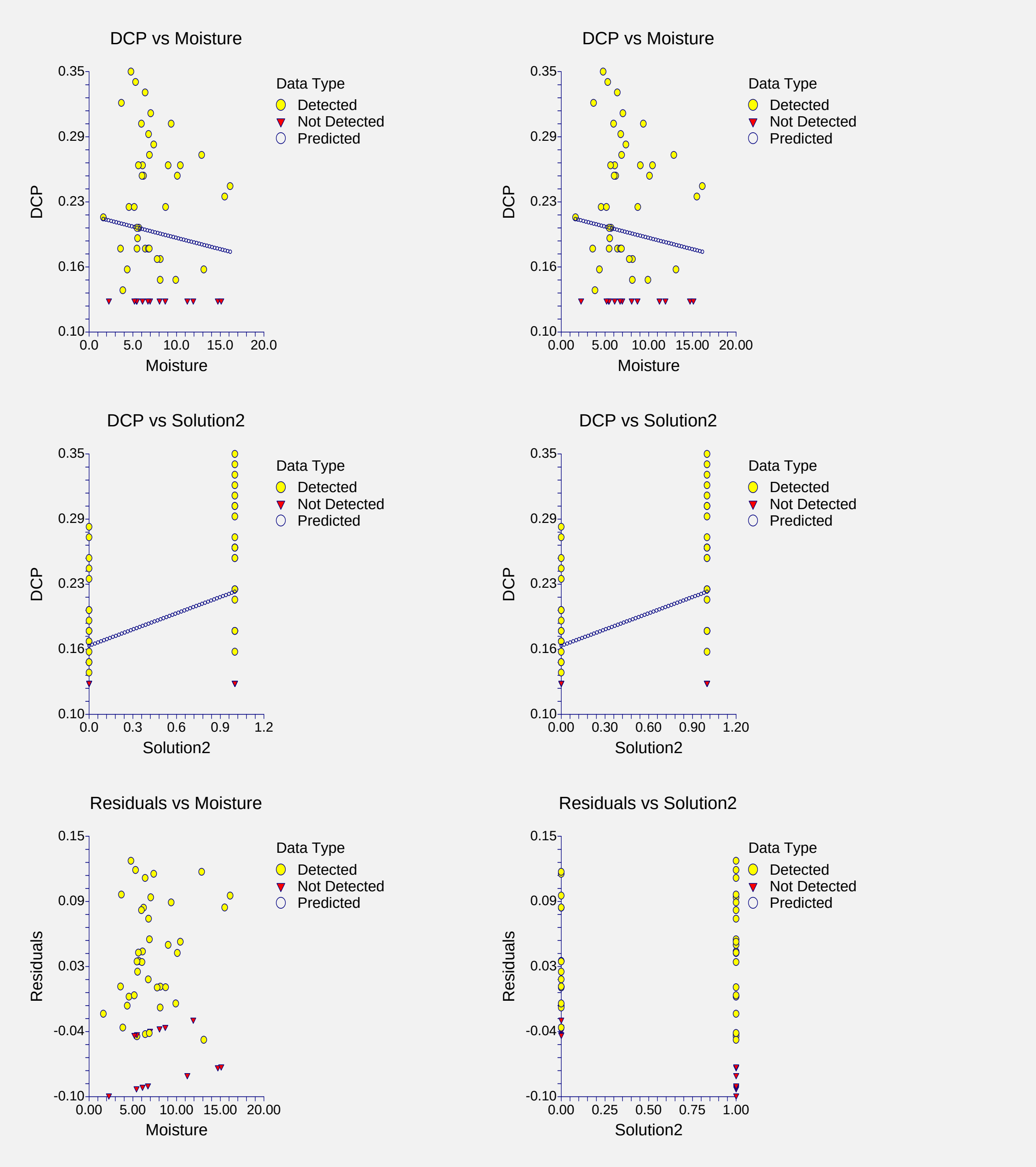
Cox-Snell Residual

The Cox-Snell residual is defined as

$$r_k'' = -\log\left\{1 - F\left(\frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}\right)\right\}$$

Here again, the residual does not have a direct interpretation for censored values.

X-Y, X-Trans(Y), and X-Resid Plots



The first two pairs of plots show the data values from which the analysis was run. The plots on the left show the response versus the independent variable in the original scale. The plots on the right show the response versus the independent variable in the transformed metric (for the normal distribution there is no transformation, so that the plots on the left and right are the same). The third pair of plots shows the residuals in the transformed scale (again, here, there is no transformation because the normal distribution is used).

Example 2 – Validation using Helsel (2005)

On pages 134-138, Helsel (2005) presents an example of using nondetects lognormal distribution regression to compare zinc concentrations among two zones. The estimate of the zone effect is given as -0.257408. The corresponding Z value and probability level are -1.60 and 0.110, respectively. The Log-likelihood is -407.296. The data are contained in the ZINC.S0 dataset.

These data can be run in this procedure to see that ***NCSS*** gets the same results. You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Nondetects Regression window.

1 **Open the ZINC dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **ZINC.S0**.
- Click **Open**.

2 **Open the Nondetects Regression window.**

- On the menus, select **Analysis**, then **Regression / Correlation**, then **Other Regression Routines**, then **Nondetects Regression**. The Nondetects Regression procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

- On the Nondetects Regression window, select the **Variables tab**.
- Double-click in the **Response Variable** box. This will bring up the variable selection window.
- Select **Zinc** from the list of variables and then click **Ok**.
- Double-click in the **Nondetection (Censor) Variable** box. This will bring up the variable selection window.
- Select **ZNondet** from the list of variables and then click **Ok**.
- Enter the values **0** and **1** for the **Detected** and **Not Detected** fields, respectively.
- Double-click in the **X's: Independent Variable** box. This will bring up the variable selection window.
- Select **Zone** from the list of variables and then click **Ok**. Note that the values of Zone are appropriate for this problem.
- Set the **Distribution** to **Lognormal**.

4 **Specify the estimation parameters.**

- On the Nondetects Regression window, select the **Estimation tab**.
- Set **Derivatives** to **0.0005**.

5 **Specify the reports.**

- On the Nondetects Regression window, select the **Reports tab**.
- Uncheck all boxes except **Parameter Report**.

6 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Parameter Estimation Section

```
Maximum Likelihood Parameter Estimation Section

Parameter          Parameter       Standard                      Prob          Lower           Upper
Name               Estimate        Error            Z Value      Level         95.0% C.L.      95.0% C.L.
Intercept          2.723747        0.1203683        22.6284      0.0000        2.48783         2.959665
Zone               -0.2574348      0.1612933        -1.5961      0.1105        -0.5735639      0.0586942
Sigma              0.8428832       6.194304E-02     13.6074      0.0000        0.7298154       0.9734681

Approximate R-Squared
Log Likelihood     -407.2973
Iterations         39
```

The results of ***NCSS*** match those of Helsel (2005) to several decimal places.

Chapter 350

Introduction to Curve Fitting

Introduction

Historians attribute the phrase *regression analysis* to Sir Francis Galton (1822-1911), a British anthropologist and meteorologist, who used the term *regression* in an address that was published in *Nature* in 1885. Galton used the term while talking of his discovery that offspring of seeds “did not tend to resemble their parent seeds in size, but to be always more mediocre [i.e., more average] than they.... The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it.”

The content of Galton’s paper would probably be called *correlation analysis* today, a term which he also coined. However, the term *regression* soon was applied to situations other than Galton’s and it has been used ever since.

Regression Analysis refers to the study of the relationship between a response (dependent) variable, Y, and one or more independent variables, the X’s. When this relationship is reasonably approximated by a straight line, it is said to be *linear*, and we talk of linear regression. When the relationship follows a curve, we call it curvilinear regression.

Usually, you assume that the independent variables are measured exactly (without random error) while the dependent variable is measured with random error. Frequently, this assumption is not completely true, but when it cannot be justified, a much more complicated fitting procedure is required. However, if the size of the measurement error in an independent variable is small relative to the range of values of that variable, least squares regression analysis may be used with legitimacy.

Linear Regression Models

Perhaps the simplest example of a regression model is the familiar straight-line regression between two variables, X and Y, expressed by the formula:

(1) $Y = B_0 + B_1 X$

where B_0 and B_1 are called parameters, which are known constants linking Y and X. B_0 is the y-intercept, B_1 is the slope.

The relationship in (1) is exact. If you know X, you can determine Y exactly. Exact relationships are hard to find in applied science. Usually, you have to deal with empirical approximations determined from observed data. These relationships are represented as follows:

(2) $Y_i = B_0 + B_1 X_i + e_i$

where Y_i and X_i are the ith observed values of the dependent variable and the explanatory (regressor, predictor, or independent) variable, respectively. B_0 and B_1 are unknown parameter constants which must be estimated. The error term, e_i, represents the error at the ith data point. It is customary to assume that $E(e_i)=0$ (unbiased) and $V(e_i)=s^2$ (constant variance).

Actually, linear models include a broader range of models than those represented by equation (2). The main requirement is that the model is linear in the parameters (the B-coefficients). Other linear models are:

(3) $\ln(Y_i) = B_0 + B_1 \ln(X_i) + e_i$

and

(4) $Y_i = e^{B_0} + \sqrt{B_1}\, e^{X_i} + e_i$

At first, (4) appears nonlinear in the parameters. However, if you set $C_0 = e^{B_0}$, $C_1 = \sqrt{B_1}$, and $Z_i = e^{X_i}$ you will notice that it reduces to the form of (2). Models which may be reduced to linear models with suitable transformations are called intrinsically linear models. Model (5) is a second example of an intrinsically linear model.

(5) $Y_i = B_0 [e^{B_1 X_i}] e_i$

Notice that applying a logarithmic transformation to both sides of (5) results in the following:

(6) $\ln(Y_i) = \ln(B_0) + B_1 X_i + \ln(e_i)$

This is now easily recognized as an intrinsically linear model.

You should note that if the errors are normally distributed in (5), their logarithms in model (6) will not be so distributed. Likewise, if the errors, $\log(e_i)$, in (6) are normally distributed, the detransformed errors, e_i, in (5) will not be. Hence, when you are applying transformations to simplify models, you should check to see that the resulting error term has the desired properties. We will come back to this point later.

Nonlinear Regression Models

Nonlinear regression models are those which are not linear in the parameters to begin with nor can they be made so by transformation. A general representation for the nonlinear regression model is:

(7) $Y_i = f(X_i, e_i ;\; B_1, B_2, \ldots, B_p)$

where B_1, B_2, ..., B_p are the *p* parameters to be estimated from your data, and e_i is the error term.

Note that e_i is not necessarily additive as in (2), although this is a common form. An example of an additive model is:

(8) $Y_i = B_0 e^{B_1 (X_i)} + e_i$

Linear models, such as those in (5), are preferred over nonlinear models, such as (8), for two main reasons. First, the linear model is mathematically easier to work with. Parameters may be estimated with explicit expressions. Nonlinear models must use iterative schemes, which may converge to several solutions. Second, often the investigator does not know the actual form of the relationship and is looking for an approximation. The linear model is an obvious place to start.

Least Squares Estimation of Nonlinear Models

The method of least squares minimizes the error sum of squares, Q, which is given by

$$(9) \qquad Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = f(X_i; \hat{B}_1, \hat{B}_2, \ldots)$ is the value predicted for a specific X_i using the parameters estimated by least squares. If the errors are normally distributed, the least squares estimates are also the maximum likelihood estimates. This is one of the reasons we strive for normally distributed errors.

The values of the B's that minimize *Q* in (9) may be found either of two ways. First, if *f()* is a simple function, such as in (2), you may find an analytic solution by differentiating *Q* with respect to B_1, B_2, ..., B_p, setting the resulting partial derivatives equal to zero, and solving the resulting p normal equations. Unfortunately, very few nonlinear models may be estimated this way.

The second method is to try different values for the parameters, calculating *Q* each time, and work towards the smallest *Q* possible. Three general procedures work toward a solution in this manner.

The *Gauss-Newton*, or *linearization*, method uses a Taylor series expansion to approximate the nonlinear model with linear terms. These may be used in a linear regression to come up with trial parameter estimates which may then be used to form new linear terms. The process iterates until a solution is reached.

The *steepest descent* method searches for the minimum *Q* value by iteratively determining the direction in which the parameter estimates should be changed. It is particularly useful when poor starting values are used.

The *Marquardt* algorithm uses the best features of both the Gauss-Newton and the steepest descent methods. This is the procedure that is implemented in this program. Note that *numerical derivatives* are used whenever derivatives are called for.

Starting Values

All iterative procedures require starting values for the parameters. This program finds the starting values for you. However, the values so found may fail to converge or you may be using a user-defined function which does not have preprogrammed starting values. Hence, you will have to supply your own starting values.

Unfortunately, there is no easy method for generating starting values for the B's in every case. However, we can provide you with some guidelines and a general method of attack that will work in many cases.

1. Try entering a 1 or 0 for each parameter and letting the program crank through a few iterations for you. You must be careful not to give impossible values (like taking the square root of a negative number), or the procedure will halt immediately. Even though the procedure may take

longer to converge, the elapsed time will often be shorter than when using steps 2 and 3 below, since they require much more time and effort on your part.

2. Pick *p* observations that spread across the range of the independent variable and solve the model ignoring the error term. The resulting solution will often provide reasonable starting values. This includes transforming the model to a simpler form.

3. Consider the behavior of *f()* as X approaches zero or infinity and substitute in appropriate observations that most closely approximate these conditions. This might be accomplished from a plot of your data or from an examination of the data directly. Once some of the parameters have been estimated in this manner, others may be found by applying step 2 above.

Inferences about Nonlinear Regression Parameters

The following results are from Seber (1989), chapter 5. They require the assumption that the errors are normally distributed with equal variance.

Confidence Intervals for Parameters

Let

(10) $$Y_i = f(X_i; B_1, B_2, \ldots) + e_i \quad (i = 1,2,\ldots,n)$$

represent the nonlinear model that we are interested in fitting. Let B represent the parameters B_1, B_2, ..., B_p. The asymptotic distribution of the estimates of B, which we call $\hat{B}$, is given by

(11) $$\hat{B} \sim N_p(B, \sigma^2 C^{-1}), \quad C = F.' F., \quad F. = [(\frac{\partial f}{\partial Bj})]$$

For large n we have, approximately,

(12) $$\hat{B}_r \pm t_{n-p}^{\alpha/2} s \sqrt{\hat{c}^{rr}}$$

which gives approximate, large-sample 100(1-a)% confidence limits for the individual parameters. Note the *s* is an estimate of σ in (11), based on the residuals from the fit of (10).

These intervals are often referred to as the asymptotic-linearization confidence intervals because they are based on a local linearization of the function (10). If the curvature of (10) is sharp near $\hat{B}_i$, then the approximation will have considerable error and (12) will be unreliable.

Confidence Intervals for a Predicted Value

Using (10) - (12) it is easy to give approximate, asymptotic 100(1-a)% confidence intervals (or prediction intervals) for predicted values. These are:

(13) $$\hat{Y}_0 \pm t_{n-p}^{\alpha/2} s[1 + f_0'(F.' F.)^{-1} f_0]^{1/2}, \quad f_0 = (\frac{\partial f(X_0)}{\partial B_1}, \frac{\partial f(X_0)}{\partial B_2}, \ldots)$$

Note that f_0 and *F.* must be estimated using the $\hat{B}$. Hence, if the fit of (10) is good and there is little curvature, these confidence intervals will be accurate. If the fit is poor or there is sharp curvature near the region of interest, these confidence limits may be unsatisfactory.

Parameterization

One of the first choices you must make is the way parameters are attached to the functional form of a model. For example, consider the following two models:

(14) $$Y_i = \frac{B_0 X_i}{X_i + B_1} + e_i$$

(15) $$Y_i = \frac{X_i}{C_0 X_i + C_1} + e_i$$

These are actually the same basic model. Note that if we let $C_0=1/B_0$ and $C_1=B_1/B_0$, model (15) is simply a rearrangement of (14). However, the statistical properties of these two models are very different. Equations (14) and (15) are two parameterizations of the same basic model.

If there is no precedent for a particular model parameterization, then you should use that model with the best statistical properties. If this case, trial-and-error methods will have to be used to find a model. Often this will include comparing a plot of your data to a plot of the functional forms that are available, until a good match is found. If there are several models possible, a careful study of the error terms (residuals) is necessary to help in your selection.

A common misconception is the view that whether a parameter appears linearly or nonlinearly in the nonlinear model relates directly to its estimation behavior. This is just not the case. (See Ratkowsky (1989) section 2.5.2.)

Another common misconception is that a complicated model is superior to a simple model. In general, the simpler the model, the better the behavior of the estimation process. Adding an extra parameter has unpredictable results on the estimation process. In some cases, it has little effect, while in others it has disastrous consequences.

Overparameterization (using too complicated a model) often leads to convergence problems. These models may have multiple solutions. The estimates from these models are usually biased and nonnormally distributed. They show high correlation among the parameter estimates. This problem may also occur when you use only a portion of a complicated function to fit a set of data. It is always better to find a simpler function that exhibits the functional behavior of your data. (See Ratkowsky (1989) section 2.5.4.)

The Stochastic Term e_i

A regression model such as (10) may be thought of as having a deterministic part *f(X;B1,B2,...)* and a stochastic (random) part e_i. Often, assumptions about the e_i are necessary. The most common are:

1. Independently distributed
2. Identically distributed with constant variance
3. Normally distributed

Independence

Independence means that the error at one value of i (say i=4) is not related to the error at another value of i (say i=5). Independence is often violated when data are taken over time and some carry-over effects are active.

Identicalness

Identicalness means that the distribution of the errors is the same for all values of i (for all data pairs X_i and Y_i). In practice, this assumption is equated with constant variance in the errors. If the variance of the e_i increases or decreases, then this assumption is violated.

Normality

The question of normality is very difficult to assess with small sample sizes (under 100). With large sample sizes, normal probability plots (discussed later) do a pretty good job. Least-squares methods (those used by this program) tend to create normality in the observed residuals even if the actual e_i's are not normal.

Some normality tests are available in the *Descriptive Statistics* module, so you can try them on your residuals. However, most technicians agree that if your observed residuals have a bell-shaped distribution with no outliers, the normality assumption is okay.

Summary

These assumptions are ideals that are only approximately met in practice. Least squares tends to be robust to minor departures from these assumptions. Only when there are major departures such as outliers, a large shift in the size of the variance, or a large serial correlation between successive residuals will estimates be significantly in error.

Interpretation of R-Squared

R-Squared is computed as

(16) $$R^2 = 1 - \frac{\sum_{1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{1}^{n}(Y_i - \bar{Y})^2}$$

That is, it measures the variance accounted for by the nonlinear model over and above that which is accounted for by the mean of Y. When the model does not contain an intercept-type term, this representation of R-Squared must be used carefully. You should also note that the predicted values, $\hat{Y}_i$, might be in the original (detransformed) metric or in a transformed metric. The program selects what we feel is the appropriate metric in each situation.

A common misconception is the view that R-Squared, the proportion of explained variation, is useful as a goodness-of-fit index in all nonlinear regression situations. Only when you have a linear model with a constant term does R-Squared represent the proportion of variation explained. (See Ratkowsky (1989) section 2.5.3.)

Transformed Y and Unequal Variance

One of the challenges of nonlinear modeling is selecting the appropriate form of the error term. For example, consider the two models (17) and (18) which differ only in the way the error term is represented.

(17) $Y = AX^{B} + e_1$

(18) $Y = CX^{D}e_2$

If you take the logs of both sides in (18) you will get

(19) $\ln(T) = \ln(C) + D\ln(X) + \ln(e_2)$

which is linear in the parameters and can be estimated using simple linear regression. Most of us would rather fit (19) with linear regression than fit (17) with a nonlinear least-squares algorithm. Does it matter? Of course it does.

The difference lies in the pattern of the residuals, the e's. If the true relationship is (17) and you fit (19), you will see a strange pattern in the plot of the residuals. They will exhibit nonconstant variance. Again, if the true relationship is (18), then using (17) will result in an improper model.

The point is, the pattern of the residuals, not convenience, dictates the form of the error term. Hence, you should not use (18) and (19) on a curve with constant variance. Instead, you should use (17). Similarly, if the variance is increasing, a variance-stabilizing transformation of Y (like the log) will be useful in making the variance constant across all values of X.

In summary, there are three reasons for transforming Y: firstly, to obtain linearity; secondly, to obtain errors that are normally distributed; and thirdly to obtain a constant error variance. An examination of the residuals both from the fits before and after transformation is the only way to assess which model is appropriate

Reducing Transformation Bias in Curve Fitting

This material is taken from Miller (1984). If possible, read his article.

One of the most common ways of modeling a nonlinear relationship between two variables is to find a transformation for either the dependent or independent variables (or both) that results in a linear relationship. This relationship may then be estimated with standard linear regression methods. The residuals, calculated from the fit after the transformations, are studied to see that the various assumptions hold. This is taught in many courses and textbooks.

In many situations, the real interest lies in the relationship between the variables in the original metric. When the dependent variable has been transformed, a reverse transformation is used to return a transformed predicted value to the original metric. Prediction intervals in the fitted (transformed) metric are detransformed to arrive at corresponding prediction intervals in the original metric. These detransformed predicted values and prediction intervals give values for the median response, not the mean response as is often supposed.

When the estimated mean response is sought and the above methods are used, the resulting estimates are severely biased. This program provides bias correction factors that may be applied when an estimate of the mean response is desired.

Without going into the details of how and why this biasing occurs, we present the following correction procedures that may be used to correct for this bias. Remember, if the median response is okay then these correction factors do not have to be applied. Note that $\hat{\sigma}^2$ is the mean square error from the transformed model. Also $\hat{Y}$ refers to the detransformed predicted value of Y.

The following table shows the dependent variable transformation and the bias correction factor used.

Transformation	Bias Correction Formula
Ln(Y)	$\hat{Y}\left(exp\left(\frac{\hat{\sigma}^2}{2}\right)\right)$
Sqrt(Y)	$\hat{Y}+\hat{\sigma}^2$
1/Y	$\hat{Y}(1+\hat{Y}^2\hat{\sigma}^2)$

Further Reading

This has been a brief introduction to curve fitting. If you want to get into the issues of variable transformations more deeply, we suggest that you begin with Box and Draper (1987), chapters 7 and 8.

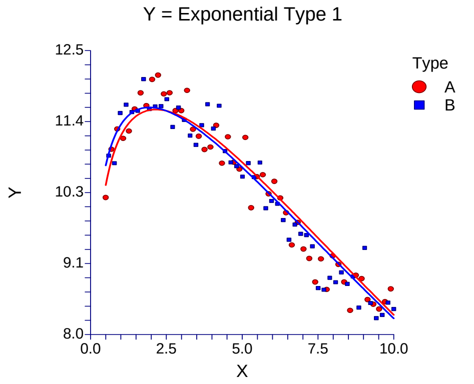
If you want to see examples of fitting curves to data, we suggest Draper and Smith (1981), Hastings (1957), Davis (1962), and Ezekiel and Fox (1967). The first of these is a modern account of nonlinear regression, which goes through several examples. The last three books were written before the computer revolution when the emphasis was on hand calculation. Even though the calculation methods are out of date in these books, they work many examples and provide a great deal of insight into the art of curve fitting.

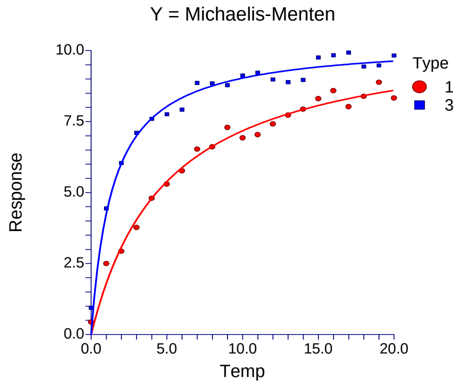
Chapter 351

Curve Fitting – General

Introduction

Curve fitting refers to finding an appropriate mathematical model that expresses the relationship between a dependent variable *Y* and a single independent variable *X* and estimating the values of its parameters using nonlinear regression. An introduction to curve fitting and nonlinear regression can be found in the chapter entitled Curve Fitting, so these details will not repeated here. Here are some examples of the curve fitting that can be accomplished with this procedure.





This program is general purpose curve fitting procedure providing many new technologies that have not been easily available. It is preprogrammed to fit over forty common mathematical models including growth models like linear-growth and Michaelis-Menten. It also fits many approximating models such as regular polynomials, piecewise polynomials and polynomial ratios. In addition to these preprogrammed models, it also fits models that you write yourself.

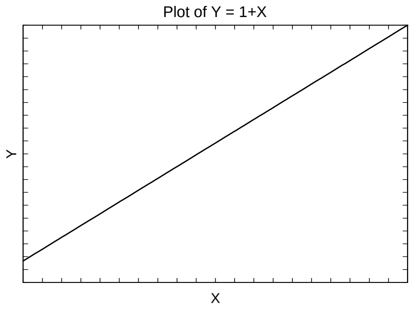
This routine includes several innovative features. First, it can fit curves to several batches of data simultaneously. Second, it compares fitted models across groups using graphics and numerical tests such as an approximate F-test for curve coincidence and a computer-intensive randomization test that compares curve coincidence and individual parameter values. Third, this routine computes bootstrap confidence intervals for parameter values, predicted means, and predicted values using the latest computer-intensive bootstrapping technology.

Selecting a Preset Model

Over thirty preset models are available. These models provide a variety of curve shapes. Several of the models were developed for quite different physical processes, but yield similar results. We now present examples and details of several of the preset models available.

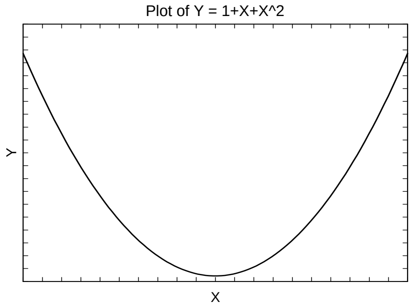
1. Linear: Y=A+BX

This common model is usually fit using standard linear regression techniques. We include it here to allow for various special forms made by transforming X and Y



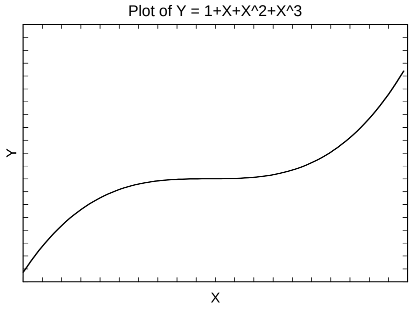
2. Quadratic: Y=A+BX+CX^2

The quadratic or second-order polynomial model results in the familiar parabola.



3. Cubic: Y=A+BX+CX^2+DX^3

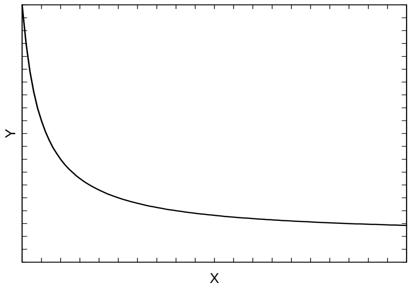
This is the cubic or third-order polynomial model.



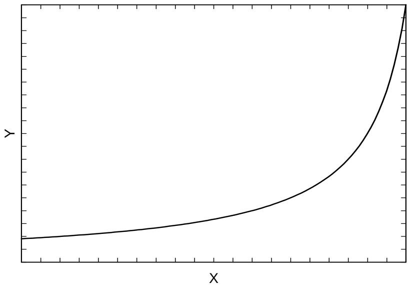
4. PolyRatio(1,1): Y=(A+BX)/(1+CX)

The ratio of first-order polynomials model is a slight extension of the Michaelis-Menten model. It may be used to approximate many more complicated models.

Plot of Y = (5+X)/(1+2*X)



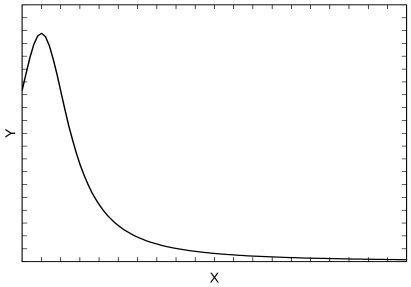
Plot of Y = (1+X)/(1-X)



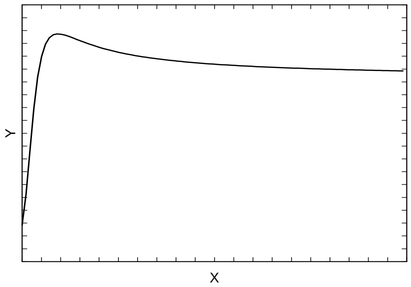
5. PolyRatio(2,2): Y=(A+BX+CX^2)/(1+DX+EX^2)

The ratio of second-order polynomials model may be used to approximate many complicated models.

Plot of Y = (1+X-X^2)/(1-X+X^2)



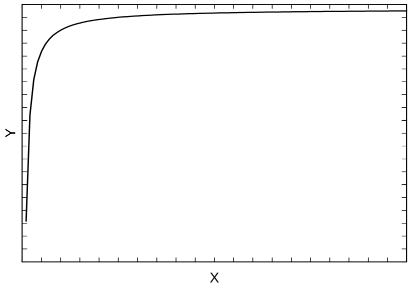
Plot of Y = (1+X+X^2)/(5-X+X^2)



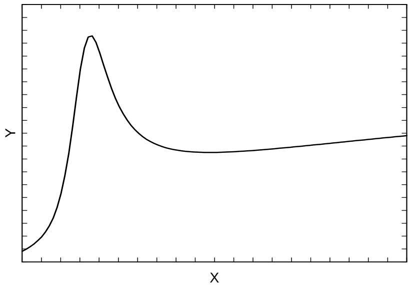
6. PolyRatio(3,3): Y=(A+BX+CX^2+DX^3)/(1+EX+FX^2+GX^3)

The ratio of third-order polynomials model may be used to approximate many complicated models. However, care must be used when estimating such high-degree models.

Plot of Y = (1+X+X^2+X^3)/(1-X+X^2-X^3)



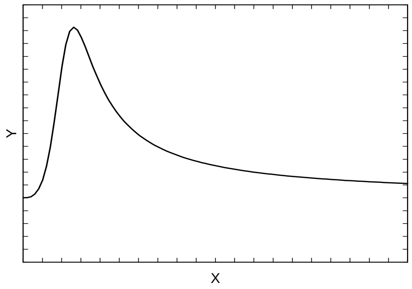
Plot of Y = (1+2*X+X^2+X^3)/(1+X+8*X^2+X^3)



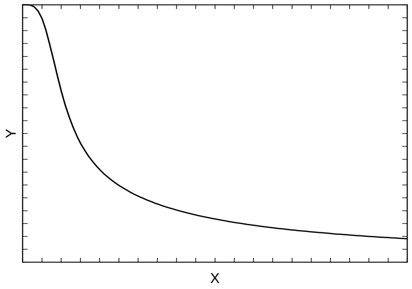
7. PolyRatio(4,4): Y=(A+BX+CX^2+DX^3+EX^4) / (1+FX+GX^2+HX^3+IX^4)

The ratio of fourth-order polynomials model may be used to approximate many complicated models. However, care must be used when estimating such high-degree models.

Plot of Y = (1+X^3+X^4)/(1-X^3+X^4)



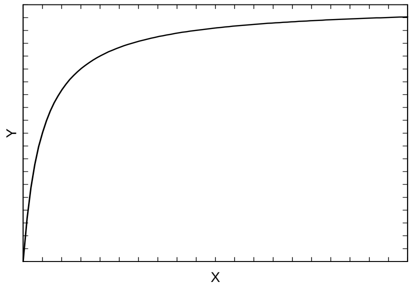
Plot of Y = (1+X^3-X^4)/(1+X^3+X^4)



8. Michaelis-Menten: Y=AX/(B+X)

This is a popular growth model.

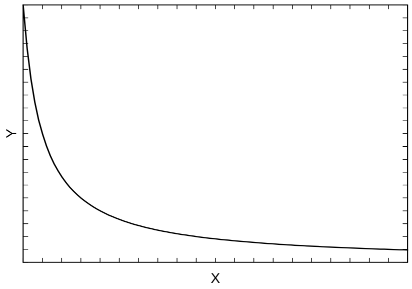
Plot of Y = X/(1+X)



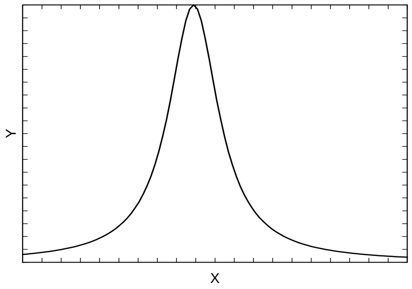
9. Reciprocal: Y=1/(A+BX)

This model, known as the reciprocal or Shinozaki and Kira model, is mentioned in Ratkowsky (1989, page 89) and Seber (1989, page 362).

Plot of Y = 1/(1+X)

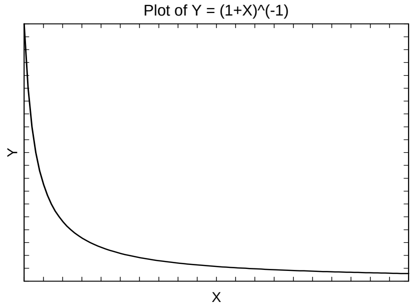


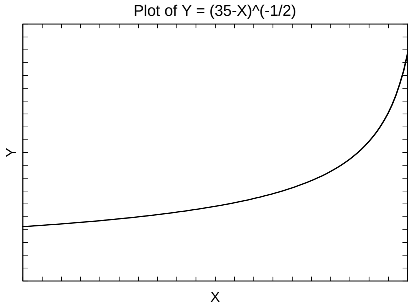
Plot of Y = 1/(4+2*X^2)



10. Bleasdale-Nelder: Y=(A+BX)^(-1/C)

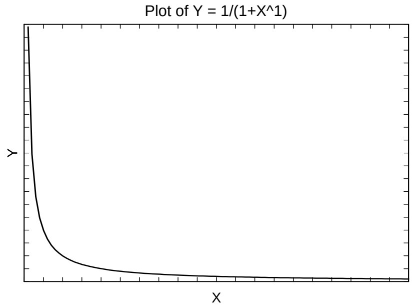
This model, known as the Bleasdale-Nelder model, is mentioned in Ratkowsky (1989, page 103) and Seber (1989, page 362).

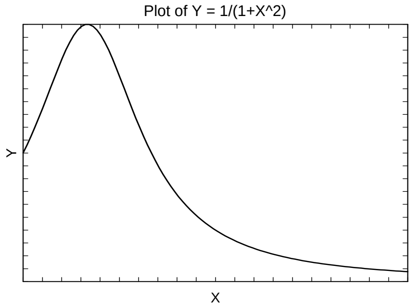

Plot of Y = (1+X)^(-1)

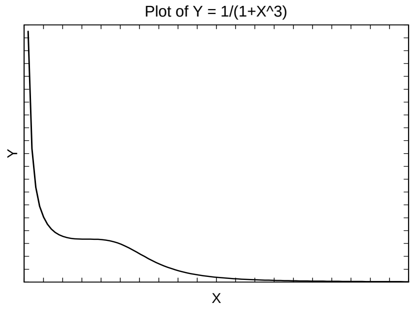

Plot of Y = (35-X)^(-1/2)

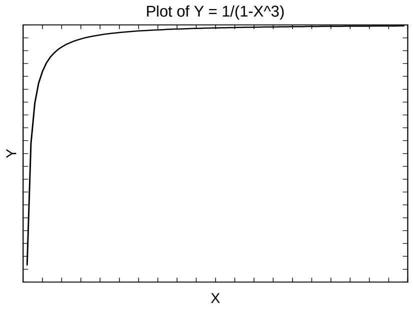
11. Farazdaghi and Harris: Y=1/(A+BX^C)

This model, known as the Farazdaghi and Harris model, is mentioned in Ratkowsky (1989, pages 99 and 104) and Seber (1989, page 362).


Plot of Y = 1/(1+X^1)

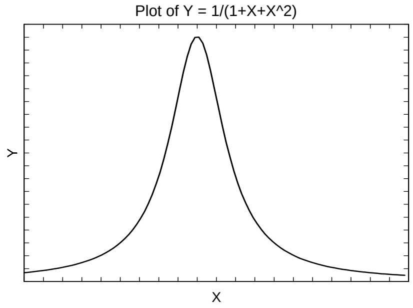

Plot of Y = 1/(1+X^2)


Plot of Y = 1/(1+X^3)


Plot of Y = 1/(1-X^3)

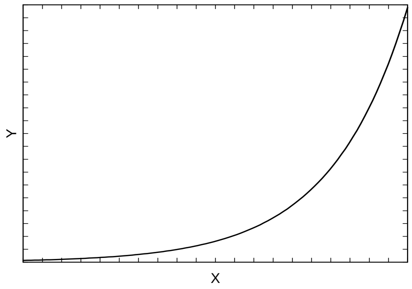
12. Holliday: Y=1/(A+BX+CX^2)

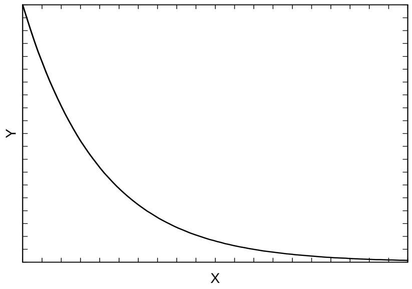
This model, known as the Holliday model, is mentioned in Seber (1989, page 362).


Plot of Y = 1/(1+X+X^2)

13. Exponential: Y=EXP(A(X-B))

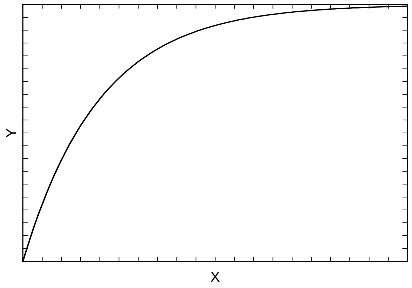
This model, known as the exponential model, is mentioned in Seber (1989, page 327). Note that taking the log of both sides reduces this equation to a linear model.

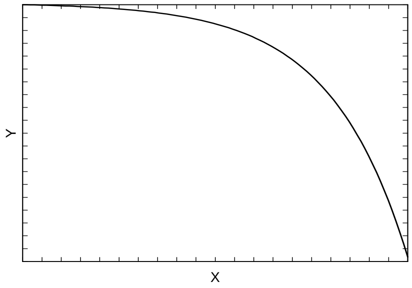

Plot of Y = EXP(X)


Plot of Y = EXP(-X)

14. Monomolecular: Y=A(1-EXP(-B(X-C)))

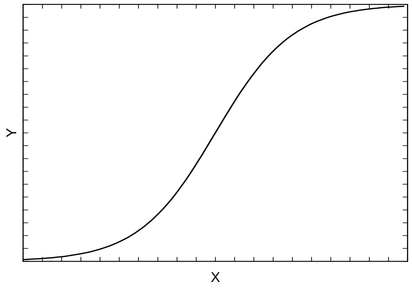
This model, known as the monomolecular model, is mentioned in Seber (1989, page 328).


Plot of Y = 1-EXP(-X)


Plot of Y = 1-EXP(X)

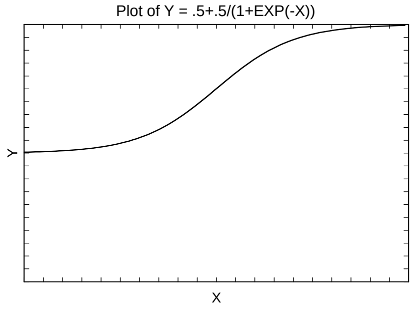
15. Three Parameter Logistic: Y=A/(1+B(EXP(-CX)))

This model, known as the three-parameter logistic model, is mentioned in Seber (1989, page 330).


Plot of Y = 1/(1+EXP(-X))

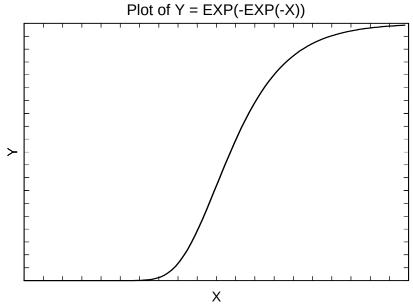
16. Four Parameter Logistic: Y=D+(A-D)/(1+B(EXP(-CX)))

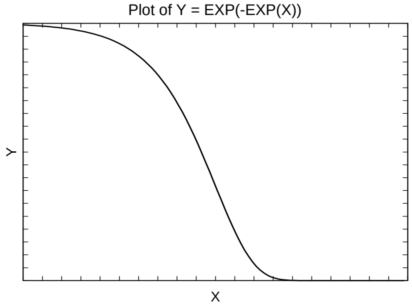
This model, known as the four-parameter logistic model, is mentioned in Seber (1989, page 338). Note that the extra parameter, D, has the effect of shifting the graph vertically. Otherwise, this plot is the same as the three-parameter logistic.



17. Gompertz: Y=A(EXP(-EXP(-B(X-C))))

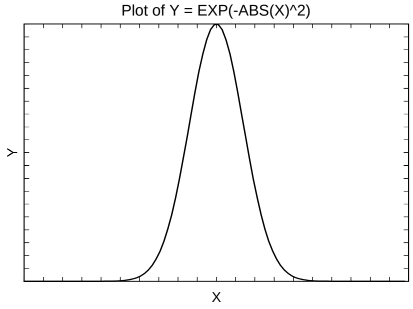
This model, known as the Gompertz model, is mentioned in Seber (1989, page 331).

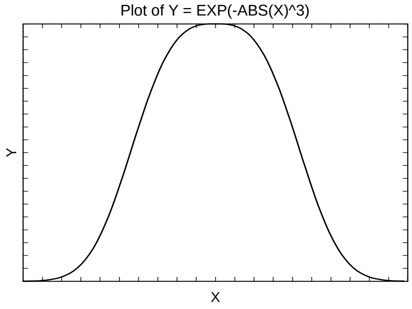




18. Weibull: Y=A-(A-B)EXP(-(C|X|)^D)

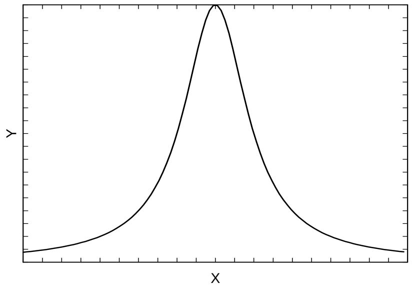
This model, known as the Weibull model, is mentioned in Seber (1989, page 338).

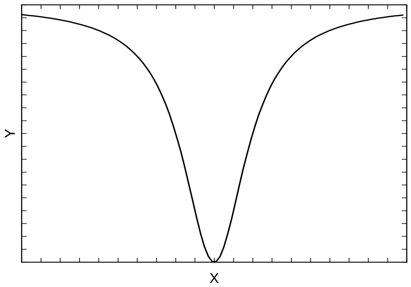




19. Morgan-Mercer-Floding: Y=A-(A-B)/(1+(C|X|)^D)

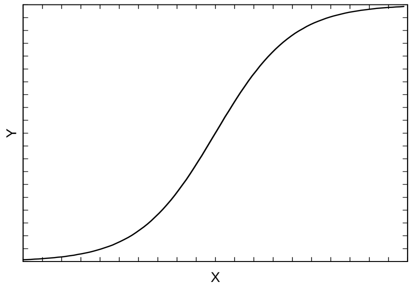
This model, known as the Morgan-Mercer-Floding model, is mentioned in Seber (1989, page 340).

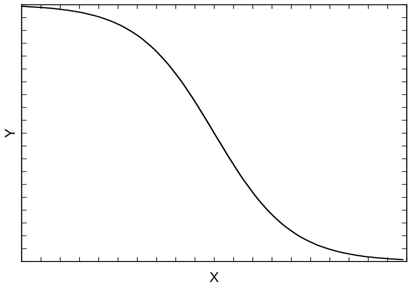

Plot of Y = 1/(1+ABS(X)^2)


Plot of Y = 1/(1+ABS(X)^(-2))

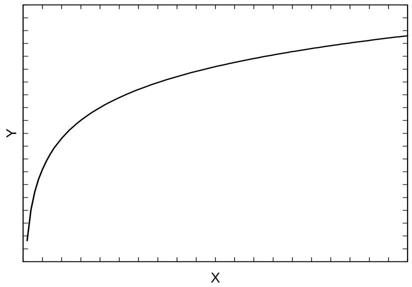
20. Richards: Y=A(1+(B-1)EXP(-C(X-D)))^(1/(1-B))

This model, known as the Richards model, is mentioned in Seber (1989, page 333).

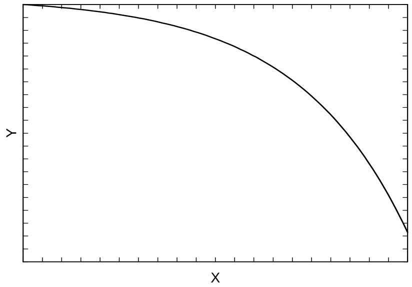

Plot of Y = 1/(1+EXP(-X))

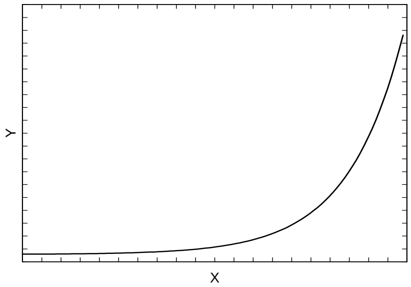

Plot of Y = 1/(1+EXP(X))

21. Logarithmic: Y=B(LN(|X|-A))

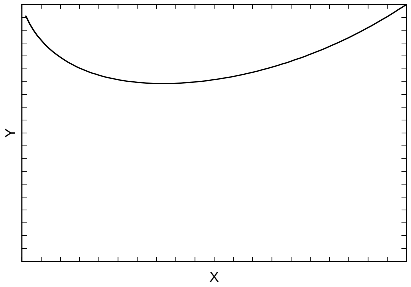

Plot of Y = LOG(ABS(X))

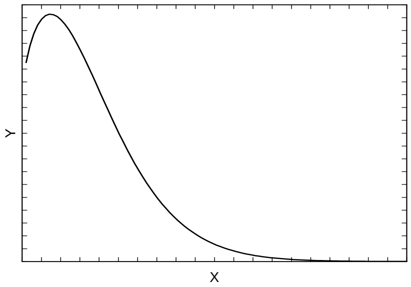
22. Power: Y=A(1-B^X)


Plot of Y = 1-2^X

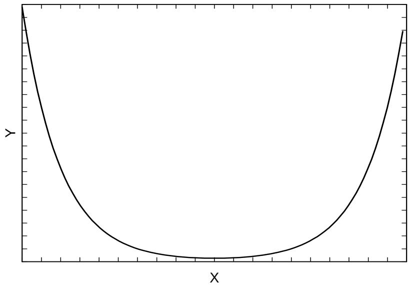

Plot of Y = 1+2^X

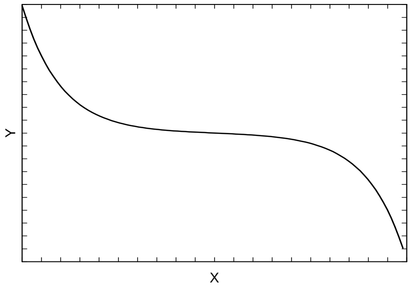
23. Power^Power: Y=AX^(BX^C)


Plot of Y = X^X

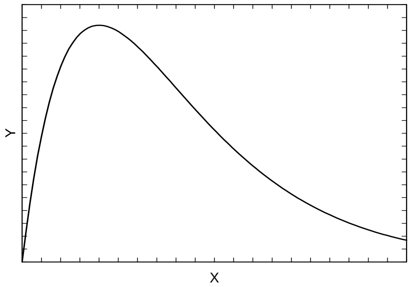

Plot of Y = X^(-X)

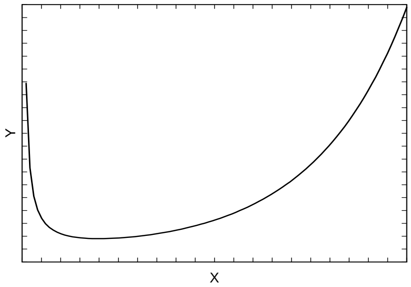
24. Sum of Exponentials: Y=A(EXP(-BX))+C(EXP(-DX))


Plot of Y = EXP(-X)+EXP(X)

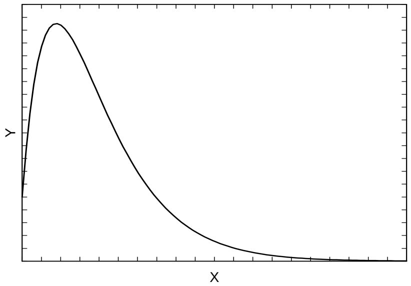

Plot of Y = EXP(-X)-EXP(X)

25. Exponential Type 1: Y=A(X^B)EXP(-CX)

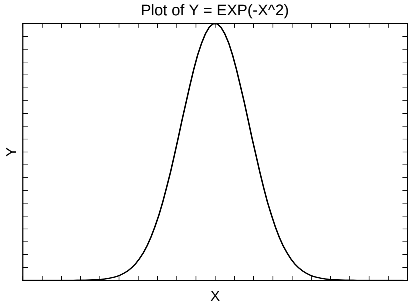

Plot of Y = X*EXP(-X)


Plot of Y = 1/X*EXP(X)

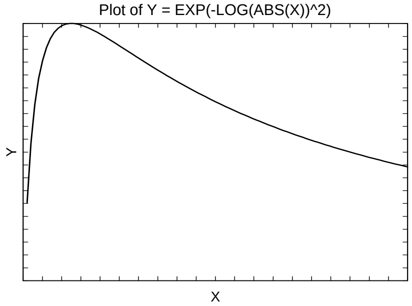
26. Exponential Type 2: Y=(A+BX)EXP(-CX)+D


Plot of Y = (1+(9*X))*EXP(-X)

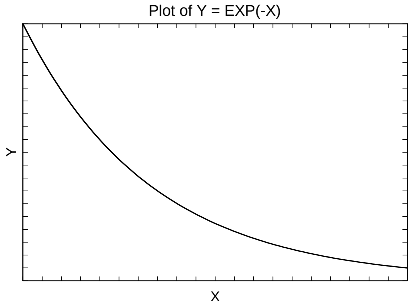
27. Normal: Y=A+B(EXP(-C(X-D)^2))



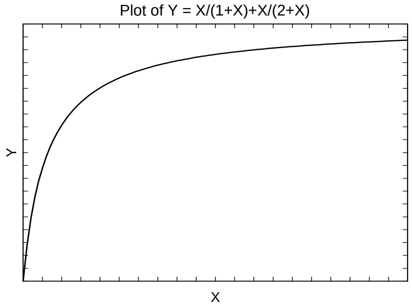
28. Lognormal: Y=A+(B/X)EXP(-C(LN(|X|)-D)^2)



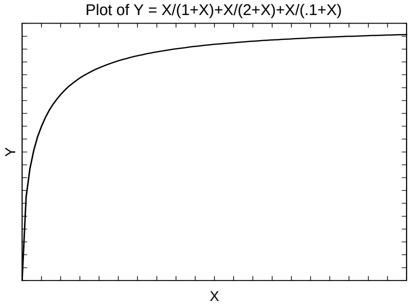
29. Exponential: Y=A Exp(-BX)



30. Michaelis-Menten(2): Y=AX/(B+X) + CX/(D+X)



31. Michaelis-Menten(3): Y=AX/(B+X) + CX/(D+X) + EX/(F+X)



32. Linear-Linear: Y=A + BX + C(X-D)SIGN(X-D)

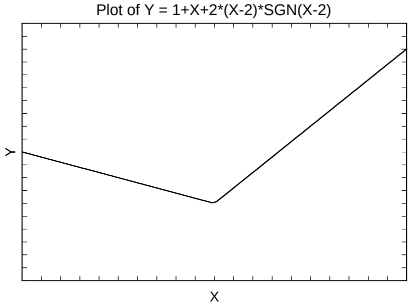
Common Equation

Y = a1 + b1X, X<J

Y = a2 + b2X, X³J

Parameter Identities

A=(a1+a2)/2	B=(b1+b2)/2	a1=A+DC	b1=B-C	J=D
C=(b2-b1)/2	D=J	a2=A-DC	b2=B+C	



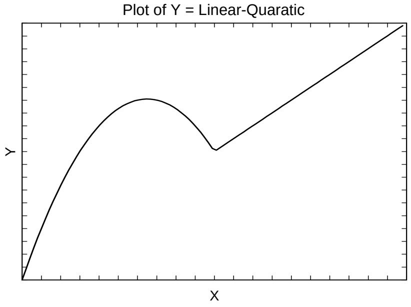
33. Linear-Quadratic: Y=A+BX+CX^2+(X-D)SIGN(X-D)[C(X+D)+E]

Common Equation

Y=a1+b1X,	X<=a
Y=a2+b2X+c2X^2,	X>a

Parameter Identities

A=(a1+a2)/2	B=(b1+b2)/2	C=c2/2
D=a	E=(b2-b1)/2	
a1=A+CD2+DE	b1=B-E	a=D
a2=A-CD2-DE	b2=B+E	c2=2C



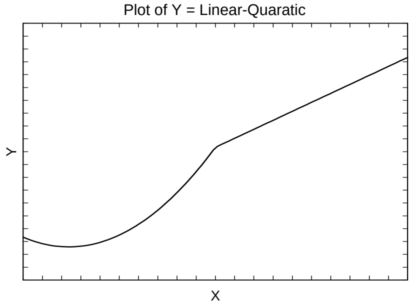
34. Quadratic-Linear: Y=A+BX+CX^2+(X-D)SIGN(X-D)[E(X+D)+F]

Common Equation

Y=a1+b1X+c1X^2,	X<=a
Y=a2+b2X,	X>a

Parameter Identities

A=(a1+a2)/2	B=(b1+b2)/2	C=c1/2
D=a		E=(b2-b1)/2
a1=A-CD2+DE	b1=B-E	a=D
a2=A+CD2-DE	b2=B+E	c1=2C


Plot of Y = Linear-Quaratic

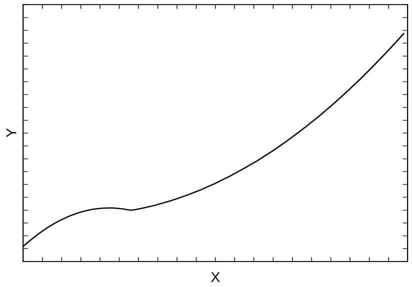
35. Quadratic-Quadratic: Y=A+BX+CX^2+(X-D)SIGN(X-D)[E(X+D)+F]

Common Equation

Y=a1+b1X+c1X^2,	X<=a
Y=a2+b2X+c2X^2,	X>a

Parameter Identities

A=(a1+a2)/2	B=(b1+b2)/2	C=(c1+c2)/2
D=a	E=(c2-c1)/2	F=(b2-b1)/2
a1=A-ED2+DF	b1=B-F	a=D
a2=A+eD2-DF	b2=B+F	
c1=C-E	c2=C+E	


Plot of Y = Quadratic-Quadratic

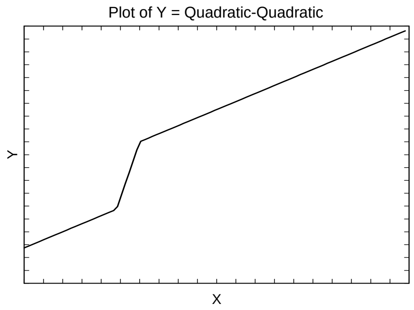
36. Linear-Linear-Linear: Y=A+BX+C(X-D)SIGN(X-D)+E(X-F)SIGN(X-F)

Common Equation

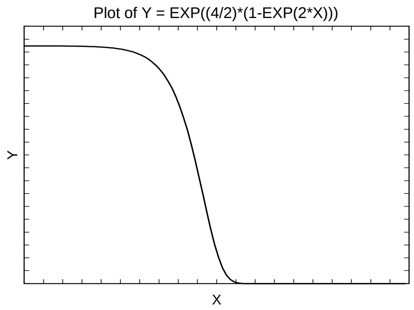
Y=a1+b1X	X<J1
Y=a2+b2X	a1<X<=J2
Y=a3+b3X	X>J2

Parameter Identities

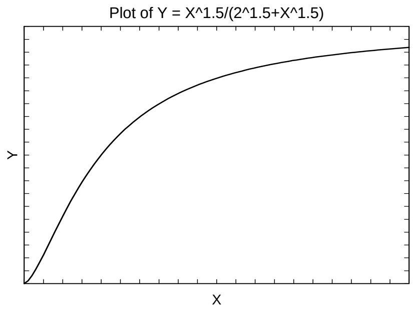
A=(a1+a3)/2	B=(b1+b3)/2	C=(b2-b1)/2
D=J1	E=(b3-b2)/2	F=J2
a1=A+CD+EF	b1=B-C-E	J1=D
a2=A-CD-EF	b2=B+C-E	J2=F
a3=A-CD+EF	b3=B+C+E	



37. Gompertz 2: Y=Exp((A/B)(1-Exp(BX)))

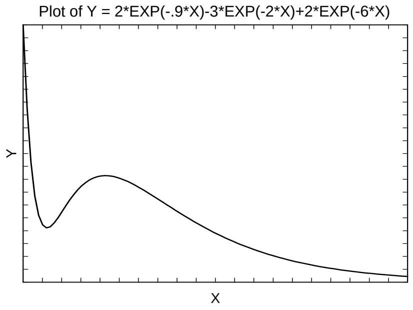


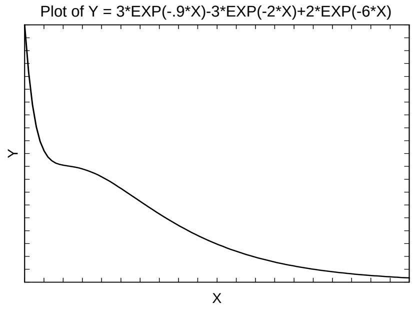
38. Hill: Y=AX^C/(B^C+X^C)



39. Sum of 3 Exponentials: Y=A(Exp(-BX))-C(Exp(-DX))+E(Exp(-FX))

This model is intended for the case when all parameters are positive. Note that the default starting values may not work for this model. You should be prepared to try different starting values.


Plot of Y = 2*EXP(-.9*X)-3*EXP(-2*X)+2*EXP(-6*X)


Plot of Y = 3*EXP(-.9*X)-3*EXP(-2*X)+2*EXP(-6*X)

Custom Models

You are not limited to the preset models that are shown above. You can enter your own custom model using standard mathematical notation. The only difference between using a preset model and using your own model is that with a preset model the starting values of the search algorithm are chosen based on the model. When using a custom model, you will have to set your own starting values based on the data you are trying to fit. When you do not specify starting values, the program uses all zeros, which may or may not lead to a reasonable solution.

Confidence Intervals

Two methods are used to calculate confidence intervals of the regression parameters and predicted values. The first method is based on the usual normality and constant variance of residuals assumption. When the data follow these assumptions, standard expressions for the confidence intervals are used based on the Student's *t* distribution. Unfortunately, nonlinear regression dataset rarely follow these assumptions.

The second method is called the *bootstrap* method. This is a modern, computer-intensive method that has only become available in recent years as extensive computer power has become available.

Bootstrap Confidence Intervals

Bootstrapping provides standard errors and confidence intervals for nonlinear-regression parameter, predicted means, and predicted values. The method is simple in concept, but it requires extensive computation time.

Bootstrap confidence intervals are based on the assumption that your sample is actually representative of the population. Beginning with this assumption, *B* samples are drawn (*B* is over 1000) of size *N* from your original sample with replacement. With replacement sampling means that each observation may be selected more than once. For each bootstrap sample, the nonlinear-regression results are computed and stored.

Suppose you want the standard error and a confidence interval of a regression parameter. The bootstrap sampling process provides *B* estimates of this parameter. The standard deviation of these *B* estimates is the bootstrap estimate of the standard error of the parameter. The bootstrap confidence interval is found by arranging the *B* values in sorted order and selecting the appropriate percentiles from the list. For example, a 90% bootstrap confidence interval for the parameter is given by fifth and ninety-fifth percentiles of the bootstrap parameter values.

The main assumption made when using the bootstrap is that your sample approximates the population. Because of this assumption, bootstrapping does not work well for small samples in which there is little likelihood that the sample is representative of the population. Bootstrapping should only be used in medium to large samples.

Bootstrap Prediction Intervals

Bootstrap confidence intervals for the mean of *Y* given *X* are generated from the bootstrap sample in the usual way. To calculate prediction intervals for the predicted value (not the mean) of *Y* given *X* requires a modification to the predicted value of *Y* to be made to account for the variation of *Y* about its mean. This modification of the predicted *Y* values in the bootstrap sample, suggested by Davison and Hinkley, is as follows.

$$\hat{y}_i = \hat{y}_i + e_r^*$$

where e_r^* is a randomly selected modified residual (see below). By adding the residual we have added an appropriate amount of variation to represent the variance of individual *Y*'s about their mean value.

Modified Residuals

Davison and Hinkley (1999) page 279 recommend the use of a special rescaling of the residuals when bootstrapping to keep results unbiased. Because of the high amount of computing involved in bootstrapping, these modified residuals are calculated using

$$e_j^* = \frac{e_j}{\sqrt{1 - \frac{1}{N}}} - \bar{e}$$

where

$$\bar{e} = \frac{\sum_{j=1}^{N} e_j}{N}$$

Note that there is a different rescaling than Davison and Hinkley recommended. We have used this rescaling because it is much quicker to calculate.

Hypothesis Testing

When curves are fit to two or more groups, it is often of interest to test whether certain regression parameters are equal and whether the fitted curves coincide. Although some approximate results have been obtained using indicator variables, these are asymptotic results and little is known about their appropriateness in sample samples. We provide a test of the hypothesis that all group curves coincide using an *F*-test that compares the residual sum of squares obtained when the grouping is

ignored with the total of the residual sum of squares obtained for each group. This test is routinely used in the analysis of variance associated with linear models and its application to nonlinear models has occasionally been suggested. However, it is based on naive assumptions that seldom occur.

Because of the availability of fast computing speed in recent years, a second method of hypothesis testing, called the *randomization test*, is now available. This test will be discussed next.

Randomization Test

Randomization testing is discussed by Edgington (1987). The details of the randomization test are simple: all possible permutations of the group variable while leaving the dependent and independent variables in their original order are investigated. For each permutation, the difference between the estimated group parameters is calculated. The number of permutations with a magnitude greater than or equal to that of the actual sample is counted. Dividing this count by the number of permutations gives the significance level of the test.

The randomization test is suggested because an exact test is achieved without making unrealistic assumptions about the data such as constant variance, normality, or model accuracy. The test was not used in the past because the amount of computations was prohibitive. In fact, the randomization test was originally proposed by Fisher and he chose his *F*-test because its distribution close approximated the randomization distribution.

The only assumption that a randomization test makes is that the data values are *exchangeable* under the null hypothesis.

For even moderate sample sizes, the total number of permutations is in the trillions, so a Monte Carlo approach is used in which the permutations are found by random selection rather than enumeration. Using this approach, a reasonable approximation to the test's probability level may be found by considering only a few thousand permutations rather than the trillions needed for complete enumeration. Edgington suggests that at least 1000 permutations by computed. We suggest that this be increased to 10000 for important results.

The program tests two types of hypotheses using randomization tests. The first is that each of the estimated model parameters is equal. The second is that the individual fitted curves coincide across all groups.

Randomization Statistics for Testing Parameter Equivalence

The test statistic for comparing a model parameter is formed by summing the difference between the group parameter estimates for each pair of groups. If there are *G* groups, the test statistic is computed using the formula

$$B_{RT} = \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} \left| \hat{\beta}_i - \hat{\beta}_j \right|$$

Randomization Statistics for Testing Curve Equivalence

The test statistic for comparing the whole curve is formed by summing the difference between the estimated predicted values for each pair of groups at several points along the curve. If there are *G* groups and *K* equally spaced test points, the test statistic is computed using the formula

$$C_{RT} = \sum_{k=1}^{K} \sum_{i=1}^{G-1} \sum_{j=i+1}^{G} \left| \hat{y}_{ki} - \hat{y}_{kj} \right|$$

Data Structure

The data are entered in two variables: one dependent variable and one independent variable. Additionally, you may specify a frequency variable containing the observation count for each row and a group variable that is used to partition the data in to independent groups.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable from the current database. This variable is being predicted using the (preset or custom) model you specify. The actual values fed into the algorithm depend on which transformation (if any) is selected for this variable.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y'=1/(Y*Y), ***Y'=1/Y***, ***Y'=1/SQRT(Y)***, ***Y'=LN(Y)***, ***Y'=SQRT(Y)***, ***Y'=Y (none)***, and ***Y'=Y*Y***

Care must be taken so that you do not apply a transformation that omits much of your data. For example, you cannot take the square root of a negative number, so if you apply this transformation to negative values, those observations will be treated as missing values and ignored. Similarly, you cannot have a zero in the denominator of a quotient and you cannot take the logarithm of a number less than or equal to zero.

X (Independent) Variable

Specify the independent (X) variable. This variable is used to predict the dependent variable using the model you have specified. This variable is referred to as 'X' in the Preset and Custom model statements. The actual values used depend on which transformation (if any) is selected for this variable.

X Transformation

Specifies a power transformation of the independent variable. Available transformations are

X'=1/(X*X), ***X'=1/X***, ***X'=1/SQRT(X)***, ***X'=LN(X)***, ***X'=SQRT(X)***, ***X'=X (none)***, and ***X'=X*X***

Care must be taken so that you do not apply a transformation that omits much of your data. For example, you cannot take the square root of a negative number, so if you apply this transformation to negative values, those observations will be treated as missing values and ignored. Similarly, you cannot have a zero in the denominator of a quotient and you cannot take the logarithm of a number less than or equal to zero.

Frequency Variable

An optional variable containing a set of counts (frequencies). Normally, each row represents one observation. On occasion, however, each row of data may represent more than one observation. This variable contains the number of observations that a row represents. Rows with zeroes and negative values are ignored.

Group Variable

This optional variable divides the observations into groups. When specified, a separate analysis is generated for each unique value of this variable. Use the Value Label option under the Format tab to specify the way in which the group values are displayed.

Model

Preset Model

Select the model that you want to fit. Select ‘Custom’ to use a model you have entered in the ‘Custom Model’ box. Whenever possible, use one of the preset models since reasonable starting values for the parameters will be calculated for you. The minimum, maximum, and starting values of each letter in the preset model are defined in the corresponding MIN START MAX box on the Options panel. The preset models available are

0	Custom	Use the custom model
1	Y=A+BX	Simple Linear
2	Y=A+BX+CX^2	Quadratic
3	Y=A+BX+CX^2+DX^3	Cubic
4	Y=(A+BX)/(1+CX)	PolyRatio(1,1)
5	Y=(A+BX+CX^2)/(1+DX+EX^2)	PolyRatio(2,2)
6	Y=(A+BX+CX^2+DX^3)/(1+EX+FX^2+GX^3)	PolyRatio(3,3)
7	Y=(A+BX+CX^2+DX^3+EX^4) / (1+FX+GX^2+HX^3+IX^4)	PolyRatio(4,4)
8	Y=AX/(B+X)	Michaelis-Menten
9	Y=1/(A+BX)	Reciprocal
10	Y=(A+BX)^(-1/C)	Bleasdale-Nelder
11	Y=1/(A+BX^C)	Farazdaghi and Harris
12	Y=1/(A+BX+CX^2)	Holliday
13	Y=EXP(A(X-B))	Exponential
14	Y=A(1-EXP(-B(X-C)))	Monomolecular
15	Y=A/(1+B(EXP(-CX)))	Three Parameter Logistic
16	Y=D+(A-D)/(1+B(EXP(-CX)))	Four Parameter Logistic
17	Y=A(EXP(-EXP(-B(X-C))))	Gompertz
18	Y=A-(A-B)EXP(-(C|X|)^D)	Weibull
19	Y=A-(A-B)/(1+(C|X|)^D)	Morgan-Mercer-Floding
20	Y=A(1+(B-1)EXP(-C(X-D)))^(1/(1-B))	Richards
21	Y=B(LN(|X|-A))	Logarithmic
22	Y=A(1-B^X)	Power

23	Y=AX^(BX^C)	Power^Power
24	Y=A(EXP(-BX))+C(EXP(-DX))	Sum of Exponentials
25	Y=A(X^B)EXP(-CX)	Exponential Type 1
26	Y=(A+BX)EXP(-CX)+D	Exponential Type 2
27	Y=A+B(EXP(-C(X-D)^2))	Normal
28	Y=A+(B/X)EXP(-C(LN(|X|)-D)^2)	Lognormal
29	Y=A Exp(-BX)	Exponential
30	Y=AX/(B+X) + CX/(D+X)	Michaelis-Menten(2)
31	Y=AX/(B+X) + CX/(D+X) + EX/(F+X)	Michaelis-Menten(3)
32	Y=A + BX + C(X-D)SIGN(X-D)	Linear-Linear
33	Y=A+BX+CX^2+(X-D)SIGN(X-D)[C(X+D)+E]	Linear-Quadratic
34	Y=A+BX+CX^2+(X-D)SIGN(X-D)[E(X+D)+F]	Quadratic-Linear
35	Y=A+BX+CX^2+(X-D)SIGN(X-D)[E(X+D)+F]	Quadratic-Quadratic
36	Y=A+BX+C(X-D)SIGN(X-D)+E(X-F)SIGN(X-F)	Linear-Linear-Linear
37	Y=Exp((A/B)(1-Exp(BX)))	Gompertz 2
38	Y=AX^C/(B^C+X^C)	Hill

Custom Model

This box is only used when the Preset Model option is set to ‘Custom Model’. When used, it contains the regression model written in standard mathematical notation.

Use ‘X’ to represent the independent variable specified in the X Variable box, not its variable name. Hence, if your independent variable is HEAT, you would enter A+B*LN(X), not A+B*LN(HEAT).

Use the letters (case ignored) A,B,C,... (except X and Y) to represent the parameters to be estimated from the data. The letters used must be specified in one of the Parameter boxes listed under the Search tab. Note that you do not include a ‘Y=’ in the expression. That is, you would enter A+B*X, not Y=A+B*X.

Expression Syntax

Construct the expression using standard mathematical syntax. Possible symbols and functions are

Symbols

+	add
-	subtract
*	multiply
/	divide
^	exponent (X^2 = X*X)
()	parentheses
<	less than.
>	greater than
=	equals
<=	less than or equal
>=	greater than or equal
<>	not equal

Functions

(a logic b)	Indicator function. If true, result is 1; otherwise, result is 0. Logic values are <, >, =, <>, <=, and >=. The symbols a and b are replaced by numbers or letters.
ABS(X)	Absolute value of X.
ARCOSH(X)	Arc cosh of X.
ARSINH(X)	Arc sinh of X.
ARTANH(X)	Arc tanh of X.
ASN(X)	Arc sine of X.
ATN (X)	Arc tangent of X.
COS(X)	Cosine of X.
COSH(X)	Hyperbolic cosine of X.
ERF(X)	The error function of X
EXP(X)	Exponential of X.
INT(X)	Integer part of X.
LN(X)	Log base e of X.
LOG(X)	Log base 10 of X.
LOGGAMMA(X)	Log of the gamma function.
NORMDENS(X)	Normal density.
NORMPROB(X)	Normal CDF (probability).
NORMVALUE(X)	Inverse normal CDF.
SGN(X)	Sign of X which is -1 if X<0, 0 if X=0, and 1 if X>0.
SIN(X)	Sine of X.
SINH(X)	Hyperbolic sine of X.
SQR(X)	Square root of X.
TAN(X)	Tangent of X.
TANH(X)	Hyperbolic tangent of X.
TNH(X)	Hyperbolic tangent of X.
TRIGAMMA(X)	Trigamma function.

Independent Variable

Use ‘X’ in your expression to represent the independent variable you have specified.

Parameters

The letters of the alphabet (except X and Y) may be used to represent the parameters. Parameters can be only one character long and case is ignored. Each parameter must be defined in the Parameter fields below.

Numbers

You can enter numbers in standard format such as 23.456 and 254.43, or you can use scientific notation such as 1E-5 (which is 0.00001) and 1E5 (which is 100000).

Examples

Standard mathematical syntax is used. This is discussed in detail in the Transformation section. Examples of valid expressions are:

```
A+B*X
```

```
C+D*X+E*X*X or G+H*X+B*X^2
```

A*EXP(B*X)

(X<=5)*A+(X>5)*B+C

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias that may occur and the bias correction procedures used.

Model Parameters

The following options control the nonlinear regression algorithm.

Parameter

Enter a letter (other than X and Y) used in the Model. Note that the case of the character is ignored. Each letter used in a Model (either Preset or Custom) must be defined in this section by entering its letter, bounds, and starting value.

For example, suppose the model is **A + B*X + C*X^2**. The parameters in this expression are A, B, and C. Each must be defined here.

Min Start Max

Enter the minimum, starting value, and maximum of this parameter by entering three numbers separated by blanks or commas. You may enter '?' as the starting value to instruct the program pick one for you (in which case a zero is often used). The program searches for the best value between the minimum and the maximum values, beginning with the starting value.

Make sure that the starting values you supply are possible. For example, if the model includes the phrase 1/B, don't start with B=0. Before taking a lot of time trying to find a starting value, make a few trial runs using starting values of 0.0, 0.1, and 1.0. Often, one of these values will work.

Examples

-1000 1 1000 which means starting value = 1, lower bound = -1000, and upper bound = 1000.

-1 ? 1E9 which means starting value is unspecified, lower bound = -1, and upper bound = 1000000000.

- **Minimum**

  This is the smallest value that the parameter can take on. The algorithm searches for a value between this and the maximum. If you want to search in an unlimited range, enter a large negative number such as -1E9, which is -1000000000.

  Since this is a search algorithm, the narrower the range that you search in, the quicker it will converge.

  Care should be taken to specify minima and maxima that keep calculations in range. Suppose, for example, that your equation includes the expression LOG(B*X) and that values of X are positive. Since you cannot take the logarithm of zero or a negative number, you should set the minimum of B as a small positive number, insuring that the estimation procedure will not fail because of impossible calculations.

- **Starting Value**

  Enter a starting value for this parameter or enter '?' to have the system estimate a starting value for you. When using a custom model, a '?' is replaced by zero.

- **Maximum**

  This is the largest value that the parameter can take on. The algorithm searches for a value between the minimum and this value, beginning at the Starting Value. If you want to search in an unlimited range, enter a large positive number such as 1E9, which is 1000000000.

  Since this is a search algorithm, the narrower the range that you search in, the quicker the process will converge.

Resampling

Bootstrap Confidence Intervals

This option causes bootstrap confidence intervals and associated bootstrap reports and plots to be generated using resampling simulation as specified under the Resampling tab.

Bootstrapping may be time consuming when the bootstrap sample size is large. A reasonable strategy is to keep this option unchecked until you have considered all other reports. Then run this option with a bootstrap size of 100 or 1000 to obtain an idea of the time needed to complete the simulation.

Randomization Hypothesis Tests

This option hypothesis tests and associated reports to be generated using Monte Carlo simulation as specified under the Resampling tab.

Randomization tests may be time consuming when the Monte Carlo sample size is large. A reasonable strategy is to keep this option unchecked until you have run and considered all other reports. Then run this option with a Monte Carlo size of 100, then 1000, and then 10000 to obtain an idea of the time needed to complete the simulation.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

This section controls which reports and plots are displayed.

Select Reports

Combined Summary Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Combined Function Plot: Y ... Probability Plot: Trans(Y)

These options specify which plots are displayed.

Predicted Values

Predict Y at these X Values

Enter an optional list of X values at which to report the predicted value of Y and corresponding confidence interval. You can enter a single number or a list of numbers. The list can be separated with commas or spaces. The list can also be of the form ‘XX:YY(ZZ)’ which means XX to YY by ZZ.

Examples

10

10 20 30 40 50

0:90(10) which means 0 10 20 30 40 50 60 70 80 90

100:950(200) which means 100 300 500 700 900

1000:5000(500) which means 1000 1500 2000 2500 3000 3500 4000 4500 5000

Legend

Show Legend

Specify whether to display the plot legend when a Group Variable is used.

Legend Text

Specify the legend title. Note that {G} is replaced by the Group Variable name.

Report Options

Alpha Level

Enter the value of alpha for the confidence limits. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Value Labels

Value Labels may be used with the Group Variable to make reports more legible by assigning meaningful labels to numbers and codes.

- **Data Values**

  All data are displayed in their original format, regardless of whether a value label has been set or not.

- **Value Labels**

  All values of variables that have a value label variable designated are converted to their corresponding value label when they are output. This does not modify their value during computation.

- **Both**

  Both data value and value label are displayed.

Example

A variable named GENDER (used as a grouping variable) contains 1's and 2's. By specifying a value label for GENDER, the printout will display Male instead of 1 and Female instead of 2 on the reports. This option specifies whether (and how) to use the value labels.

Reminder

Value Labels are formed as two adjacent variables. The variable on the left contains the original values and the variable on the left contains the labels. A value label is assigned to a variable on the Variable Info sheet by designating the left variable of the pair as the Value Label variable.

Skip Line After

When writing a row of information to a report, some names and labels may be too long to fit in the space allocated. If the name (or label) contains more characters than this, the rest of the output for that line is moved down to the next line. Most reports are designed to hold a label of up to ‘15’ characters.

Enter ‘1’ when you always want each row’s output to by printed on two lines. Enter ‘100’ when you want each row printed on only one line. Note that this may cause some columns to be miss-aligned.

Report Options – Decimal Places

B ... SS & MS Decimals

Specify the number of decimal places used when displaying this item. Use ‘General’ to display the entire number without special formatting using the number of digits specified in the Precision box.

Function Plot and Residual Plot Tabs

This section controls the plot(s) showing the data with the fitted function line overlain on top and the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Display Prediction Limits (Function Plot)

This option controls whether the prediction limits (confidence limits on the predicted values) are displayed.

C.L. Line Width (Function Plot)

Specify the width of the confidence limit lines. Note that the color of this line is the color of the corresponding symbol as defined under the Symbols tab.

Function Line Width (Function Plot)

Specify the width of the function lines. Note that the color of this line is the color of the corresponding symbol as defined under the Symbols tab.

Number of Points (Function Plot)

Specify the number of points along the function at which it is evaluated for plotting. This affects the granularity of the line that represents the fitted function. Although valid values are from 20 to 2000, we recommend 200.

Symbol (Residual Plot)

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* are replaced by the name of the variable. Press the button on the right of the field to specify the font of the text.

Symbols Tab

These options control the shape, color and size of the symbols plotted on the function plot.

Plotting Symbols

Group 1 - 15

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color of points plotted on the function plot. Note that the color you specify will be used for the function line and confidence limits.

Resampling Tab

The following options control the bootstrapping and randomization tests.

Bootstrap Options – Sampling

Samples (N)

This is the number of bootstrap samples used. A general rule of thumb is that you use at least 100 when standard errors are your focus or at least 1000 when confidence intervals are your focus. If computing time is available, it does not hurt to do 10000.

We recommend setting this value to at least 3000.

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. We recommend setting the parameter to at least 50.

Bootstrap Options – Estimation

Percentile Type

The method used to create the percentiles when forming bootstrap confidence limits. You can read more about the various types of percentiles in the Descriptive Statistics chapter. We suggest you use the Ave *X*(p[n+1]) option.

C.I. Method

This option specifies the method used to calculate the bootstrap confidence intervals. The reflection method is recommended.

- **Percentile**

  The confidence limits are the corresponding percentiles of the bootstrap values.

- **Reflection**

  The confidence limits are formed by reflecting the percentile limits. If *X*0 is the original value of the parameter estimate and *XL* and *XU* are the percentile confidence limits, the Reflection interval is (2 *X*0 - *XU*, 2 *X*0 - *XL*).

Bootstrap Confidence Coefficients

These are the confidence coefficients of the bootstrap confidence intervals. Since bootstrapping calculations may take several minutes, it may be useful to obtain confidence intervals using several different confidence coefficients.

All values must be between 0.50 and 1.00. You may enter several values, separated by blanks or commas. A separate confidence interval is given for each value entered.

Examples

0.90 0.95 0.99

0.90:0.99(0.01)

0.90

Bootstrap Options – Histograms

Vertical and Horizontal Axis Labels

These are the labels of the vertical and horizontal axes of the bootstrap histograms.

Plot Style File

This is the histogram style file. We have provided several different style files to choose from, or you can create your own in the Histogram procedure.

Number of Bars

The number of bars shown in a bootstrap histogram. We recommend setting this value to at least 25 when the number of bootstrap samples is over 1000.

Histogram Title

This is the title used on the bootstrap histograms.

Randomization Test Options

Monte Carlo Samples

Specify the number of Monte Carlo samples used when running randomization tests. Somewhere between 1000 and 100000 are usually necessary. Although we use 1000 as the default value, a better value for routine use is 10000.

You also need to check the ‘Randomization Hypothesis Tests’ box on the Variables tab to run these tests.

Comparative Points

Specify the number of *X* values at which the difference between group curves is computed. This is the value of *K* in the formula given earlier. The sum of the absolute values of these differences is use in the randomization test of whether the group curves coincide.

Random Number Seed

Random Number Seed

This option specifies a random seed for the random number generator. Possible values are all integers between 1 and 32000. If you want to obtain the same results from one run to the next, use the same seed value. If you want to let the program select a random seed based on the time-of-day, enter ‘RANDOM SEED’.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Curve Fitting

This section presents an example of how to fit and compare a Michaelis-Menten model (model 8) to two groups of data. This example will use the data in the **FNREG5** database. In this example, the dependent variable is RESPONSE and the independent variable is TEMP. The groups are defined by the values of TYPE.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Curve Fitting – General window.

1 Open the FNREG5 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG5.S0**.
- Click **Open**.

2 Open the Curve Fitting – General window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **Curve Fitting** – **General**. The Curve Fitting – General procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- Select the **Variables tab**.
- Set the **Y Variable** to **RESPONSE**.
- Set the **X Variable** to **TEMP**.
- Set the **Group Variable** to **TYPE**.
- Set the **Preset Model** to **8 Y=AX/(B+X) Michaelis-Menten**.
- Check the **Bootstrap Confidence Intervals** box.
- Check the **Randomization Hypothesis Tests** box.

4 Specify the reports.

- Select the **Reports tab**.
- Check **all reports and plots except** the **Iteration Detail Report**.
- Set the **Predict Y at these X Values** to **5 10 15 20**.

5 Specify the resampling.

- Select the **Resampling tab**.
- Set **Samples (N)** to **200**. (We are using a small value for illustrative purposes. You should use at least 3000 when actually using the results.)
- Set **Monte Carlo Samples** to **200**. (We are using a small value for illustrative purposes. You should use at least 1000 when actually using the results.)
- Set **Random Number Seed** to **17448**. (Use this number so that our reports agree. Usually you would leave this set to ‘RANDOM START’.)

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Parameter Estimates for All Groups

```
Type        Count    Iter's        R2          A          B
1              21         4   0.98356   10.72798    4.95941
3              21         6   0.97645   10.31200    1.42325
Combined       42         4   0.81153   10.20315    2.54358
```

This report displays a summary of the results for each group and then for the case in which all groups are combined into one group.

Group Name (Type)

This column, headed by the name of the Group Variable, lists the group value that is displayed on this line. Note that the Value Labels option may be used to give more meaningful names to these values.

Count

This is the number of observations used by the nonlinear regression algorithm.

Iter's

This is the number of iterations used by the nonlinear regression algorithm to find the estimates. You should note whether the maximum number of iterations has been reached (in which case the algorithm did not converge).

R2

This is the value of the pseudo R-squared value. A value near one indicates that the model fits the data well. A value near zero indicates that the model does not fit the data well.

A B

The final values of the estimated parameters are displayed so that you may compare them across groups.

Analysis of Variance Across Groups

```
                                Model     Error   Sum Squares   Mean Square
Type        Count    Iter's        R2        DF         Error         Error
1              21         4   0.98356        19       1.73157       0.09114
3              21         6   0.97645        19       2.16427       0.11391
Combined       42         4   0.81153        40      43.74009       1.09350
Ignored                       0.98321        38       3.89585       0.10252
```

This report displays goodness of fit results for each group and then for the case in which all groups are combined into one dataset. The final row of the report, labeled 'Ignored', gives the goodness of fit statistics for the model in which a separate curve is fit for each group.

Group Name (Type)

This column, headed by the name of the Group Variable, lists the group value that is displayed on this line.

Count

This is the number of observations used by the nonlinear regression algorithm.

Iter's

This is the number of iterations used by the nonlinear regression algorithm to find the estimates. You should note whether the maximum number of iterations has been reached (in which case the algorithm did not converge).

R2

This is the value of the pseudo R-squared value. A value near one indicates that the model fits the data well. A value near zero indicates that the model does not fit the data well. Note

Error DF

The degrees of freedom are the number of observations minus the number of parameters fit.

Sum Squares Error

This is the sum of the squared residuals for this group.

Mean Square Error

This is a rough estimate of the variance of the residuals for this group.

Curve Inequality F-Test

```
Curves                                                  F-Test
Tested        DF       Mean Square      F Ratio     Prob Level
All            2          19.92212     194.3200        0.00000
Error         38           0.10252
```

This report displays an F-Test of whether all of the group curves are equal. This test compares the residual sum of squares obtained when the grouping is ignored with the total of the residual sum of squares obtained for each group. This test is routinely used in analysis linear models and its application to nonlinear models has occasionally been suggested. However, it is based on normality assumptions which seldom occur. When testing curve coincidence is important, we suggest you use a randomization test.

Curves Tested

This column indicates the term presented on this row.

DF

The degrees of freedom of this term.

Mean Square

The mean square associated with this term.

F Ratio

The F-ratio for testing the hypothesis that all curves coincide.

F-Test Prob Level

This is the probability level of the F-ratio. When this value is less than 0.05 (a common value for alpha), the test is 'significant' meaning that the hypothesis of equal curves is rejected. If this value is larger than the nominal level (0.05), the null hypothesis cannot be rejected. We do not have enough evidence to reject.

Curve Inequality Randomization Tests

```
                                                    Number of Points
Curves        Randomization     Monte Carlo            Compared
Tested           Prob Level         Samples         Along Curve
1 vs. 3             0.00000             200                  10
```

This report displays the results of a randomization test whose null hypothesis is that the all the group curves coincide. When more than two groups are present, a separate test is provided for each pair of groups, plus a combined test of the equality of all groups.

Curves Tested

This column indicates the groups whose equality is being test on this row.

Randomization Prob Level

This is the two-sided probability level of the randomization test. When this value is less than 0.05, the test is ‘significant’ meaning that the null hypothesis of equal curves is rejected. If this value is larger than the nominal level (0.05), there is not enough evidence in the data to reject the null hypothesis of equality.

(Note: because this is a Monte Carlo test, your results may vary from those displayed here.)

Monte Carlo Samples

The number of Monte Carlo samples.

Number of Points Compared Along the Curve

The number of values along the X axis at which a comparison between curves is made. Of course, the more X values used, the more accurate (and time consuming) will be the test.

Parameter Inequality Randomization Tests

```
Curves            Parameter     Randomization        Monte Carlo
Compared             Tested        Prob Level         Iterations
1 vs. 3                   A           0.74500                200
1 vs. 3                   B           0.03500                200
```

This report displays the results of randomization tests about the equality of each parameter across groups. When more than two groups are present, a separate test is provided for each pair of groups, plus a combined test of parameter equality of all groups.

Curves Compared

This column indicates the groups being test on this row.

Parameter Test

This column indicates model parameter whose equality is being tested.

Randomization Prob Level

This is the two-sided probability level of the randomization test. When this value is less than 0.05, the test is ‘significant’ meaning that the null hypothesis of equal parameter values across groups is rejected. If this value is larger than the nominal level (0.05), there is not enough evidence in the data to reject the null hypothesis of equality.

(Note: because this is a Monte Carlo test, your results may vary from those displayed here.)

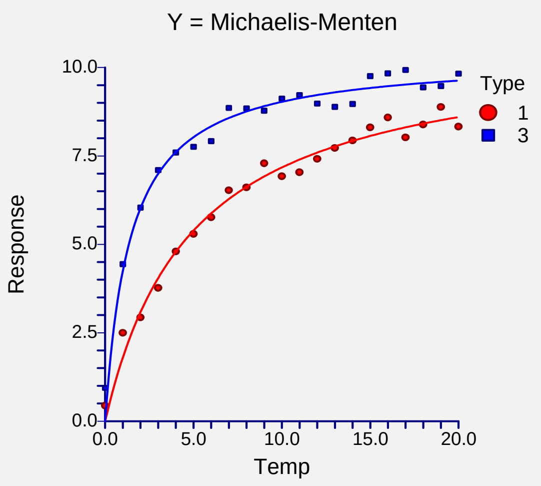
Monte Carlo Samples

The number of Monte Carlo samples.

Number of Points Compared Along the Curve

The number of values along the X axis at which a comparison between curves is made. Of course, the more X values used, the more accurate (and time consuming) will be the test.

Combined Plot Section



This plot displays all of the data and fitted curves, allowing you to quickly assess the quality of the results.

Iteration Summary Section for Type=1

Itn No.	Residual Sum of Squares	A	B
1	1.81547	10.51692	4.58046
2	1.73188	10.71254	4.93394
3	1.73157	10.72751	4.95871
4	1.73157	10.72798	4.95941

This report displays the progress of the search algorithm in its search for a solution. It allows you to assess whether the algorithm had indeed converged or whether the program should be re-run with the Maximum Iterations increased or the model changed.

Note that if over ten iterations were needed, the program does not display every iteration.

Model Estimation Section for Type = 1

```
Parameter     Parameter       Asymptotic          Lower          Upper
Name           Estimate   Standard Error       95% C.L.       95% C.L.
A              10.72798          0.30895       10.08135       11.37461
B               4.95941          0.44270        4.03282        5.88599

Iterations            4        Rows Read             21
R-Squared      0.983564        Rows Used             21
Random Seed       17448      Total Count             21

Estimated Model
(10.7279796048293)*(x)/((4.95940560335216)+(x))
```

This report displays the details of the estimation of the model parameters.

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit. In most cases, the bootstrap confidence interval will be more accurate.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit. In most cases, the bootstrap confidence interval will be more accurate.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

R-Squared

There is no direct R-squared defined for nonlinear regression. This is a pseudo R-squared constructed to approximate the usual R-squared value used in multiple regression. We use the following generalization of the usual R-squared formula:

$$\textit{R-Squared} = (\textit{ModelSS} - \textit{MeanSS})/(\textit{TotalSS}-\textit{MeanSS})$$

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-squared that you use in multiple regression, it will serve well for comparative purposes.

Random Seed

This is the value of the random seed that was used when running the bootstrap confidence intervals and randomization tests. If you want to duplicate your results exactly, enter this random seed into the Random Seed box under the Simulation tab.

Estimated Model

This is the model that was estimated with the parameters replaced with their estimated values. This expression may be copied and pasted as a variable transformation in the spreadsheet. This will allow you to predict for additional values of X. Note that to insure accuracy, the parameter estimates are always given to double-precision accuracy.

Analysis of Variance Table for Type = 1

```
                                         Sum of           Mean
Source                  DF              Squares         Square
Mean                     1            847.88494      847.88494
Model                    2            951.50296      475.75148
Model (Adjusted)         1            103.61802      103.61802
Error                   19              1.73157        0.09114
Total (Adjusted)        20            105.34959
Total                   21            953.23453
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just “SSE.”

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Bootstrap Section

```
---  Estimation Results    ------    | --- Bootstrap Confidence Limits----
Parameter                 Estimate   | Conf. Level Lower              Upper
Intercept
A
Original Value            10.72798   | 0.90000   10.21652          11.26251
Bootstrap Mean            10.73831   | 0.95000   10.11747          11.32328
Bias (BM - OV)             0.01033   | 0.99000    9.81792          11.47991
Bias Corrected            10.71765
Standard Error             0.30969
B
Original Value             4.95941   | 0.90000    4.30466           5.70459
Bootstrap Mean             4.97616   | 0.95000    4.07184           5.90947
Bias (BM - OV)             0.01676   | 0.99000    3.87871           6.09763
Bias Corrected             4.94265
Standard Error             0.43834

Predicted Mean and Confidence Limits of Response When Temp = 5.00000
Original Value             5.38585   | 0.90000    5.21330           5.53273
Bootstrap Mean             5.38588   | 0.95000    5.15965           5.54946
Bias (BM - OV)             0.00003   | 0.99000    5.06827           5.58565
Bias Corrected             5.38582
Standard Error             0.09954
Predicted Value and Confidence Limits of Response When Temp = 5.00000
Original Value             5.38585   | 0.90000    4.76670           5.77922
Bootstrap Mean             5.41128   | 0.95000    4.70845           5.83435
Bias (BM - OV)             0.02542   | 0.99000    4.53484           5.89363
Bias Corrected             5.36043
Standard Error             0.30921

(Report continues for the other values of Temp)

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.
```

This report provides bootstrap estimates and confidence intervals for the parameters, predicted means, and predicted values. Note that bootstrap confidence intervals and prediction intervals are provided for each of the *X* (Temp) value requested. Details of the bootstrap method were presented earlier in this chapter.

Original Value

This is the parameter estimate obtained from the complete sample without bootstrapping.

Bootstrap Mean

This is the average of the parameter estimates of the bootstrap samples.

Bias (BM - OV)

This is an estimate of the bias in the original estimate. It is computed by subtracting the original value from the bootstrap mean.

Bias Corrected

This is an estimated of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate.

Standard Error

This is the bootstrap method's estimate of the standard error of the parameter estimate. It is simply the standard deviation of the parameter estimate computed from the bootstrap estimates.

Conf. Level

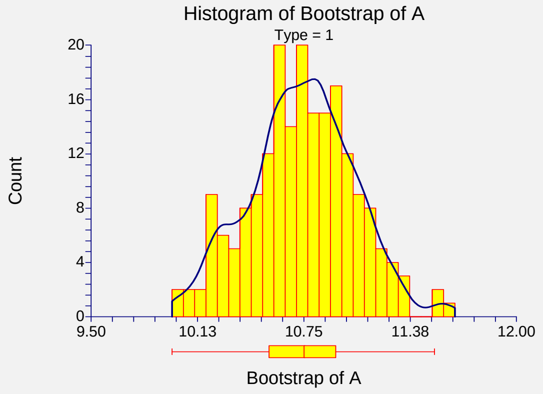
This is the confidence coefficient of the bootstrap confidence interval given to the right.

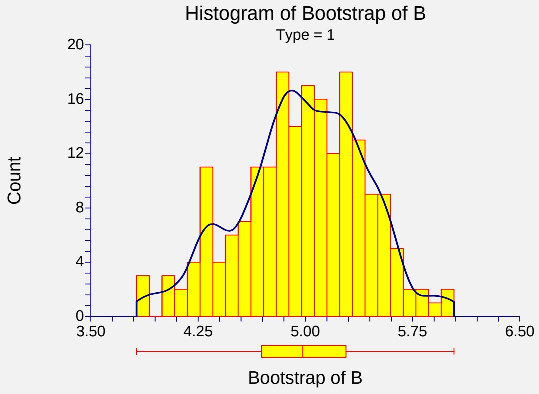
Bootstrap Confidence Limits - Lower and Upper

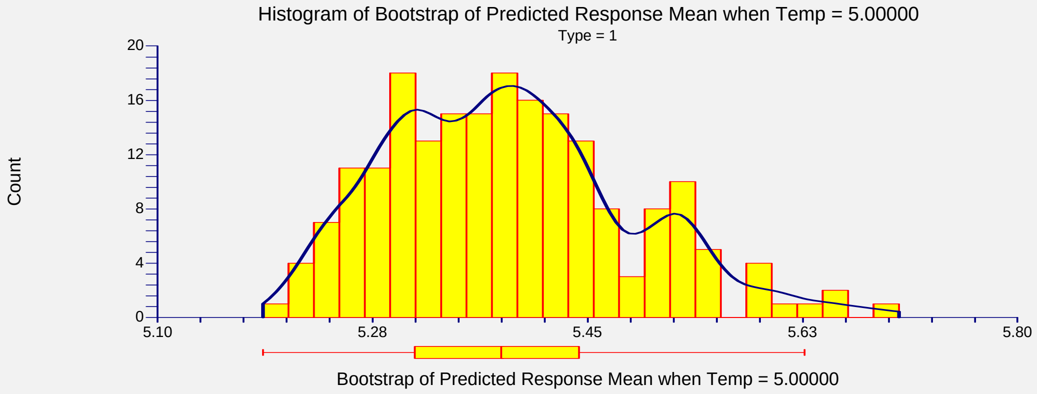
These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

Bootstrap Histograms Section







(Several more histograms are displayed.)

Each histogram shows the distribution of the corresponding estimate.

Note that the number of decimal places shown in the horizontal axis is controlled by which histogram style file is selected. In this example, we selected Bootstrap2 which was created to provide two decimal places.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters for Type = 1
        A                   B
A       1.000000            0.940484
B       0.940484            1.000000
```

This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.98), the precision of the parameter estimates is suspect.

Predicted Values for Specified X Values for Type=1

```
                 Predicted       Lower 95.0%       Upper 95.0%
                  Value of        Prediction        Prediction
Temp              Response             Limit             Limit
5.00000            5.38585           4.71548           6.05623
10.00000           7.17139           6.52162           7.82116
15.00000           8.06235           7.40400           8.72069
20.00000           8.59634           7.91914           9.27355
```

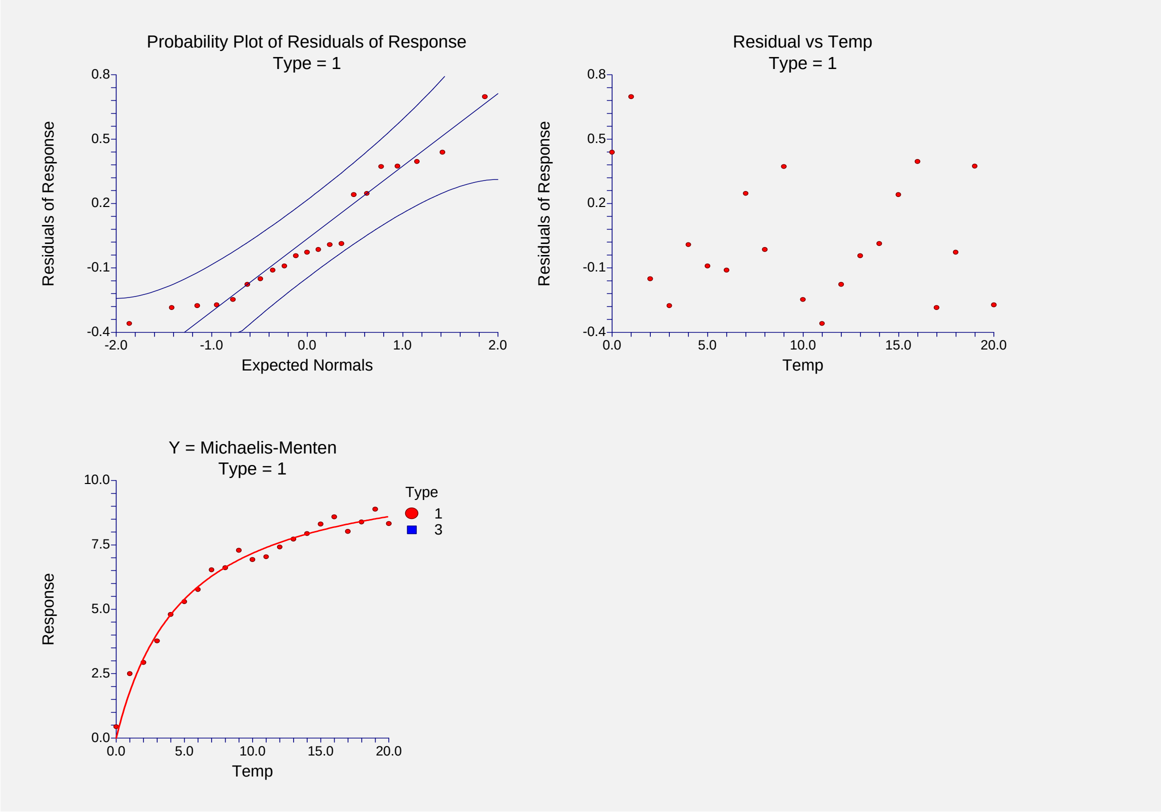
This section shows the predicted mean values and asymptotic (large sample) prediction intervals for the X values that were specified. Note that these are prediction limits for a new value, not confidence limits for the mean of the values.

Predicted Values and Residuals Section

```
                                                Lower 95.0%  Upper 95.0%
Row                                  Predicted   Prediction   Prediction
No.        Temp       Response          Value         Limit        Limit       Residual
1          0.00000     0.43846        0.00000      -0.63186      0.63186        0.43846
2          1.00000     2.49732        1.80018       1.14295      2.45740        0.69714
3          2.00000     2.93207        3.08302       2.40603      3.76000       -0.15094
4          3.00000     3.76707        4.04351       3.36238      4.72464       -0.27644
5          4.00000     4.79763        4.78959       4.11244      5.46675        0.00803
6          5.00000     5.29474        5.38585       4.71548      6.05623       -0.09111
.                .           .              .             .            .              .
.                .           .              .             .            .              .
.                .           .              .             .            .              .
```

This section shows the values of the predicted values, prediction limits, and residuals. If you have observations in which the independent variable is given, but the dependent (Y) variable is blank, a predicted value and prediction limits will be generated and displayed in this report.

Plots



Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefinition of the model.

Function Plot

This plot displays the data along with the estimated function. It is useful in deciding if the fit is adequate and the prediction limits are appropriate.

Chapter 360

Growth and Other Models

Introduction

Models for the study of growth have been studied for a long time. Historically, these models have been divided into two classifications: empirical and mechanistic. An empirical model is one that was developed to be flexible enough to fit many sets of data well. However, its parameters usually do not have direct interpretation in terms of the physical process being modeled. A mechanistic model is derived from the mathematics of the physical process producing the data. Recently, the differentiation between these two classes of models has narrowed. Mechanistic models are usually based on overly simplistic assumptions and some would argue that they are really empirical.

Because of these two competing classifications, many mathematical models have been developed that have similar shapes and characteristics. Often, the selection of a model is arbitrary and several of the available curves will do an excellent job of fitting the data.

This program provides thirteen growth models and an additional eight miscellaneous models for use in fitting data. The parameters of these models are estimated using the Levenberg-Marquardt nonlinear least-squares algorithm as presented in Nash (1987).

Starting Values

Starting values may be provided. If you want, the procedure will calculate reasonable starting values from your data.

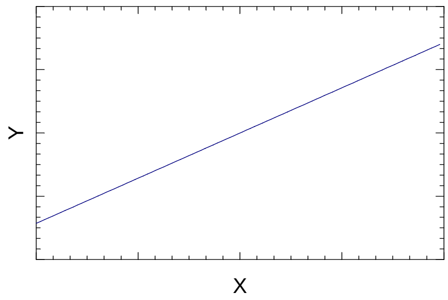
List of Models

The models available provide a variety of shapes and forms. Several of the models find their roots in quite different physical processes, but yield similar results. We will now present examples and details of each of the twenty-one models available. Note that you will often only need part of the curve.

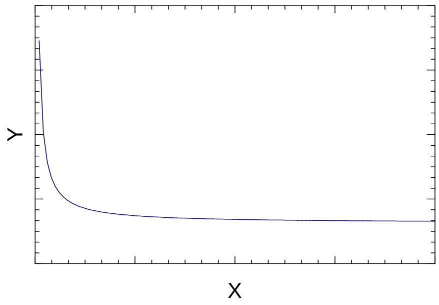
1. Linear: Y=A+BX

This common model is usually fit using standard linear regression techniques. We include it here to allow for various special forms made by transforming X and Y

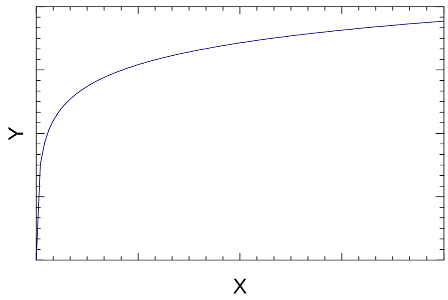
Plot of Y = 1+X



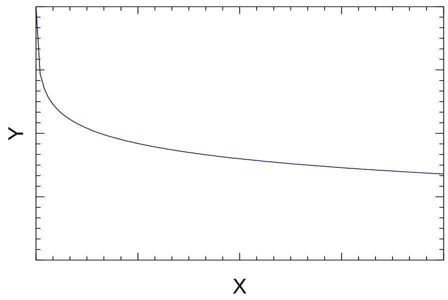
Plot of Y = 1+1/X



Plot of Y = 1+LOG(X)



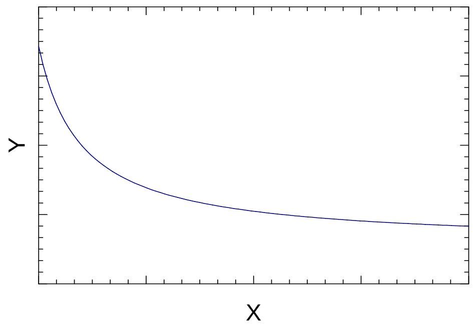
Plot of Y = 1-LOG(X)



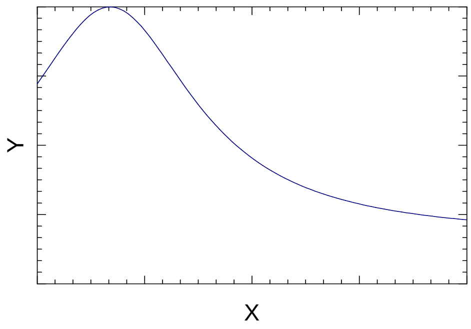
2. Reciprocal: Y=1/(A+BX)

This model, known as the reciprocal or Shinozaki and Kira model, is mentioned in Ratkowsky (1989, page 89) and Seber (1989, page 362).

Plot of Y = 1/(1+X)



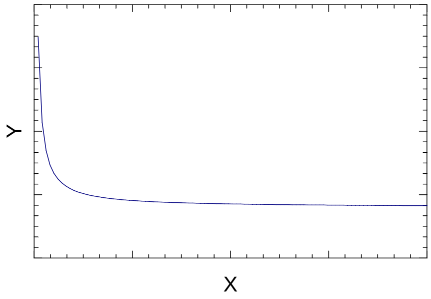
Plot of Y = 1/(4+2*X^2)



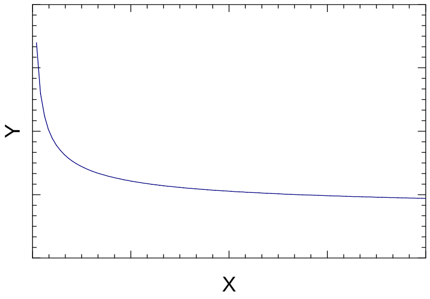
3. Bleasdale-Nelder: Y=(A+BX)^(-1/C)

This model, known as the Bleasdale-Nelder model, is mentioned in Ratkowsky (1989, page 103) and Seber (1989, page 362).

Plot of Y = (1+X)^(-1)



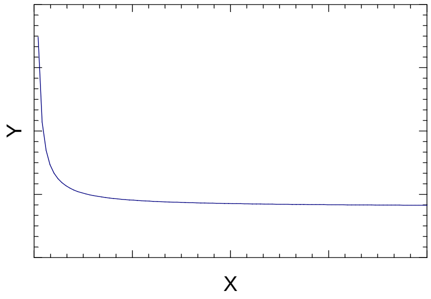
Plot of Y = (1+X)^(-1/2)



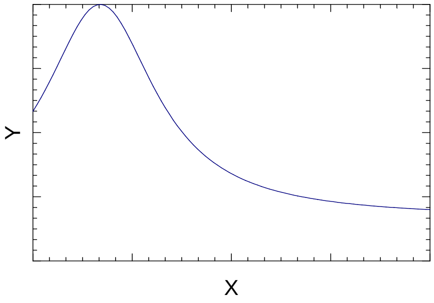
4. Farazdaghi and Harris: Y=1/(A+BX^C)

This model, known as the Farazdaghi and Harris model, is mentioned in Ratkowsky (1989, pages 99 and 104) and Seber (1989, page 362).

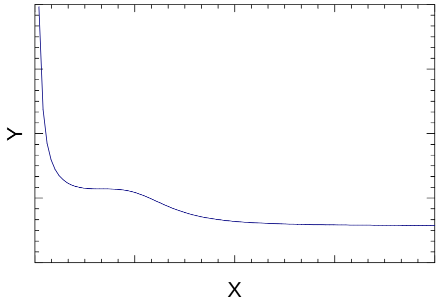
Plot of Y = 1/(1+X^1)



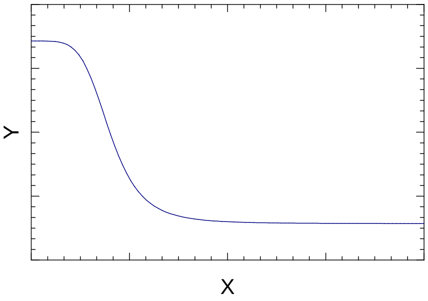
Plot of Y = 1/(1+X^2)



Plot of Y = 1/(1+X^3)

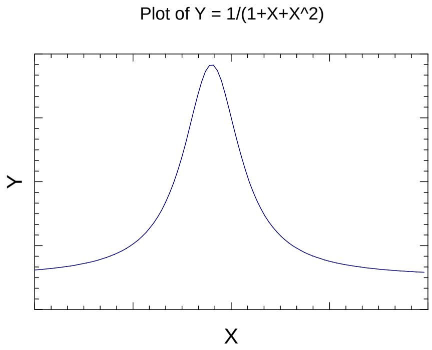


Plot of Y = 1/(1+X^5)



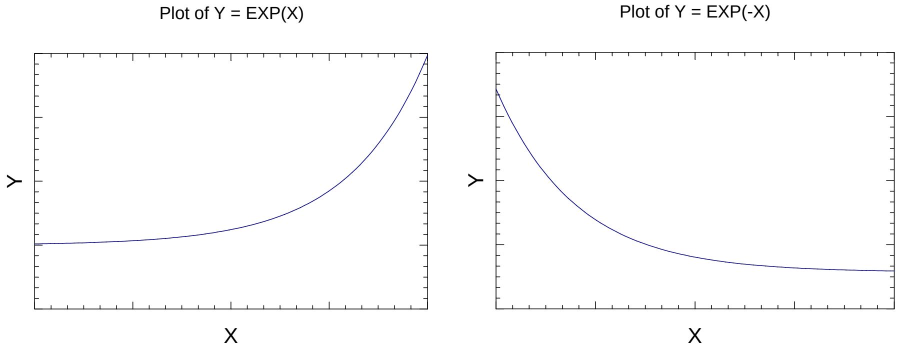
5. Holliday: Y=1/(A+BX+CX^2)

This model, known as the Holliday model, is mentioned in Seber (1989, page 362).



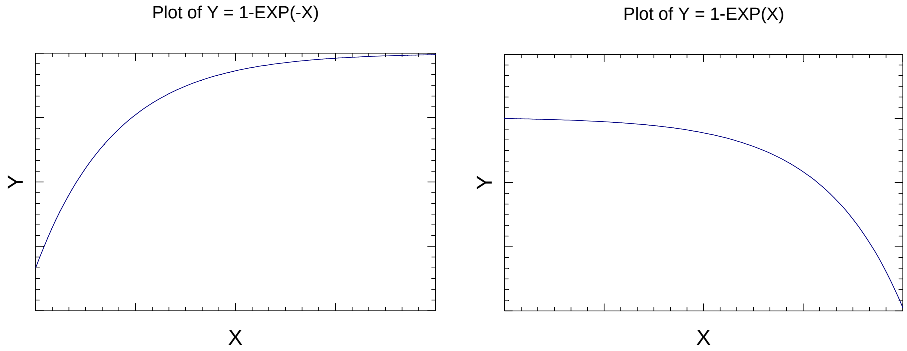
6. Exponential: Y=EXP(A(X-B))

This model, known as the exponential model, is mentioned in Seber (1989, page 327). Note that taking the log of both sides reduces this equation to a linear model.



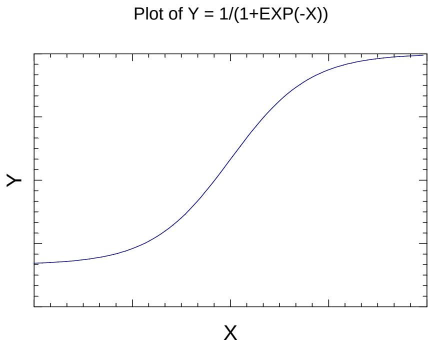
7. Monomolecular: Y=A(1-EXP(-B(X-C)))

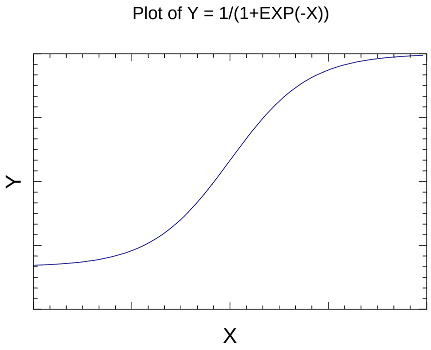
This model, known as the monomolecular model, is mentioned in Seber (1989, page 328).



8. Three Parameter Logistic: Y=A/(1+B(EXP(-CX)))

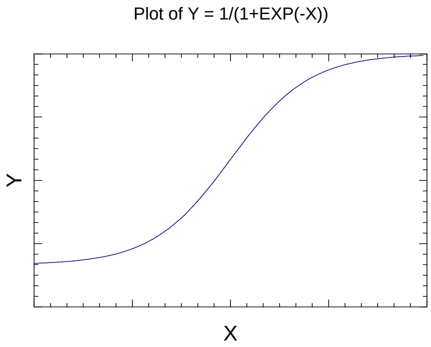
This model, known as the three-parameter logistic model, is mentioned in Seber (1989, page 330).

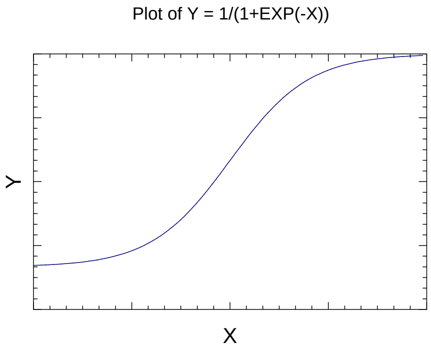

Plot of Y = 1/(1+EXP(-X))


Plot of Y = 1/(1+EXP(-X))

9. Four Parameter Logistic: Y=D+(A-D)/(1+B(EXP(-CX)))

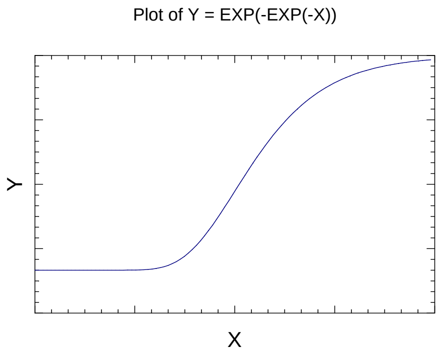
This model, known as the four-parameter logistic model, is mentioned in Seber (1989, page 338). Note that the extra parameter, D, has the effect of shifting the graph vertically. Otherwise, this plot is the same as the three-parameter logistic.

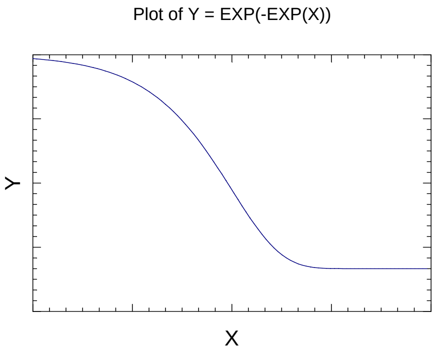

Plot of Y = 1/(1+EXP(-X))


Plot of Y = 1/(1+EXP(-X))

10. Gompertz: Y=A(EXP(-EXP(-B(X-C))))

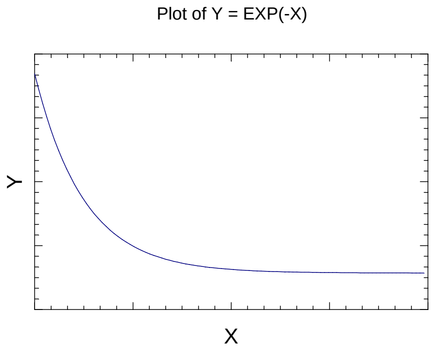
This model, known as the Gompertz model, is mentioned in Seber (1989, page 331).

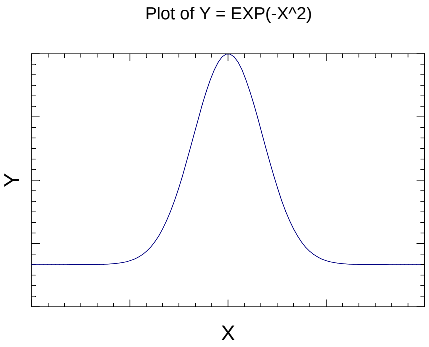

Plot of Y = EXP(-EXP(-X))

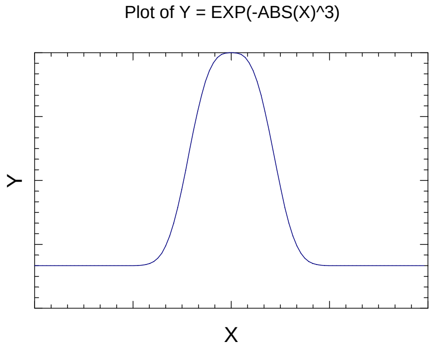

Plot of Y = EXP(-EXP(X))

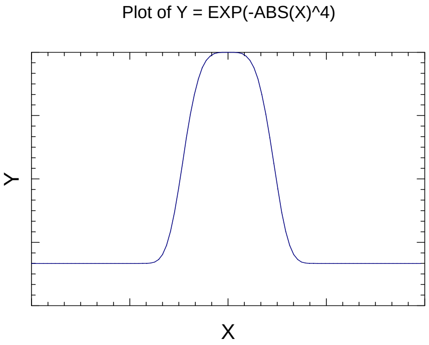
11. Weibull: Y=A-(A-B)EXP(-(C|X|)^D)

This model, known as the Weibull model, is mentioned in Seber (1989, page 338).


Plot of Y = EXP(-X)

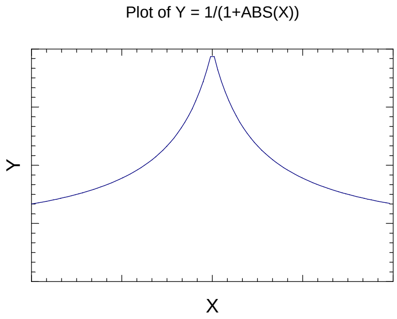

Plot of Y = EXP(-X^2)

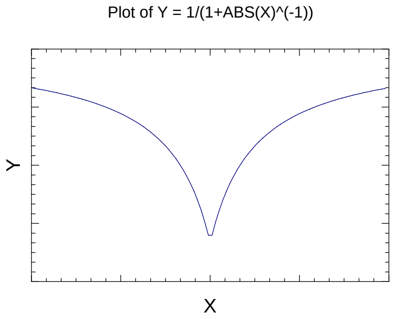

Plot of Y = EXP(-ABS(X)^3)

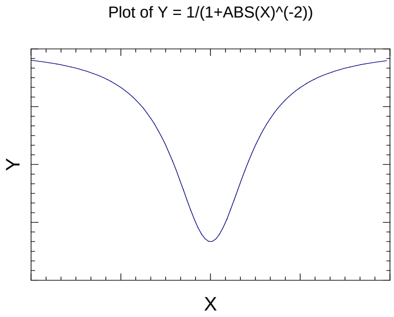

Plot of Y = EXP(-ABS(X)^4)

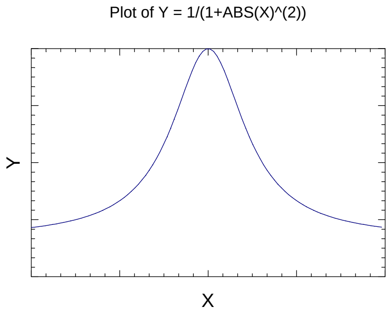
12. Morgan-Mercer-Floding: Y=A-(A-B)/(1+(C|X|)^D)

This model, known as the Morgan-Mercer-Floding model, is mentioned in Seber (1989, page 340).


Plot of Y = 1/(1+ABS(X))

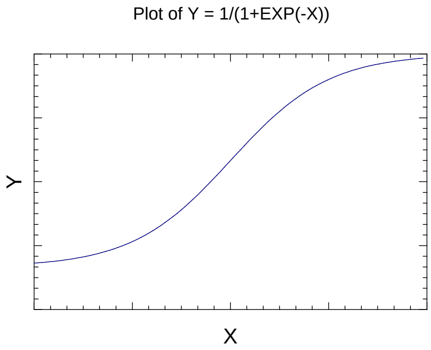

Plot of Y = 1/(1+ABS(X)^(-1))

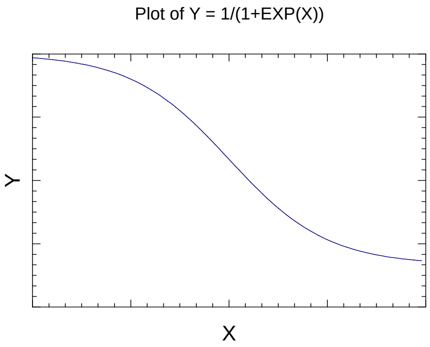

Plot of Y = 1/(1+ABS(X)^(-2))


Plot of Y = 1/(1+ABS(X)^(2))

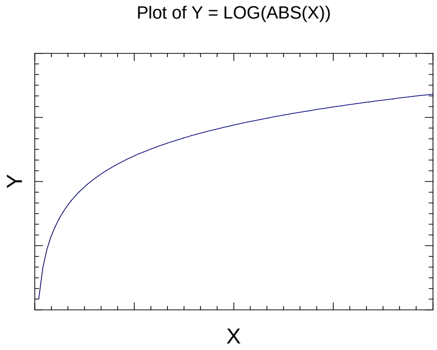
13. Richards: Y=A(1+(B-1)EXP(-C(X-D)))^(1/(1-B))

This model, known as the Richards model, is mentioned in Seber (1989, page 333).

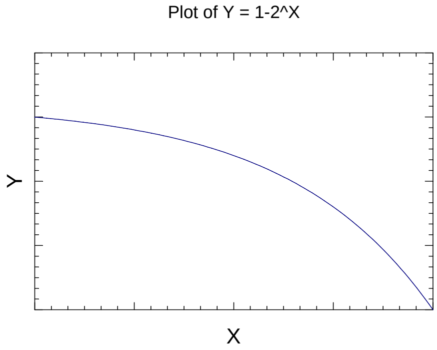




14. Y=B(LN(|X|-A))

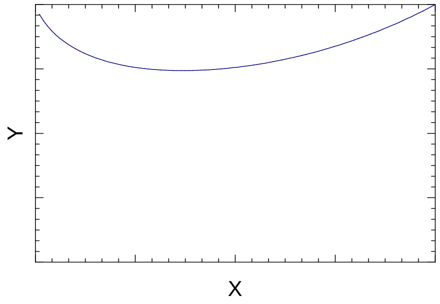


15. Y=A(1-B^X)

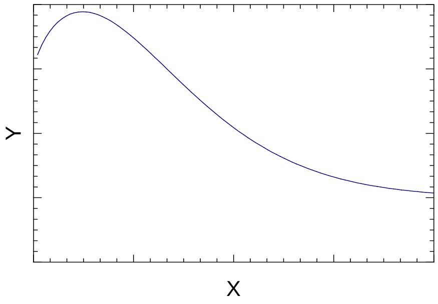


16. Y=AX^(BX^C)

Plot of Y = X^X

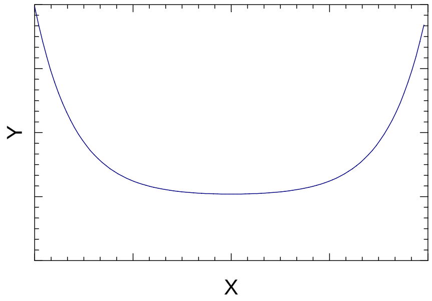


Plot of Y = X^(-X)

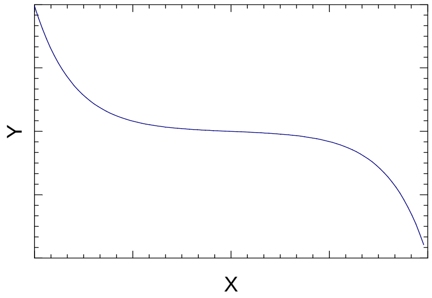


17. Sum of Exponentials: Y=A(EXP(-BX))+C(EXP(-DX))

Plot of Y = EXP(-X)+EXP(X)

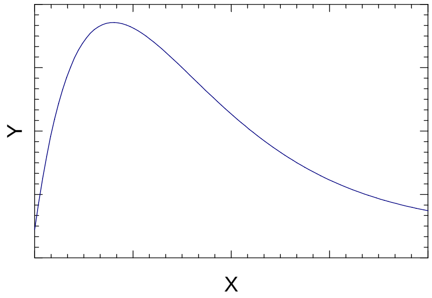


Plot of Y = EXP(-X)-EXP(X)

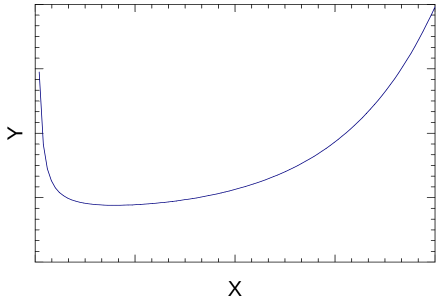


18. Y=A(X^B)EXP(-CX)

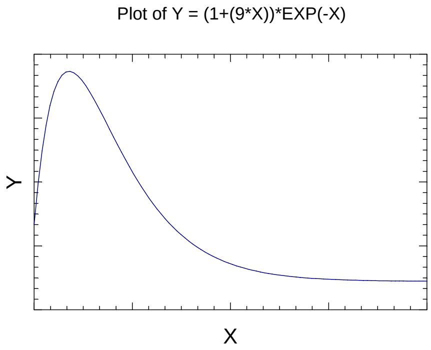
Plot of Y = X*EXP(-X)



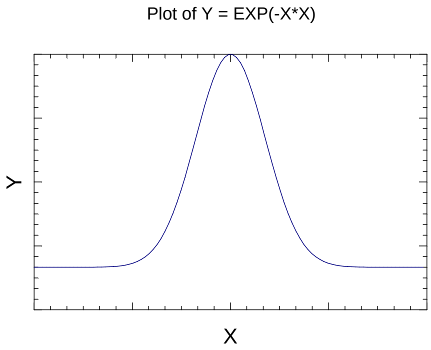
Plot of Y = 1/X*EXP(X)



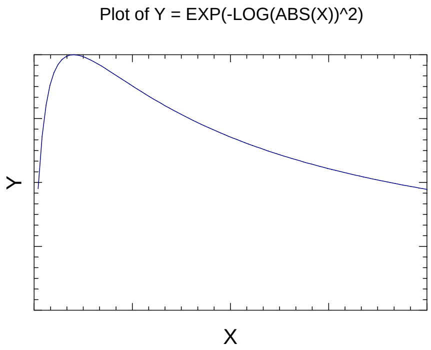
19. Y=(A+BX)EXP(-CX)+D

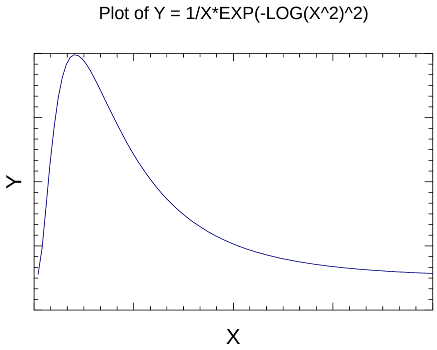


20. Normal: Y=A+B(EXP(-C(X-D)^2))



21. Lognormal: Y=A+(B/X)EXP(-C(LN(X)-D)^2)





Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two variables: one dependent variable and one independent variable.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable from the current database.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y’=1/(Y*Y), ***Y’=1/Y***, ***Y’=1/SQRT(Y)***, ***Y’=LN(Y)***, ***Y’=SQRT(Y)***, ***Y’=Y (none)***, and ***Y’=Y*Y***

X (Independent) Variable

Specifies a single independent (*X*) variable from the current database.

X Trans

Specifies a power transformation of the independent variable. Available transformations are

X’=1/(X*X), ***X’=1/X***, ***X’=1/SQRT(X)***, ***X’=LN(X)***, ***X’=SQRT(X)***, ***X’=X (none)***, and ***X’=X*X***

Model

Model Type

Specify the model that you want to fit. The available models are:

1	Y=A+BX	Linear
2	Y=1/(A+BX)	Reciprocal
3	Y=(A+BX)^(-1/C)	Bleasdale-Nelder
4	Y=1/(A+BX^C)	Farazdaghi and Harris
5	Y=1/(A+BX+CX^2)	Holliday
6	Y=EXP(A(X-B))	Exponential
7	Y=A(1-EXP(-B(X-C)))	Monomolecular
8	Y=A/(1+B(EXP(-CX)))	Three Parameter Logistic

<table>
<tr><td>9</td><td>Y=D+(A-D)/(1+B(EXP(-CX)))</td><td>Four Parameter Logistic</td></tr>
<tr><td>10</td><td>Y=A(EXP(-EXP(-B(X-C))))</td><td>Gompertz</td></tr>
<tr><td>11</td><td>Y=A-(A-B)EXP(-(C|X|)^D)</td><td>Weibull</td></tr>
<tr><td>12</td><td>Y=A-(A-B)/(1+(C|X|)^D)</td><td>Morgan-Mercer-Floding</td></tr>
<tr><td>13</td><td>Y=A(1+(B-1)EXP(-C(X-D)))^(1/(1-B))</td><td>Richards</td></tr>
<tr><td>14</td><td>Y=B(LN(|X|-A))</td><td></td></tr>
<tr><td>15</td><td>Y=A(1-B^X)</td><td></td></tr>
<tr><td>16</td><td>Y=AX^(BX^C)</td><td></td></tr>
<tr><td>17</td><td>Y=A(EXP(-BX))+C(EXP(-DX))</td><td>Sum of Exponentials</td></tr>
<tr><td>18</td><td>Y=A(X^B)EXP(-CX)</td><td></td></tr>
<tr><td>19</td><td>Y=(A+BX)EXP(-CX)+D</td><td></td></tr>
<tr><td>20</td><td>Y=A+B(EXP(-C(X-D)^2))</td><td>Normal</td></tr>
<tr><td>21</td><td>Y=A+(B/X)EXP(-C(LN(X)-D)^2)</td><td>Lognormal</td></tr>
</table>

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias and the bias correction procedures used.

Model – Parameter Starting Values

A ... D

This is the beginning value of the parameter during the search procedure. If left blank, the program will calculate starting values based on simple formulas that use only a few observations. If the calculated starting values do not converge, you will have to enter your own (or try a different model).

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Iteration Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Function Plot with Actual Y ... Probability Plot with Transformed Y

These options specify which plots are displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates and predicted values. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Function Plot and Residual Plot Tabs

This section controls the plot(s) showing the data with the fitted function line overlain on top and the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Line (Function Plot)

This option controls the attributes of the line representing the fitted function. Click this box to bring up the line specification dialog box. This window will let you set the line pattern, size, and color.

Display Prediction Limits (Function Plot)

This option controls whether the prediction limits (confidence limits on the predicted values) are displayed.

Number of Points (Function Plot)

This option specifies at how many points the estimated function is calculated to create the overlay function that is displayed on the Function Plots. A value between 50 and 150 is usually sufficient.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{M}* are replaced by the name of the variable and the model expression, respectively. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Growth Model

This section presents an example of how to fit a growth model. In this example, we will fit the three-parameter logistic growth model (model 8) to variables Y1 and X1 of the FNREG1 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Growth and Other Models window.

1 Open the FNREG1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.

- Click on the file **FNREG1.S0**.
- Click **Open**.

2 Open the Growth and Other Models window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **Growth and Other Models**. The Growth and Other Models procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Growth and Other Models window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Y1** from the list of variables and then click **Ok**.
- Double-click in the **X Variable** box. This will bring up the variable selection window.
- Select **X1** from the list of variables and then click **Ok**.
- Select **8 Y=A/(1+B(EXP(-CX))) 3 Parameter Logistic** in the Model Type list box.

4 Specify the reports.

- On the Growth and Other Models window, select the **Reports tab**.
- Check the **Residual Report** box. Leave all other reports and plots checked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section
Itn      Error Sum
No.      Lambda              Lambda          A               B               C
0        40.60853            0.00004         0               1               0
1        17.50267            0.000016        0.961331        1               0
2        2.502684            0.0000064       0.9613522       0.9999608       0.5602523
3        1.460298            2.56E-06        0.8592459       0.6802145       1.055472
4        0.9565874           1.024E-06       0.9553283       0.8540083       1.291937
5        0.9563748           4.096E-07       0.955538        0.8655906       1.284907
6        0.9563742           1.6384E-07      0.9556665       0.8664535       1.284215
7        0.9563742           0.65536         0.9556671       0.8664643       1.284214
Convergence criterion met.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm's progress.

Model Estimation Section

```
Model Estimation Section

Parameter           Parameter      Asymptotic       Lower          Upper
Name                Estimate       Standard Error   95% C.L.       95% C.L.
A                   0.9556671      2.056039E-02     0.9148604      0.9964738
B                   0.8664643      0.1013374        0.6653376      1.067591
C                   1.284214       0.1215541        1.042962       1.525465

Dependent           Y1
Independent         X1
Model               Y1=A/(1+B*EXP{-C*(X1)})
R-Squared           0.945358
Iterations          7
Estimated Model
(.9556671)/(1+(.8664643)*EXP{-(1.284214)*(X1)})
```

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Model

The model that was estimated. Use this to double check that the model estimated was what you wanted.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

R-Squared = (ModelSS - MeanSS)/(TotalSS-MeanSS)

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

Estimated Model

The model that was estimated with the parameters replaced with their estimated values. This expression may be copied and pasted as a variable transformation in the spreadsheet. This will allow you to predict for additional values of X.

Analysis of Variance Table

```
Analysis of Variance Table
                                    Sum of          Mean
Source              DF              Squares         Square
Mean                1               23.10585        23.10585
Model               3               39.65216        13.21739
Model (Adjusted)    2               16.5463         8.27315
Error               97              0.9563742       9.859528E-03
Total (Adjusted)    99              17.50267
Total               100             40.60853
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just “SSE.”

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters
        A              B              C
A       1.000000       0.611455       -0.518521
B       0.611455       1.000000       -0.401338
C       -0.518521      -0.401338      1.000000
```

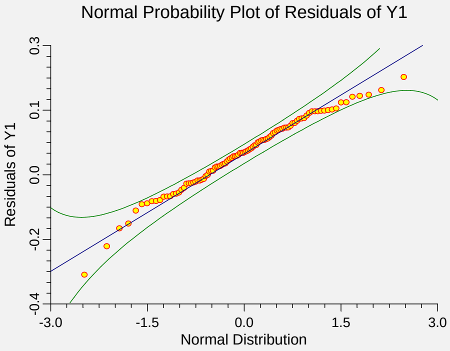
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

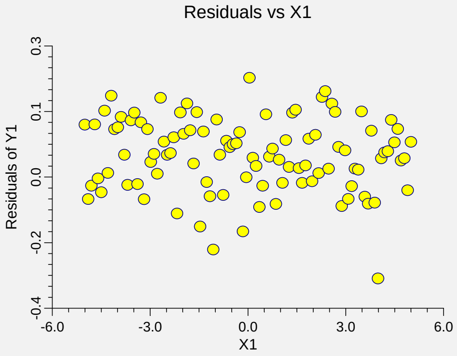
Predicted Values and Residuals Section

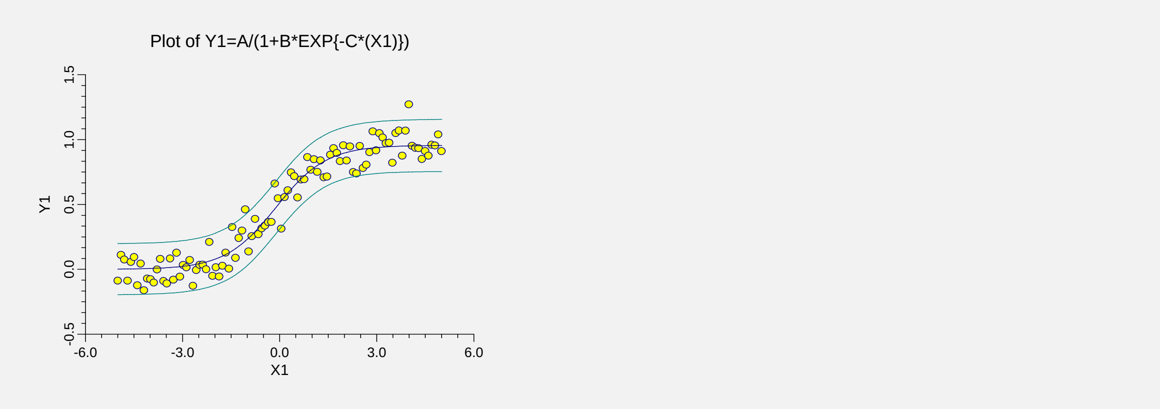
```
Predicted Values and Residuals Section
Row                                Predicted       Lower 95.0%    Upper 95.0%
No.   X1           Y1              Value           Value          Value         Residual
1     -5           -8.793363E-02   1.791046E-03    -0.1952929     0.198875      -8.972467E-02
2     -4.89899     0.1107571       2.038591E-03    -0.1950479     0.1991251     0.1087185
3     -4.79798     7.545952E-02    2.320266E-03    -0.1947694     0.1994099     7.313926E-02
4     -4.69697     -8.788628E-02   2.640754E-03    -0.1944527     0.1997343     -9.052704E-02
.     .            .               .               .
.     .            .               .               .
.     .            .               .               .
```

This section shows the values of the residuals and predicted values. If you have observations in which the independent variable is given, but the dependent (Y) variable is blank, a predicted value and prediction limits will be generated and displayed in this report.

Residual Plots







Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

Function Plot

This plot displays the data along with the estimated function and prediction limits. It is useful in deciding if the fit is adequate and the prediction limits are appropriate.

In poorly fit models, we have found that it is often necessary to disable the prediction limits so that the data will show up. In these cases, the prediction limits may be so wide that the scale of the plot does not allow the data values to be separated.

Predicting for New Values

You can use your model to predict Y for new values of X. Here’s how. Add new rows to the bottom of your database containing the values of the independent variable that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.

Chapter 365

Piecewise Polynomial Models

Introduction

Piecewise-polynomial models, sometimes referred to as *multiphase models* (see Seber [1989] chapter 9 for a complete discussion), are constructed by combining straight lines and quadratics. For example, the multiphase model linear-linear refers to a model made up of two linear equations, each active over a different range of X.

This program fits five types of multiphase models. A special feature is that you do not have to enter the change (connecting) point. The algorithm calculates change points for you.

The change points in multiphase models are often sharp corners. The program allows the use of two types of smoothing functions to smooth out these sharp corners.

Starting Values

Starting values may be provided. If you want, the procedure will calculate reasonable starting values from your data.

List of Models

The models available provide a variety of shapes and forms. Several of the models find their roots in quite different physical processes, but yield similar results. We will now present examples and details of each of the five models available. Note that you will often need only part of the curve.

1. Linear-Linear

Estimation Equation

Y=A + BX + C(X-D)SIGN(X-D)

Common Equation

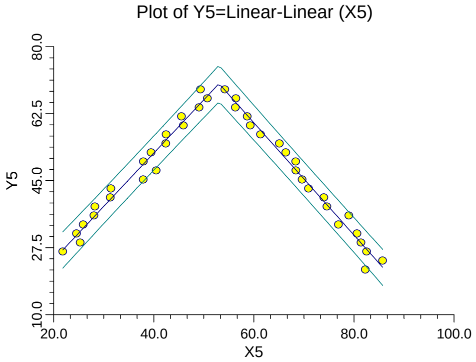
Y = a1 + b1X, X<J

Y = a2 + b2X, X³J

Parameter Identities

A=(a1+a2)/2 B=(b1+b2)/2 a1=A+DC b1=B-C J=D

C=(b2-b1)/2 D=J a2=A-DC b2=B+C



2. Linear-Quadratic

Estimation Equation

Y=A+BX+CX^2+(X-D)SIGN(X-D)[C(X+D)+E]

Common Equation

Y=a1+b1X, X<=a

Y=a2+b2X+c2X^2, X>a

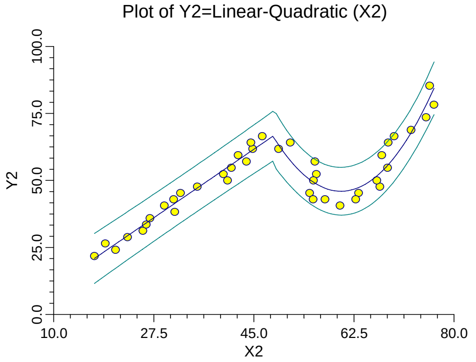
Parameter Identities

A=(a1+a2)/2 B=(b1+b2)/2 C=c2/2

D=a E=(b2-b1)/2

a1=A+CD2+DE b1=B-E a=D

a2=A-CD2-DE b2=B+E c2=2C



3. Quadratic-Linear

Estimation Equation

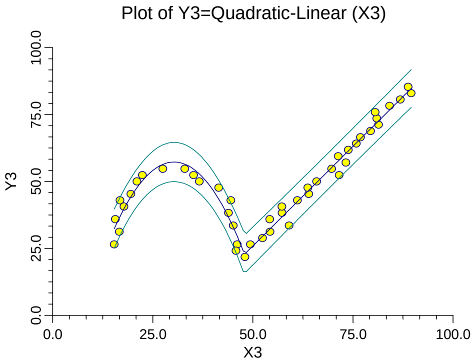
Y=A+BX+CX^2+(X-D)SIGN(X-D)[E-C(X+D)]

Common Equation

Y=a1+b1X+c1X^2,	X<=a
Y=a2+b2X,	X>a

Parameter Identities

A=(a1+a2)/2	B=(b1+b2)/2	C=c1/2
D=a	E=(b2-b1)/2	
a1=A-CD2+DE	b1=B-E	a=D
a2=A+CD2-DE	b2=B+E	c1=2C



4. Quadratic-Quadratic

Estimation Equation

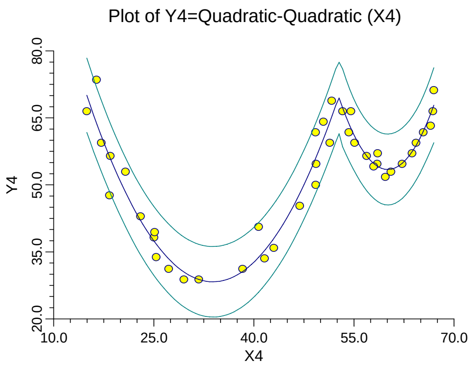
Y=A+BX+CX^2+(X-D)SIGN(X-D)[E(X+D)+F]

Common Equation

Y=a1+b1X+c1X^2,	X<=a
Y=a2+b2X+c2X^2,	X>a

Parameter Identities

A=(a1+a2)/2	B=(b1+b2)/2	C=(c1+c2)/2
D=a	E=(c2-c1)/2	F=(b2-b1)/2
a1=A-ED2+DF	b1=B-F	a=D
a2=A+eD2-DF	b2=B+F	
c1=C-E	c2=C+E	



5. Linear-Linear-Linear

Estimation Equation

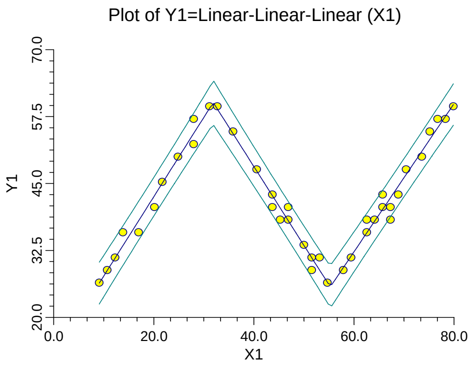
Y=A+BX+C(X-D)SIGN(X-D)+E(X-F)SIGN(X-F)

Common Equation

Y=a1+b1X	X<J1
Y=a2+b2X	a1<X<=J2
Y=a3+b3X	X>J2

Parameter Identities

A=(a1+a3)/2	B=(b1+b3)/2	C=(b2-b1)/2
D=J1	E=(b3-b2)/2	F=J2
a1=A+CD+EF	b1=B-C-E	J1=D
a2=A-CD-EF	b2=B+C-E	J2=F
a3=A-CD+EF	b3=B+C+E	



Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two variables: one dependent variable and one independent variable.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable from the current database.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y'=1/(Y*Y), ***Y'=1/Y***, ***Y'=1/SQRT(Y)***, ***Y'=LN(Y)***, ***Y'=SQRT(Y)***, ***Y'=Y (none)***, and ***Y'=Y*Y***

X (Independent) Variable

Specifies a single independent (*X*) variable from the current database.

X Transformation

Specifies a power transformation of the independent variable. Available transformations are

X'=1/(X*X), ***X'=1/X***, ***X'=1/SQRT(X)***, ***X'=LN(X)***, ***X'=SQRT(X)***, ***X'=X (none)***, and ***X'=X*X***

Model

Model Type

Specify the model that you want to fit. The available models are:

1 Linear-Linear
2 Linear-Quadratic
3 Quadratic-Linear
4 Quadratic-Quadratic
5 Linear-Linear-Linear

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias and the bias correction procedures used.

Model – Curve Transition

Transition Type

Specifies the smoothness of transition from one curve to the next.

Sharp means a corner is used to join the two curves. This is accomplished by using the SIGN(z) function. Mathematically, this function is defined as follows:

$$SIGN(X - D) = \begin{cases} -1, & X < D \\ 0, & X = D \\ 1, & X > D \end{cases}$$

Inside means a smooth transition is made inside the intersecting curves using the HYP(z) function. As z, the *transition value*, approaches 0, HYP(Z) approaches SIGN(z). This function is represented mathematically as follows:

$$HYP(z) = \frac{1}{(X - D)} \sqrt{(X - D)^2 + z}$$

Outside means a smooth transition is made outside the intersecting curves. Here TNH(z) stands for the hyperbolic tangent, tanh(z). As z, the *transition value*, approaches 0, TNH(Z) approaches SIGN(z). This function is represented mathematically as follows:

$$TNH(z) = \frac{e^{\frac{(X-D)}{z}} - e^{-\frac{(X-D)}{z}}}{e^{\frac{(X-D)}{z}} + e^{-\frac{(X-D)}{z}}}$$

Transition Value

A value defining the smoothness of the transition from one curve to the next. Enter a value close to zero for very sharp corner or a value close to one for a very smooth corner.

Note that this option is only used when the Transition Type is Inside or Outside.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Iteration Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Function Plot with Actual Y ... Probability Plot with Transformed Y

These options specify which plots are displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates and predicted values. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Function Plot and Residual Plot Tabs

This section controls the plot(s) showing the data with the fitted function line overlain on top and the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Line (Function Plot)

This option controls the attributes of the line representing the fitted function. Click this box to bring up the line specification dialog box. This window will let you set the line pattern, size, and color.

Display Prediction Limits (Function Plot)

This option controls whether the prediction limits (confidence limits on the predicted values) are displayed.

Number of Points (Function Plot)

This option specifies at how many points the estimated function is calculated to create the overlay function that is displayed on the Function Plots. A value between 50 and 150 is usually sufficient.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{M}* are replaced by the name of the variable and the model expression, respectively. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Piecewise Polynomial Model

This section presents an example of how to fit a piecewise polynomial model. In this example, we will fit the linear-linear-linear model (model 5) to the variables Y1 and X1 of the FNREG2 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Piecewise Polynomial Models window.

1 Open the FNREG2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG2.S0**.
- Click **Open**.

2 Open the Piecewise Polynomial Models window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **Piecewise Polynomial Models**. The Piecewise Polynomial Models procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Piecewise Polynomial Models window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Y1** from the list of variables and then click **Ok**.
- Double-click in the **X Variable** box. This will bring up the variable selection window.
- Select **X1** from the list of variables and then click **Ok**.
- Select **5 Linear-Linear-Linear** in the **Model Type** list box.

4 Specify the reports.

- On the Piecewise Polynomial Models window, select the **Reports tab**.
- Check the **Residual Report** box. Leave all other reports and plots checked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section

Itn   Error Sum
No.   Lambda         Lambda         A              B           C           D
0     91978.17       0.00004        13.73333       1.396706    0           32.7044
1     128.7423       0.000016       -22.38974      1.439365    -1.41563    32.7044
2     98.84669       0.0000064      -18.73867      1.417875    -1.451389   31.90939
3     97.88834       2.56E-06       -18.97554      1.433407    -1.477436   31.87549
4     97.88829       1.024E-06      -18.97609      1.43342     -1.477446   31.87606
Convergence criterion met.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm's progress.

Model Estimation Section

```
Model Estimation Section

Parameter               Parameter       Asymptotic       Lower           Upper
Name                    Estimate        Standard Error   95% C.L.        95% C.L.
A                       -18.97609       2.371069         -23.79468       -14.1575
B                       1.43342         4.792372E-02     1.336028        1.530813
C                       -1.477446       5.087956E-02     -1.580845       -1.374046
D                       31.87606        0.5111715        30.83723        32.91488
E                       1.428675        4.940346E-02     1.328275        1.529075
F                       55.26949        0.4470693        54.36093        56.17804

Dependent               Y1
Independent             X1
Model                   Y1=Linear-Linear-Linear (X1)
R-Squared               0.975292
Iterations              4
Estimated Model
(-18.97609)+(1.43342)*(X1)+(-1.477446)*((X1)-(31.87606))*SIGN((X1)-(31.87606))+(1.428675)*((X1)-(55.26949))
 *SIGN((X1)-(55.26949))

Common Model
Y1 = a1 + b1(X1)     if X1<=J1
Y1 = a2 + b2(X1)     if J1<X1<=J2
Y1 = a3 + b3(X1)     if X1>J2
where
a1 =12.89089         a2 =107.0812       a3 =-50.84307
b1 =1.482191         b2 =-1.4727        b3 =1.384649
J1 =31.87606         J2 =55.26949
```

This section reports the parameter estimates.

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Model

The model that was estimated. Use this to double check that the model estimated was what you wanted.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

$$\textit{R-Squared} = (\textit{ModelSS} - \textit{MeanSS})/(\textit{TotalSS}-\textit{MeanSS})$$

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

Estimated Model

The model that was estimated with the parameters replaced with their estimated values. This expression may be copied and pasted as a variable transformation in the spreadsheet. This will allow you to predict for additional values of X.

Common Model

This section shows the model in the usual format giving the individual linear equations and the change points. The range of X is shown for which each of the equations is valid. The change points are represented by J1 and J2.

Analysis of Variance Table

```
Analysis of Variance Table
                                          Sum of           Mean
Source               DF                   Squares          Square
Mean                 1                    70560            70560
Model                6                    74423.91         12403.99
Model (Adjusted)     5                    3863.911         772.7822
Error                34                   97.88829         2.879067
Total (Adjusted)     39                   3961.799
Total                40                   74521.8
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just “SSE.”

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters

      A           B           C           D           E           F
A     1.000000    -0.859286   0.199576    0.127212    -0.628010   -0.671159
B     -0.859286   1.000000    -0.501307   -0.411324   0.453763    0.463477
C     0.199576    -0.501307   1.000000    -0.110813   -0.543587   0.394458
D     0.127212    -0.411324   -0.110813   1.000000    0.513127    -0.257576
E     -0.628010   0.453763    -0.543587   0.513127    1.000000    0.043351
F     -0.671159   0.463477    0.394458    -0.257576   0.043351    1.000000
```

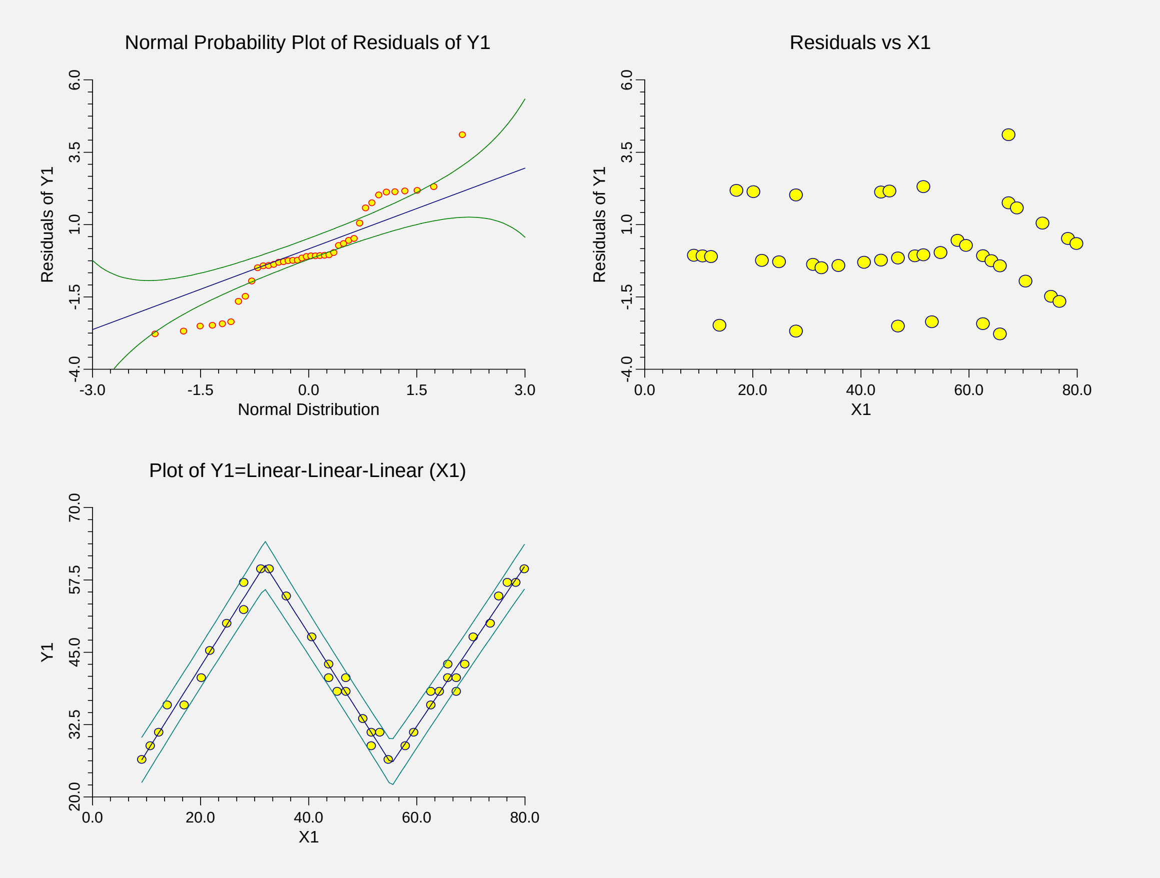
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section

Row                                     Predicted     Lower 95.0%   Upper 95.0%
No.   X1             Y1                 Value         Value         Value          Residual
1     9.119497       26.47059           26.40772      22.50505      30.3104        6.286404E-02
2     10.69182       28.82353           28.73821      24.91603      32.5604        8.531865E-02
3     12.26415       31.17647           31.0687       27.31541      34.82199       0.1077657
4     13.83648       35.88235           33.39919      29.70256      37.09583       2.483162
.     .              .                  .             .
.     .              .                  .             .
.     .              .                  .             .
```

The section shows the values of the residuals and predicted values. If you have observations in which the independent variable is given, but the dependent (Y) variable was left blank, a predicted value and prediction limits will be generated and displayed in this report.

Residual Plots



Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

In this example, we see a very striking nonrandom pattern. It appears that there might be another variable that causes a shift of about two units from one value to the next. This other variable has three values. If this were an actual study, we would now hunt for this third variable.

Function Plot

This plot displays the data along with the estimated function and prediction limits. It is useful in deciding if the fit is adequate and the prediction limits are appropriate.

In poorly fit models, we have found that it is often necessary to disable the prediction limits so that the data will show up. In these cases, the prediction limits may be so wide that the scale of the plot does not allow the data values to be separated.

Predicting for New Values

You can use your model to predict Y for new values of X. Here's how. Add new rows to the bottom of your database containing the values of the independent variable that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.

Chapter 370

Ratio of Polynomials Search – One Variable

Introduction

This procedure searches through hundreds of potential curves looking for the model that fits your data the best. The procedure is heuristic in nature, but seems to do well with the data we have tried.

A general class of models called the ratio of polynomials (see the previous chapter) provides a wide variety of curves to search from. Normally, fitting these models is a slow, iterative process. However, using a shortcut, an approximate solution may be found very quickly so that a large number of models may be searched in a short period of time. After the best fitting model is found, use the procedure discussed in the Ratio of Polynomials Fit chapter to provide a detailed analysis of it.

For each model, various transformations of X and Y can be tried. This expands the number of models that may be tried to several hundred.

The general ratio of polynomials model fit is

$$g(Y) = \frac{A0 + A1\,f(X) + A2\,f^{2}(X) + A3\,f^{3}(X) + A4\,f^{4}(X) + A5\,f^{5}(X)}{1 + B1\,f(X) + B2\,f^{2}(X) + B3\,f^{3}(X) + B4\,f^{4}(X) + B5\,f^{5}(X)} + e\,.$$

Here *g(Y)* and *f(X)* represent power transformations of Y and X such as LOG(X), SQRT(X), etc. The parameters *A0*, *A1*, *A2*, ..., *B5* are constants that are estimated from the data. The value *e* represents the error or residual of that observation. By setting some constants to zero, various simplified models are obtained. For example, if only *A0* and *A1* are nonzero, the familiar linear model, *Y=A0+A1X+e,* is obtained.

A Shortcut

Consider the simple model

$$Y = \frac{A0 + A1X}{1 + B1X} + e\,.$$

If you ignore *e* (set it to zero for a moment) and multiply both sides of this equation by *(1+B1X)* you will get

$$Y+B1XY=A0+A1X.$$

Now if you subtract *B1XY* from both sides you will get

$$Y=A0+A1X-B1XY.$$

Finally, if you relabel *XY* as *Z* you get

$$Y=A+BX+CZ.$$

Note that the variable *Z* is a direct transformation of *X* and *Y*. This last equation is in standard linear form. The parameters *A*, *B*, and *C* may be estimated using standard multiple regression! Note that the parameter *B1* in our original equation is equal to -*C* in the final equation.

One catch in using this procedure is that you have to assume the *e* to be zero. When the model fits well, the *e* will be near zero. When the model does not fit well, these *e* will be relatively large and our method breaks down. However, the large *e* will warn us that the model has not fit well.

Parsimony

One of the main principles in model building is that you should never use three parameters when two parameters will do. Hence, one of our tasks will be to find a model with the fewest number of parameters. A second principle in dealing with the ratio-of-polynomials model is that you should not fit a model with a numerator of higher polynomial order than that of the denominator. The models tried by this program follow these rules. A third rule is that all terms in a polynomial up to the desired order must be included. Hence, you would not use $Y=A+CX^2$. Instead you would fit $Y=A+BX+CX^2$.

The program tries the five models having a fifth-order polynomial in the denominator. The numerator polynomials are $A0+A1X$, $A0+A1X+A2X^2$, ..., $A0+A1X+A2X^2+A3X^3+A4X^4+A5X^5$. Next the four models having a fourth-order polynomial denominator are tried. This continues on down to the simple equation *Y=(A0+A1X)/(1+B1X)*. This process is repeated for each combination of transformations that are specified for *Y* and *X*.

Goodness-of-Fit

The final issue measuring of how well a given model fits the data so that the various models can be compared. This is tough since the goodness-of-fit statistics you are familiar with (like R^2) do not have the same meaning in this setting. However, because of the lack of other general, goodness-of-fit indices, we have chosen to base our selection on the value of R^2. We justify this because this procedure is only an intermediate step in the modeling process. You must take several steps before making your final model selection.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two variables: one dependent variable and one independent variable.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Y (Dependent) Variable

Variable

Specifies a single dependent (Y) variable from the current database. This is the variable being predicted.

Y (Dependent) Variable – Select Y Transformations

1/Y^2, 1/Y, 1/SQRT(Y), LN(Y), SQRT(Y), Y, Y^2

Specifies whether this transformation of Y should be searched.

X (Independent) Variable

Variable

Specifies a single independent (X) variable. This is the variable used to predict Y.

X (Independent) Variable – Select X Transformations

1/X^2, 1/X, 1/SQRT(X), LN(X), SQRT(X), X, X^2

Specifies whether this transformation of X should be searched.

Zero Cutoff

Zero

This is the value used as zero by the algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16.

Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control the reports and plots output.

Select Reports

Models Reported

This option limits the number of models that are reported on. For example, if you select 20 here, then the report shows the 20 best models.

Select Plots

Function Plot with Actual Y

Specifies whether the plot in the actual scale of Y and X should be displayed.

Function Plot with Transformed Y

Specifies whether the plot in the transformed scale of Y and X should be displayed.

Models Plotted

This option specifies how many of the best models are plotted.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Function Plot Tab

This section controls the plot(s) showing the data with the fitted function line on top.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Line

This option controls the attributes of the line representing the fitted function. Click this box to bring up the line specification dialog box. This window will let you set the line pattern, size, and color.

Number of Points

This option specifies at how many points the estimated function is calculated to create the overlay function that is displayed on the Function Plots. A value between 50 and 150 is usually sufficient.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Searching for the Best Ratio of Polynomials Model

This section presents an example of how to search for the best fitting ratio of polynomials model. In this example, we will search for the best fitting model using the variables Y and X of the FNREG3 database. We will also consider the log transformation of each variable in our search.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Ratio of Polynomials Search – One Variable window.

1 Open the FNREG3 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG3.S0**.
- Click **Open**.

2 Open the Ratio of Polynomials Search – One Variable window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **Ratio of Polynomials Search**. The Ratio of Polynomials Search – One Variable procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Ratio of Polynomials Search – One Variable window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Double-click in the **X Variable** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Search Summary Section

```
Search Summary Section

Model                                    Current      Best         Percent
No.    F(Y)     F(X)      Model          R-Squared    R-Squared    of Best
1      y        x         3 / 4          0.990115     0.990115     100.00
2      y        x         1 / 4          0.989091     0.990115     99.90
3      y        x         2 / 4          0.989078     0.990115     99.90
4      LN(y)    x         3 / 4          0.988342     0.990115     99.82
5      y        x         0 / 5          0.987972     0.990115     99.78
6      y        x         4 / 4          0.984444     0.990115     99.43
7      LN(y)    x         0 / 5          0.984241     0.990115     99.41
8      LN(y)    x         2 / 4          0.984015     0.990115     99.38
9      LN(y)    x         1 / 4          0.983513     0.990115     99.33
10     y        x         0 / 4          0.983109     0.990115     99.29
11     LN(y)    x         0 / 4          0.977900     0.990115     98.77
12     LN(y)    LN(x)     4 / 5          0.975901     0.990115     98.56
13     y        x         1 / 5          0.975564     0.990115     98.53
14     LN(y)    x         5 / 0          0.974421     0.990115     98.41
15     y        x         2 / 5          0.972910     0.990115     98.26
16     LN(y)    x         1 / 5          0.970396     0.990115     98.01
17     LN(y)    x         4 / 0          0.967638     0.990115     97.73
18     y        LN(x)     4 / 5          0.956059     0.990115     96.56
19     y        x         5 / 0          0.948378     0.990115     95.78
20     LN(y)    LN(x)     3 / 3          0.945165     0.990115     95.46
```

This report displays a separate line for each model tried. Note that the results have been sorted by R-Squared so that the best model is displayed at the top.

For this example, the best model is the ratio of a third order numerator polynomial and a fourth order denominator polynomial, with no transformations of Y or X needed. We would now fit this model using the Ratio of Polynomial Fit procedure.

Model No.

The ranking of the model displayed on this line.

F(Y)

The transformation (if any) applied to the Y (dependent) variable.

F(X)

The transformation (if any) applied to the X (independent) variable.

Model

The ratio of polynomial model whose results are displayed on this row. The syntax of the model statement is N/D where N represents the order of the numerator polynomial and D represents the order of the denominator polynomial. If N or D is set to zero, that polynomial is ignored.

For example, the model 1/2 means A0+A1X in the numerator and 1+B1X+B2X^2 in the denominator.

Current R-Squared

The value of pseudo R-Squared for this model and transformations.

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

$$\textit{\textbf{R-Squared = (ModelSS - MeanSS)/(TotalSS-MeanSS)}}$$

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

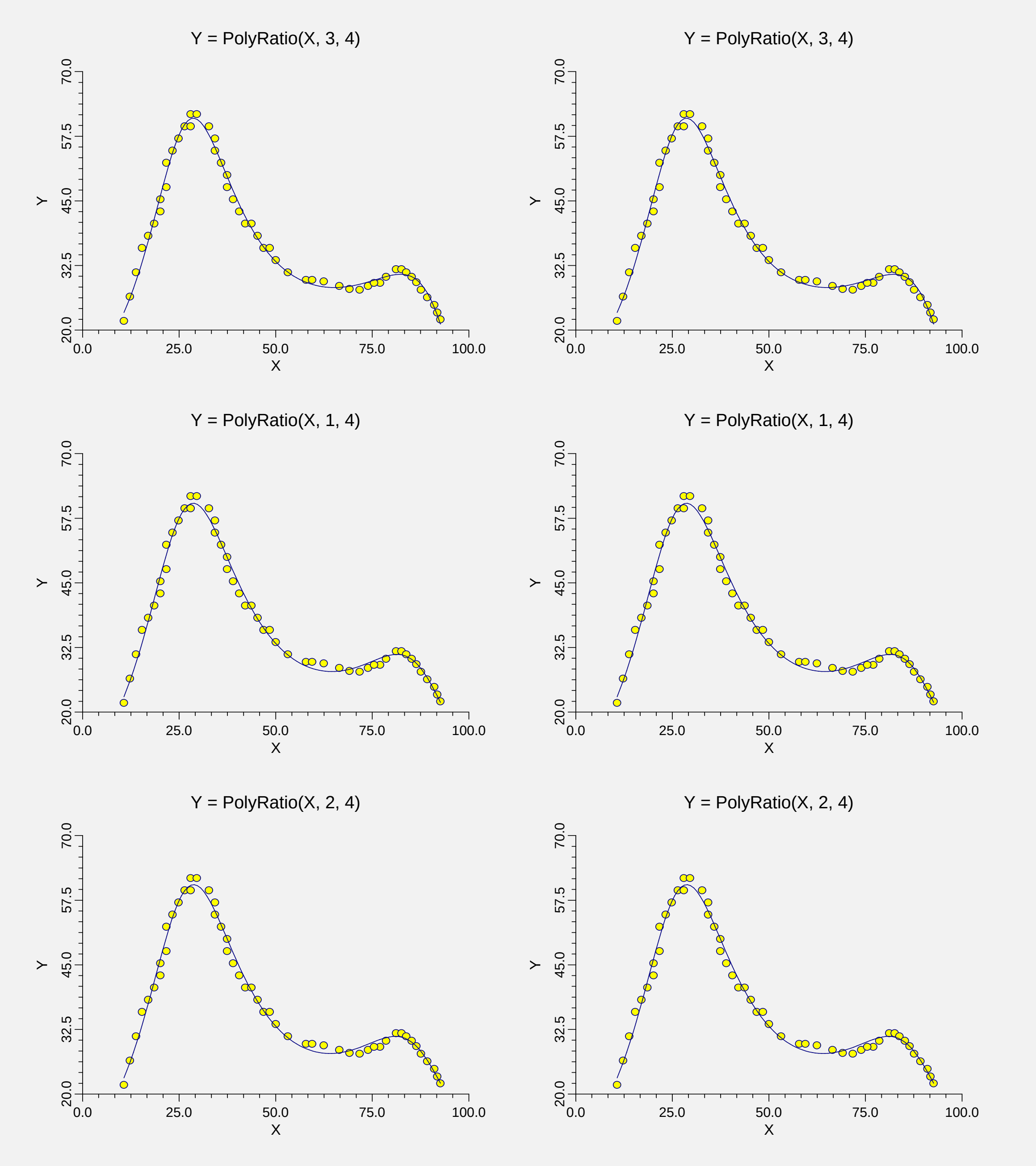
Best R-Squared

The pseudo R-Squared of the first (best) model.

Percent of Best

The percent that the pseudo R-Squared of this model is of the overall best model. Often you will be able to find models that are nearly as good as the best model, but have many fewer parameters.

Function Plots



These plots show the best few models plotted in the original (on the left) and transformed (on the right) scales. They will help you determine which model (or models) you want to evaluate further using the Ratio of Polynomial Fit procedure.

Chapter 371

Ratio of Polynomials Search – Many Variables

Introduction

This procedure searches through hundreds of potential curves looking for the model that fits your data the best. The procedure is heuristic in nature, but seems to do well with the data we have tried. This procedure is more general that the Ratio of Polynomials Search procedure because it models up to four independent variables.

A general class of models called the ratio of polynomials (see Multivariate Ratio of Polynomials Fit chapter) provides a wide variety of curves to search from. Normally, fitting these models is a slow, iterative process. However, using a shortcut, an approximate solution may be found very quickly so that a large number of models may be searched in a short period of time. After the best fitting model is found, use the procedure discussed in the Multivariate Ratio of Polynomials Fit chapter to obtain a detailed analysis.

Parsimony

One of the main principles in model building is that you never use three parameters when two parameters will do. Hence, one of our tasks will be to find a model with the fewest number of parameters. A second principle in dealing with the ratio-of-polynomials model is that you should not fit a model with a numerator of higher polynomial order than that of the denominator. The models tried by default by this program follow these rules. A third rule is that all terms in a polynomial up to the desired order must be included. Hence, you would not use $Y=A+CX^2$. Instead you would fit $Y=A+BX+CX^2$.

Goodness-of-Fit

Measuring how well a given model fits the data so that the various models can be compared is an important part of the search. This is tough since the goodness-of-fit statistics you are familiar with (like R-Squared) do not have the same meaning in this setting. However, because of the lack of other general, goodness-of-fit indices, we have chosen to base our selection on the value of R-Squared.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two or more variables: one dependent variable and up to four independent variables.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (Y) variable. This is the variable being predicted.

Variables – Transformations (Y)

1/Y^2, 1/Y, 1/SQRT(Y), LN(Y), SQRT(Y), Y, Y^2

Specifies whether this transformation of Y should be searched.

Variables – Independent Variables

U, V, W, X

Each option specifies a single independent variable. When a variable is selected, designate which of the transformations of that variable are searched.

Variables – Transformations (U,V,W,X)

1/U^2, 1/U, 1/SQRT(U), LN(U), SQRT(U), U, U^2

Specifies whether this transformation of U, V, W, and/or X should be searched.

Model Specification

These options specify which models are searched. Care must be taken when selecting models because it is very easy to overwhelm the algorithm by selecting too many candidate models.

For each model type, select one of the following options.

- **Omit**

  Do not add this type of model to the pool of models searched.

- **Numerator**

  Add models involving the numerator polynomial only to the pool of models that is searched.

- **Denominator**

  Add models involving the denominator polynomial only to the pool of models that is searched.

- **Numer. + Denom.**

  Add both numerator only and denominator only models to the pool of models that is searched.

- **Ratio**

  Add ratio of polynomial models to the pool of models that is searched.

- **Numer. + Ratio**

  Add numerator only and ratio models to the pool of models that is searched.

- **All Three**

  Add numerator only, denominator only, and ratio models to the pool of models that is searched.

(1-5) Single Variable

This option selects candidate models consisting of one variable only, up to the order specified to the left of the list box (1-5). For example, selecting Numerator in the second row (order 2) would include quadratic, single variable models such as

Y=B0+B1X+B2X^2, Y=B0+B1U+B2U^2, etc.

(1-5) Hierarchical

This option selects hierarchical-polynomial models (see Hi below) to the order specified by the number on the left (1 to 5).

1 Max Power

This option selects models using the “E”-notation (see below). These models would be of the form

Y = E1.

2 Max Power

This option selects models using the “E”-notation (see below). These models would be of the form

Y = E1,E2.

3 Max Power

This option selects models using the “E”-notation (see below). These models would be of the form

Y = E1,E2,E3.

4 Max Power

This option selects models using the “E”-notation (see below). These models would be of the form

Y = E1,E2,E3,E4.

5 Max Power

This option selects models using the “E”-notation (see Ei below). These models would be of the form

Y = E1,E2,E3,E4,E5.

Pairs - (1 and 2)

This option selects models consisting of two independent variables with no cross products, up to the order first (Pairs -1) or second (Pairs -2) order. For example, selecting Numerator in the Pairs-1 box would include models like

Y=B0+B1X+B2U.

Triplets - 1

This option selects models consisting of three independent variables with no cross products, up to a first order. For example, selecting Numerator in the Triplets -1 box would include models like

Y=B0+B1X+B2U+B3V.

Models Reported

This option limits the number of models that are reported on. For example, if you select 20 here, then the report shows the 20 best models.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Custom Models Tab

Up to ten custom models may be specified using the basic syntax for model input described below by specifying the numerator and denominator polynomials.

The syntax of these lists of terms follows these rules:

1. Individual terms may be listed as UiVj. If i or j is one, it may be omitted. For example, UV2X3 means (U)(V*V)(X*X*X) and U2 means U^2, which means U*U. A list of individual terms is formed by separating such terms with commas.
2. The **Oi** notation includes all terms of a particular order. The order is the sum of the exponents of the variables in a term. For example, the order of the term U2VW3 is six. If you had selected three variables and included “O2” in the list of terms, you would include the terms U3, V3, W3, U2V, U2W, V2W, UV2, VW2, and UVW in your model.
3. The **Si** notation includes all single variables to the power i. For example, if you had selected three variables and included “S2” in the list of terms, you would include the terms U, V, W, U2, V2, and W2 in your model.
4. The **Ei** notation includes all combinations of variables with at least one variable to the power i and none of the other variables to a power greater than i. For example, if you had selected three variables and included “E2” in the list of terms, you would include the terms U2, V2, W2, U2V, U2W, U2V2, U2W2, UV2, UW2, VW2, V2W, and V2W2 in your model.
5. The **Hi** notation includes all terms in the hierarchical model of order i. For example, if you had selected two variables and included “H2” in the list of terms, you would include the terms U, V, U2, V2, and UV in your model.
6. The **P** notation includes all simple paired terms. For example, if you had selected three variables and included “P” in the list of terms, you would include the terms UV, UW, and VW in your model.
7. The **T** notation includes all triplet terms. For example, if you had selected four variables and included “T” in the list of terms, you would include the terms UVW, UVX, UWX, and UWX in your model.

You can combine these notations however you like. If a term is specified twice, it will be included in the model only once. The order in which you specify terms is arbitrary. Examples are:

E2

U,V,E2,O1

O1,U2V2

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Multivariate Ratio of Polynomials Model

This section presents an example of how to fit a multivariate ratio of polynomials model. In this example, we will search for a model relating the dependent variable Y to the independent variables U and X of the FNREG4 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Ratio of Polynomials Search – Many Variables window.

1 Open the FNREG4 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG4.S0**.
- Click **Open**.

2 Open the Ratio of Polynomials Search – Many Variables window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **Many Independent Variables**, then **Ratio of Polynomials Search**. The Ratio of Polynomials Search – Many Variables procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Ratio of Polynomials Search – Many Variables window, select the **Variables tab**.
- Double-click in the **Y (Dependent)** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Check the box next to **Ln(Y)**.
- Double-click in the **U Variable** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**.
- Check the box next to **Ln(U)**.
- Double-click in the **V Variable** box. This will bring up the variable selection window.
- Select **U** from the list of variables and then click **Ok**.
- Check the box next to **Ln(V)**.
- Leave all other options at their default settings.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Multivariate Search Report

```
Results of Multivariate Search

Variables            Y=Y, U=X, V=U
Models Fit           192

              Percent    Transformations           No.     Model
R-Squared     of Best    Y    U    V    W     X    of X's  Numerator / Denominator
0.991356      100.00     N    L    L               10      (H2) / (H2)
0.991349      100.00     L    L    L               10      (H2) / (H2)
0.982629      99.12      L    N    L               10      (H2) / (H2)
0.982347      99.09      N    N    L               10      (H2) / (H2)
0.980970      98.95      N    N    L               8       (E1,E2) / 1
0.979765      98.83      N    L    L               8       (E1,E2) / 1
0.978064      98.66      L    N    L               8       (E1,E2) / 1
0.977188      98.57      N    L    L               5       (H2) / 1
0.976036      98.45      L    L    L               8       (E1,E2) / 1
0.975872      98.44      N    N    L               5       (H2) / 1
0.972253      98.07      L    N    L               5       (H2) / 1
0.971677      98.01      L    L    L               5       (H2) / 1
0.899874      90.77      L    N    N               8       (E1,E2) / 1
0.898429      90.63      N    L    N               10      (H2) / (H2)
0.897845      90.57      L    L    N               8       (E1,E2) / 1
0.893617      90.14      N    N    N               8       (E1,E2) / 1
0.892422      90.02      N    L    N               8       (E1,E2) / 1
0.887341      89.51      L    N    L               4       (U1,V1,U2,V2) / 1
0.884633      89.23      L    L    L               4       (U1,V1,U2,V2) / 1
0.884020      89.17      N    N    L               4       (U1,V1,U2,V2) / 1

Transformation key:
2=1/X^2, I=1/X, Q=1/SQRT(X), L=LN(X), R=SQRT(X), N=X, and S=X^2.
```

This report displays the best models (in terms of R-Squared) found. Each row describes the results for a single model.

Models Fit

This value is the total number of models that were evaluated.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

$$\textit{R-Squared} = (\textit{ModelSS} - \textit{MeanSS})/(\textit{TotalSS}-\textit{MeanSS})$$

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Percent of Best

This is the percent that the R-Squared value of this model is of the best (top) model.

Transformations Y U V W X

The letters correspond to the transformations that were used for each variable. The transformation key is listed at the bottom of the report. For example, the entry for the first row is N L L. This means that Y was regressed on ln(X) and ln(U).

No. of X’s

This is the number of parameters fit in this model. Of course, we want models with only a few parameters and a large value of R-Squared.

Model Numerator/Denominator

This gives the model using the shorthand notation described in Models - Custom section above. You can apply this shorthand notation directly in Multivariate Ratio of Polynomial Fit procedure to obtain detailed results for a particular model.

Note that the numeral one (1) is used when no polynomial is specified.

Chapter 375

Ratio of Polynomials Fit – One Variable

Introduction

This program fits a model that is the ratio of two polynomials of up to fifth order. Examples of this type of model are:

$$Y = \frac{A0 + A1X + A2X^2}{1 + B1X + B2X^2}$$

and

$$Y = \frac{A0 + A1X + A2X^2 + A3X^3 + A4X^4 + A5X^5}{1 + B1X + B2X^2 + B3X^3 + B4X^4 + B5X^5}$$

These models approximate many different curves. They offer a much wider variety of curves than the usual polynomial models. Since these are approximating curves and have no physical interpretation, care must be taken outside the range of the data. You must study the resulting model graphically to determine that the model behaves properly between data points.

Usually you would use the Ratio of Polynomials Search procedure first to find an appropriate model and then fit that model with this program.

Starting Values

Starting values are determined by the program. You do not have to supply starting values.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two variables: one dependent variable and one independent variable.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable from the current database.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y'=1/(Y*Y), ***Y'=1/Y***, ***Y'=1/SQRT(Y)***, ***Y'=LN(Y)***, ***Y'=SQRT(Y)***, ***Y'=Y (none)***, and ***Y'=Y*Y***

X (Independent) Variable

Specifies a single independent (*X*) variable from the current database.

X Transformation

Specifies a power transformation of the independent variable. Available transformations are

X'=1/(X*X), ***X'=1/X***, ***X'=1/SQRT(X)***, ***X'=LN(X)***, ***X'=SQRT(X)***, ***X'=X (none)***, and ***X'=X*X***

Model

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias and the bias correction procedures used.

These options specify the polynomials as well as certain values used to control the convergence of the nonlinear fitting algorithm.

Model – Numerator and Denominator Terms

Numerator and Denominator Terms (A1 X^1 ... B5 X^5)

These options specify which terms are in the model. The A-terms refer to the numerator and the B-terms refer to the denominator. Note that X^3 means X*X*X (X cubed).

Hence, checking A1 X^1, A2 X^2, B1 X^1, and B2 X^2 specifies the polynomial ratio model:

$$Y = \frac{A0 + A1X + A2X^2}{1 + B1X + B2X^2}$$

Note that the constant term, A0, is always included in the model.

We encourage you to follow a hierarchical approach to model building which is that you never fit a term without fitting all terms of lesser order. Hence, if you include A3 in your model, you also include A1 and A2. Also, never fit a higher order in the numerator than in the denominator. You can fit a higher order polynomial in the denominator than in the numerator.

If you follow these simple rules, you will usually be happy with the performance of your models.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Iteration Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Function Plot with Actual Y ... Probability Plot with Transformed Y

These options specify which plots are displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates and predicted values. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Function Plot and Residual Plot Tabs

This section controls the plot(s) showing the data with the fitted function line overlain on top and the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Line (Function Plot)

This option controls the attributes of the line representing the fitted function. Click this box to bring up the line specification dialog box. This window will let you set the line pattern, size, and color.

Display Prediction Limits (Function Plot)

This option controls whether the prediction limits (confidence limits on the predicted values) are displayed.

Number of Points (Function Plot)

This option specifies at how many points the estimated function is calculated to create the overlay function that is displayed on the Function Plots. A value between 50 and 150 is usually sufficient.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{M}* are replaced by the name of the variable and the model expression, respectively. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Ratio of Polynomials Model

This section presents an example of how to fit a ratio of polynomials model. In this example, we will fit a third order polynomial in the numerator and a fourth order polynomial in the denominator to the variables Y and X of the FNREG3 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Ratio of Polynomials Fit – One Variable window.

1 Open the FNREG3 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG3.S0**.
- Click **Open**.

2 Open the Ratio of Polynomials Fit – One Variable window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **Ratio of Polynomials Fit**. The Ratio of Polynomials Fit – One Variable procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Ratio of Polynomials Fit – One Variable window, select the **Variables tab**.
- Double-click in the **Y (Dependent) Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Double-click in the **X (Independent) Variable** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**.
- Under the **Numerator Terms** heading, check **A1 X^1**, **A2 X^2**, and **A3 X^3**.
- Under the **Denominator Terms** heading, check **B1 X^1**, **B2 X^2**, **B3 X^3**, and **B4 X^4**.

4 Specify the options.

- On the Ratio of Polynomials Fit – One Variable window, select the **Options tab**.
- Enter **10** in the **Max Iterations** box.

5 Specify the reports.

- On the Ratio of Polynomials Fit – One Variable window, select the **Reports tab**.
- Check the **Residual Report** box. Leave all other reports and plots checked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section

Itn   Error Sum
No.   Lambda     Lambda       A0         A1          A2            A3
0     70.81162   0.00004      11.78766   -0.2798519  5.809555E-03  -4.172031E-05
1     70.73225   0.16         11.78725   -0.2798166  5.810308E-03  -4.172031E-05
2     70.68134   0.064        11.78944   -0.2797599  5.810308E-03  -4.172031E-05
3     70.61962   0.256        11.78874   -0.2797441  5.810308E-03  -4.172031E-05
4     70.58868   0.1024       11.78977   -0.2797041  5.810308E-03  -4.172031E-05
5     70.58413   0.4096       11.78925   -0.2796919  5.810308E-03  -4.172031E-05
6     70.5548    0.16384      11.78946   -0.2796662  5.810308E-03  -4.172031E-05
7     70.54813   0.065536     11.79198   -0.2796135  5.810308E-03  -4.172031E-05
8     70.51613   0.262144     11.7916    -0.2795997  5.810308E-03  -4.172031E-05
9     70.50236   0.1048576    11.79268   -0.2795659  5.810308E-03  -4.172031E-05
10    70.49983   0.4194304    11.79228   -0.279557   5.810308E-03  -4.172031E-05
Maximum iterations before convergence.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm’s progress.

Model Estimation Section

```
Model Estimation Section

Parameter            Parameter       Asymptotic       Lower            Upper
Name                 Estimate        Standard Error   95% C.L.         95% C.L.
A0                   11.79254        1.012715         9.750206         13.83488
A1                   -0.2795364      9.166186E-02     -0.4643901       -9.468261E-02
A2                   5.810308E-03    3.516727E-03     -1.281847E-03    1.290246E-02
A3                   -4.172031E-05   2.863182E-05     -9.946187E-05    1.602125E-05
B1                   -0.0783304      2.020337E-03     -8.240479E-02    -0.074256
B2                   2.391857E-03    4.689093E-05     2.297292E-03     2.486421E-03
B3                   -2.86472E-05    4.332001E-06     -3.738351E-05    -1.991088E-05
B4                   1.171313E-07    1.025966E-06     -1.951927E-06    2.186189E-06

Dependent            Y
Independent          X
Model                Y=(A0+A1X^1+A2X^2+A3X^3) / (1+B1X^1+B2X^2+B3X^3+B4X^4)
R-Squared            0.990159
Iterations           10
Estimated Model
((11.79254-(.2795364)*(X)+(5.810308E-03)*(X)^2-(4.172031E-05)*(X)^3)/(1-(.0783304)*(X)+(2.391857E-03)*
(X)^2-(2.86472E-05)*(X)^3+(1.171313E-07)*(X)^4))
```

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Model

The model that was estimated. Use this to double check that the model estimated was what you wanted.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

$$R\text{-}Squared = (ModelSS - MeanSS)/(TotalSS\text{-}MeanSS)$$

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

Estimated Model

The model that was estimated with the parameters replaced with their estimated values. This expression may be copied and pasted as a variable transformation in the spreadsheet. This will allow you to predict for additional values of X.

Analysis of Variance Table

```
Analysis of Variance Table
                                        Sum of          Mean
Source                 DF               Squares         Square
Mean                   1                76141.53        76141.53
Model                  8                83234.82        10404.35
Model (Adjusted)       7                7093.291        1013.327
Error                  43               70.49983        1.639531
Total (Adjusted)       50               7163.791
Total                  51               83305.32
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just “SSE.”

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters
          A0          A1          A2          A3          B1          B2
A0        1.000000    -0.668342   0.251252    -0.126446   0.045937    0.033491
A1        -0.668342   1.000000    -0.826039   0.725516    0.586512    -0.664699
A2        0.251252    -0.826039   1.000000    -0.986996   -0.937223   0.967391
A3        -0.126446   0.725516    -0.986996   1.000000    0.978499    -0.991658
B1        0.045937    0.586512    -0.937223   0.978499    1.000000    -0.989097
B2        0.033491    -0.664699   0.967391    -0.991658   -0.989097   1.000000
B3        -0.444426   -0.127771   0.655601    -0.765633   -0.867018   0.801153
B4        0.769239    -0.364811   0.217490    -0.176658   -0.095811   0.123138
          B3          B4
A0        -0.444426   0.769239
A1        -0.127771   -0.364811
A2        0.655601    0.217490
A3        -0.765633   -0.176658
B1        -0.867018   -0.095811
B2        0.801153    0.123138
B3        1.000000    -0.139159
B4        -0.139159   1.000000
```

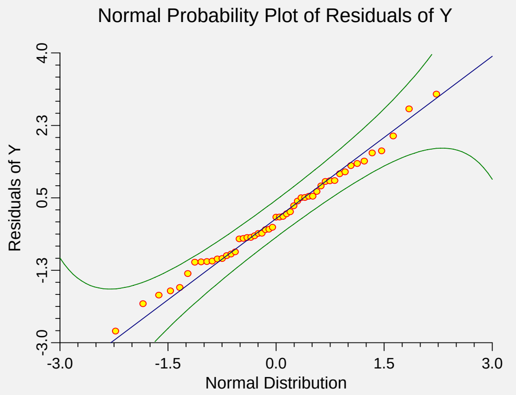
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

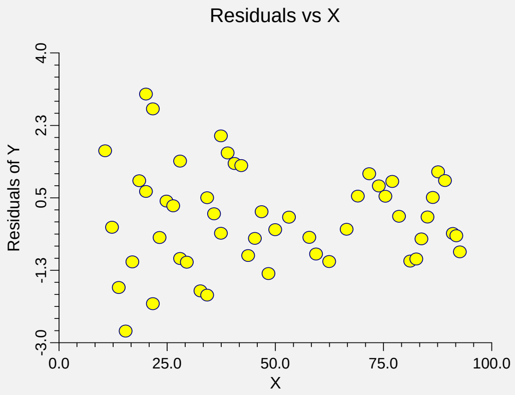
Predicted Values and Residuals Section

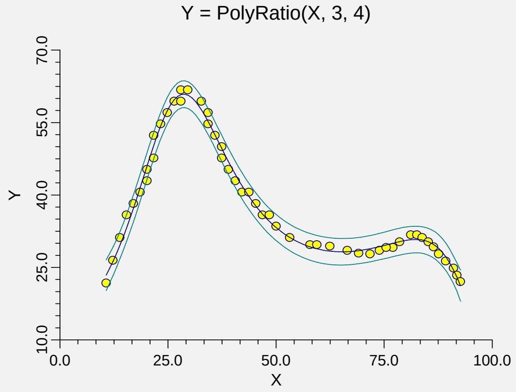
```
Predicted Values and Residuals Section

Row                                       Predicted       Lower 95.0%   Upper 95.0%
No.   X              Y                    Value           Value         Value        Residual
1     10.69182       21.76471             23.39933        20.2282       26.57045     -1.634617
2     12.26415       26.47059             26.25687        23.30506      29.20868     0.2137172
3     13.83648       31.17647             29.50871        26.70471      32.31271     1.667763
4     15.4088        35.88235             33.16126        30.38701      35.93551     2.721093
.     .              .                    .               .
.     .              .                    .               .
.     .              .                    .               .
```

The section shows the values of the residuals and predicted values. If you have observations in which the independent variable is given, but the dependent (Y) variable was left blank, a predicted value and prediction limits will be generated and displayed in this report.

Residual Plots







Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

In this example, it appears that the variance of the residuals decreases as X increases—suggesting that the equal variance assumption is violated. If this pattern was very drastic, we might want to try a variance stabilizing transformation of Y or using weighted nonlinear regression. However, the pattern does not appear severe in this case, so we probably would not take further action.

Function Plot

This plot displays the data along with the estimated function and prediction limits. It is useful in deciding if the fit is adequate and the prediction limits are appropriate.

In poorly fit models, we have found that it is often necessary to disable the prediction limits so that the data will show up. In these cases, the prediction limits may be so wide that the scale of the plot does not allow the data values to be separated.

Predicting for New Values

You can use your model to predict Y for new values of X. Here’s how. Add new rows to the bottom of your database containing the values of the independent variable that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.

Chapter 376

Ratio of Polynomials Fit – Many Variables

Introduction

This program fits a model that is the ratio of two polynomials of up to fifth order. Instead of a single independent variable, these polynomials may involve up to four independent variables (U, V, W, and X). An example of this type of model is:

$$Y = \frac{B0 + B1X + B2X^2 + B3U + B4U^2}{1 + B5X + B6X^2 + B7U + B8U^2}$$

These models approximate many different curves. They offer a much wider variety of curves than the usual polynomial models. Since these are approximating curves and have no physical interpretation, care must be taken outside the range of the data. You must study the resulting model graphically to determine that the model behaves properly between data points.

Usually you would use the Multivariate Ratio of Polynomials Search procedure first to find an appropriate model and then fit that model with this program.

Starting Values

Starting values are determined by the program. You do not have to supply starting values.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two or more variables: one dependent variable and up to four independent variables.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y'=1/(Y*Y), ***Y'=1/Y***, ***Y'=1/SQRT(Y)***, ***Y'=LN(Y)***, ***Y'=SQRT(Y)***, ***Y'=Y (none)***, and ***Y'=Y*Y***

U, V, W, X (Independent) Variable

Each option specifies an independent variable. At least one independent variable must be designated.

U, V, W, X Transformation

Specifies a power transformation of this independent variable. When the variable is referenced in the model, it refers to the transformed variable. Available transformations are

X'=1/(X*X), ***X'=1/X***, ***X'=1/SQRT(X)***, ***X'=LN(X)***, ***X'=SQRT(X)***, ***X'=X (none)***, and ***X'=X*X***

Model

These options specify the polynomials to be fit.

Numerator Terms

These options specify a list of terms that become the numerator polynomial of the model.

The syntax of these lists of terms follows these rules:

1. Individual terms may be listed as UiVj. If i or j is one, it may be omitted. For example, UV2X3 means (U)(V*V)(X*X*X) and U2 means U^2 which means U*U. A list of individual terms is formed by separating such terms with commas.
2. The **Oi** notation includes all terms of a particular order. The order is the sum of the exponents of the variables in a term. For example, the order of the term U2VW3 is six. If you had selected three variables and included "O2" in the list of terms, you would include the terms U3, V3, W3, U2V, U2W, V2W, UV2, VW2, and UVW in your model.

3. The **Si** notation includes all single variables to the power i. For example, if you had selected three variables and included "S2" in the list of terms, you would include the terms U, V, W, U2, V2, and W2 in your model.
4. The **Ei** notation includes all combinations of variables with at least one variable to the power i and none of the other variables to a power greater than i. For example, if you had selected three variables and included "E2" in the list of terms, you would include the terms U2, V2, W2, U2V, U2W, U2V2, U2W2, UV2, UW2, VW2, V2W, and V2W2 in your model.
5. The **Hi** notation includes all terms in the hierarchical model of order i. For example, if you had selected two variables and included "H2" in the list of terms, you would include the terms U, V, U2, V2, and UV in your model.
6. The **P** notation includes all simple paired terms. For example, if you had selected three variables and included "P" in the list of terms, you would include the terms UV, UW, and VW in your model.
7. The **T** notation includes all triplet terms. For example, if you had selected four variables and included "T" in the list of terms, you would include the terms UVW, UVX, UWX, and UWX in your model.

You can combine these notations however you like. If a term is specified twice, it will be included in the model only once. The order in which you specify terms is arbitrary. Examples are:

E2

U,V,E2,O1

O1,U2V2

Denominator Terms

These options specify a list of terms that become the denominator polynomial of the model. The syntax of these options follow the same rules as those given for Numerator Terms above.

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias and the bias correction procedures used.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Iteration Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Residual Plot with Actual Y ... Probability Plot with Transformed Y

These options specify which plots are displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates and predicted values. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Residual Plot Tab

This section controls the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{M}* are replaced by the name of the variable and the model expression, respectively. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Multivariate Ratio of Polynomials Model

This section presents an example of how to fit a multivariate ratio of polynomials model. In this example, we will fit a custom model to the variables Y, U, and X of the FNREG4 database. The numerator will include the terms SQRT(X), SQRT(U), and UX. The denominator will include the terms SQRT(UX), XU, and U.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Ratio of Polynomials Fit – Many Variables window.

1 Open the FNREG4 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG4.S0**.
- Click **Open**.

2 Open the Ratio of Polynomials Fit – Many Variables window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **Many Independent Variables**, then **Ratio of Polynomials Fit**. The Ratio of Polynomials Fit – Many Variables procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Ratio of Polynomials Fit – Many Variables window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Double-click in the **U Variable** box. This will bring up the variable selection window.
- Select **U** from the list of variables and then click **Ok**.
- Select **SQRT(z)** in the U Transformation box.
- Double-click in the **X Variable** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**.
- Select **SQRT(z)** in the X Transformation box.

4 Specify the reports.

- On the Ratio of Polynomials Fit – Many Variables window, select the **Reports tab**.
- Check the **Residual Report** box. Leave all other reports and plots checked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section

Itn   Error Sum
No.   Lambda          Lambda       B0          B1           B2           B3
0     0.0256219       0.00004      2.050828    0.7681349    -1.027562    1.102767
1     2.246299E-02    0.000016     2.002186    0.960939     -1.019122    1.408448
2     2.238311E-02    0.0000064    1.995856    0.9876834    -1.019637    1.488163
3     2.238232E-02    2.56E-06     1.99552     0.989055     -1.019808    1.495897
4     2.238232E-02    1.024E-06    1.99551     0.9890978    -1.019816    1.496198
Convergence criterion met.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm's progress.

Model Estimation Section

```
Model Estimation Section
Parameter                   Parameter    Asymptotic        Lower          Upper
Name         Term           Estimate     Standard Error    95% C.L.       95% C.L.
B0           Intercept      1.99551      0.0128966         1.970092       2.020928
B1           U              0.9890978    5.595055E-02      0.8788245      1.099371
B2           X              -1.019816    1.188279E-02      -1.043235      -0.9963957
B3           U2X2           1.496198     0.1342063         1.23169        1.760706
B4           u2             1.009521     2.531414E-02      0.9596294      1.059413
B5           ux             -1.081743    2.945009E-02      -1.139787      -1.0237
B6           u2x2           1.36935      0.1003615         1.171546       1.567153

R-Squared                0.993757
Iterations               4

Symbolic Model
Y = P1(U,X) / P2(U,X)
P1(U,X) = B0+B1*U+B2*X+B3*U2X2
P2(U,X) = 1+B4*U2+B5*UX+B6*U2X2
where
Y = Y
U = SQRT(U)
X = SQRT(X)

Estimated Model
((1.99551)+(.9890978)*(SQRT(U))-(1.019816)*(SQRT(X))+(1.496198)*(SQRT(U))^2
*(SQRT(X))^2) / (1+(1.009521)*(SQRT(U))^2-(1.081743)*(SQRT(U))*(SQRT(X))
 +(1.36935)*(SQRT(U))^2*(SQRT(X))^2)
```

This section reports the parameter estimates.

Parameter Name

The name of the parameter whose results are shown on this line.

Term

The name of the term in the model. Note that upper case letters are used for numerator terms and lower case letters are used for denominator terms.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

R-Squared = (ModelSS - MeanSS)/(TotalSS-MeanSS)

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

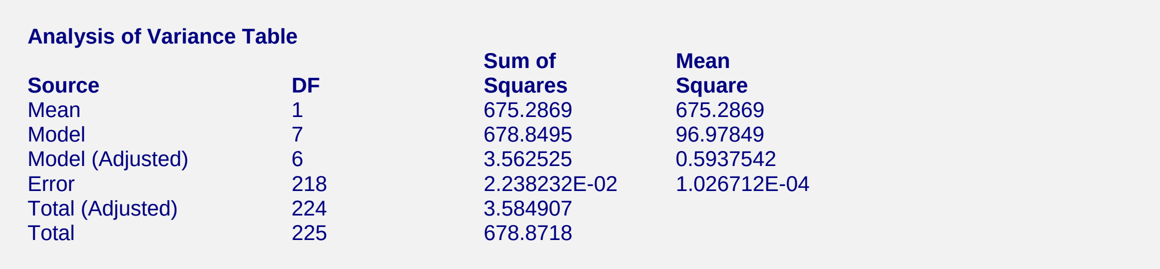
Symbolic Model

The expanded model that was fit. Any of the shortcut terms like O1 and E2 are replaced by the individual terms that they represent. Note that one list is presented for the numerator and one for the denominator. Any transformations that were applied are also listed.

Estimated Model

This is a copy of the symbolic model in which the parameter names have been replaced by their estimates. This expression may be used as a variable transformation by copying it and pasting it into the Variable Info section of the database.

Analysis of Variance Table



```
Analysis of Variance Table
                                          Sum of           Mean
Source                 DF                 Squares          Square
Mean                   1                  675.2869         675.2869
Model                  7                  678.8495         96.97849
Model (Adjusted)       6                  3.562525         0.5937542
Error                  218                2.238232E-02     1.026712E-04
Total (Adjusted)       224                3.584907
Total                  225                678.8718
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean	The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.
Model	The sum of squares associated with the model.
Model (Adjusted)	The model sum of squares minus the mean sum of squares.
Error	The sum of the squared residuals. This is often called the sum of squares error or just "SSE."
Total	The sum of the squared Y values.
Total (Adjusted)	The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters

      B0          B1          B2          B3          B4          B5
B0    1.000000    -0.915182   -0.635169   -0.014330   -0.850024   -0.465655
B1    -0.915182   1.000000    0.505466    -0.065987   0.969403    0.577334
B2    -0.635169   0.505466    1.000000    -0.520124   0.301986    0.802626
B3    -0.014330   -0.065987   -0.520124   1.000000    0.093673    -0.751426
B4    -0.850024   0.969403    0.301986    0.093673    1.000000    0.405675
B5    -0.465655   0.577334    0.802626    -0.751426   0.405675    1.000000
B6    0.041458    -0.150335   -0.553984   0.991271    0.008270    -0.819380
```

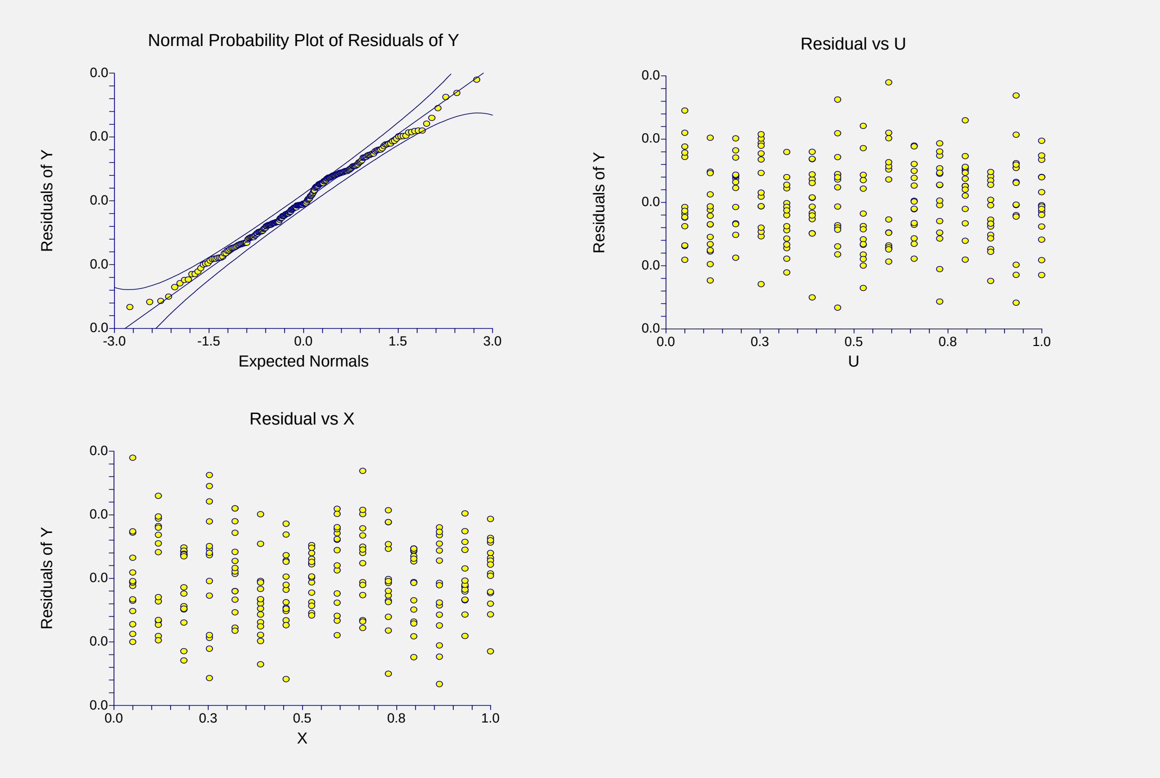
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section
Row                     Predicted      Lower 95.0%    Upper 95.0%
No.   Y                 Value          Value          Value          Residual
1     1.981996          1.992756       1.971384       2.014127       -1.075969E-02
2     2.028455          2.026659       2.006006       2.047311       1.796112E-03
3     2.027451          2.014322       1.993759       2.034884       1.312937E-02
.     .                 .              .              .
.     .                 .              .              .
.     .                 .              .              .
```

The section shows the values of the residuals and predicted values. If you have observations in which the independent variables are given, but the dependent (Y) variable was left blank, a predicted value and prediction limits will be generated and displayed in this report.

Residual Plots



Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus each of the independent variables. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

Predicting for New Values

You can use your model to predict Y for new values of the independent variables. Here’s how. Add new rows to the bottom of your database containing the values of the independent variables that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.

Chapter 380

Sum of Functions Models

Introduction

This program fits models that are the ratio of two linear expressions. The general form of a model is:

$$g(Y) = \frac{A0 + A1\,f_1(X) + A2\,f_2(X) + A3\,f_3(X) + A4\,f_4(X) + A5\,f_5(X)}{1 + B1h_1(X) + B2h_2(X) + B3h_3(X) + B4h_4(X) + B5h_5(X)} + e$$

where $f_i(X)$, $g(Y)$, and $h_i(X)$ are standard functions such as SIN(X), LN(X+1), SQRT(X/2), etc. The A0, A1, ..., B5 are constants (parameters) to be estimated from the data.

These models approximate many different curves. They offer a much wider variety of curves than the usual polynomial models.

Since these are approximating curves and have no physical interpretation, care must be taken outside the range of the data. You must study the resulting model graphically to determine that the model behaves properly between data points.

Starting Values

Starting values are determined by the program from the data. You do not have to supply starting values.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in two variables: one dependent variable and one independent variable.

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y'=1/(Y*Y), ***Y'=1/Y***, ***Y'=1/SQRT(Y)***, ***Y'=LN(Y)***, ***Y'=SQRT(Y)***, ***Y'=Y (none)***, and ***Y'=Y*Y***

X (Independent) Variable

Specifies a single independent (*X*) variable.

Model

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias and the bias correction procedures used.

Model – Numerator and Denominator Terms

These options specify up to five terms for use as the numerator and/or denominator of the model. You do not have to have a denominator.

Function

Select one of the eighteen possible transformations for this term.

f(z)=1/(z^2)	***f(z)=1/z***	***f(z)=1/SQRT(z)***
f(z)=LN(z)	***f(z)=SQRT(z)***	***f(z)=z (none)***
f(z)=z^2=z*z	***f(z)=z^3***	***f(z)=z^4***
f(z)=z^5	***f(z)=EXP(z)***	***f(z)=EXP(-z)***

f(z)=SIN(z)	***f(z)=COS(z)***	***f(z)=TAN(z)***
f(z)=SINH(z)	***f(z)=COSH(z)***	***f(z)=TANH(z)***

where

z = MX+A; M and A are constants that are supplied in the two options below.

Add (A)

X may be scaled using the equation z=MX+A. This option sets the value of A. If you want to ignore this option, set A=0.

Multiply (M)

X may be scaled using the equation z=MX+A. This option sets the value of M. If you want to ignore this option, set M=1.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt’s procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Iteration Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Function Plot with Actual Y ... Probability Plot with Transformed Y

These options specify which plots are displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates and predicted values. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Function Plot and Residual Plot Tabs

This section controls the plot(s) showing the data with the fitted function line overlain on top and the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Line (Function Plot)

This option controls the attributes of the line representing the fitted function. Click this box to bring up the line specification dialog box. This window will let you set the line pattern, size, and color.

Display Prediction Limits (Function Plot)

This option controls whether the prediction limits (confidence limits on the predicted values) are displayed.

Number of Points (Function Plot)

This option specifies at how many points the estimated function is calculated to create the overlay function that is displayed on the Function Plots. A value between 50 and 150 is usually sufficient.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{M}* are replaced by the name of the variable and the model expression, respectively. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a Sum of Functions Model

This section presents an example of how to fit a sum of functions model. In this example, we will fit the model

Y=A0+A1/(X+0.5)+SIN(X/2)+A3TANH(X)

to the variables Y and X of the FNREG1 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Sum of Functions Models window.

1 Open the FNREG1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **FNREG1.S0**.
- Click **Open**.

2 Open the Sum of Functions Models window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **Sum of Functions Models**. The Sum of Functions Models procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Sum of Functions Models window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Double-click in the **X Variable** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**.

- Under the Numerator Terms heading, select **1/z** in the **first Function** box.
- Under the Numerator Terms heading, enter **0.5** in the **first Add (A)** box.
- Under the Numerator Terms heading, select **SIN(z)** in the **second Function** box.
- Under the Numerator Terms heading, enter **0.5** in the **second Mult (M)** box.
- Under the Numerator Terms heading, select **Tanh(z)** in the **third Function** box.

4 Specify the reports.

- On the Sum of Functions Models window, select the **Reports tab**.
- Check the **Residual Report** box. Leave all other reports and plots checked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section

Itn   Error Sum
No.   Lambda      Lambda       A0          A1           A2             A3
0     5.915745    0.00004      2.070158    4.885924     1.059697       7.084624
Convergence criterion met.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm’s progress. Since no denominator terms were selected, the model was solved on the first iteration using standard multiple linear regression.

Model Estimation Section

```
Model Estimation Section

Parameter               Parameter       Asymptotic        Lower             Upper
Name                    Estimate        Standard Error    95% C.L.          95% C.L.
A0                      2.070158        1.075332          -6.435943E-02     4.204676
A1                      4.885924        0.5628729         3.768631          6.003218
A2                      1.059697        6.202012E-02      0.9365882         1.182806
A3                      7.084624        0.9798195         5.139698          9.029551

Dependent               Y
Independent             X
Model                   Y=(A0+A1*(1/(X+.5))+A2*(SIN((.5*X)))+A3*(TANH(X))) / (1)
R-Squared               0.956784
Iterations              0
Estimated Model
(2.070158+(4.885924)*1/(X+.5)+(1.059697)*SIN((.5*X))+(7.084624)*TANH(X))
```

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Model

The model that was estimated. Use this to double check that the model estimated was what you wanted. Note that the “/(1)” at the end emphasizes that there was no denominator specified.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

R-Squared = (ModelSS - MeanSS)/(TotalSS-MeanSS)

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

Estimated Model

The model that was estimated with the parameters replaced with their estimated values. This expression may be copied and pasted as a variable transformation in the spreadsheet. This will allow you to predict for additional values of X.

Analysis of Variance Table

```
Analysis of Variance Table
                                          Sum of          Mean
Source               DF                   Squares         Square
Mean                 1                    10559.48        10559.48
Model                4                    10690.45        2672.613
Model (Adjusted)     3                    130.9726        43.65752
Error                96                   5.915745        6.162234E-02
Total (Adjusted)     99                   136.8883
Total                100                  10696.37
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just “SSE.”

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters
      A0          A1          A2          A3
A0    1.000000    -0.972405   0.786167    -0.999209
A1    -0.972405   1.000000    -0.831952   0.964719
A2    0.786167    -0.831952   1.000000    -0.779440
A3    -0.999209   0.964719    -0.779440   1.000000
```

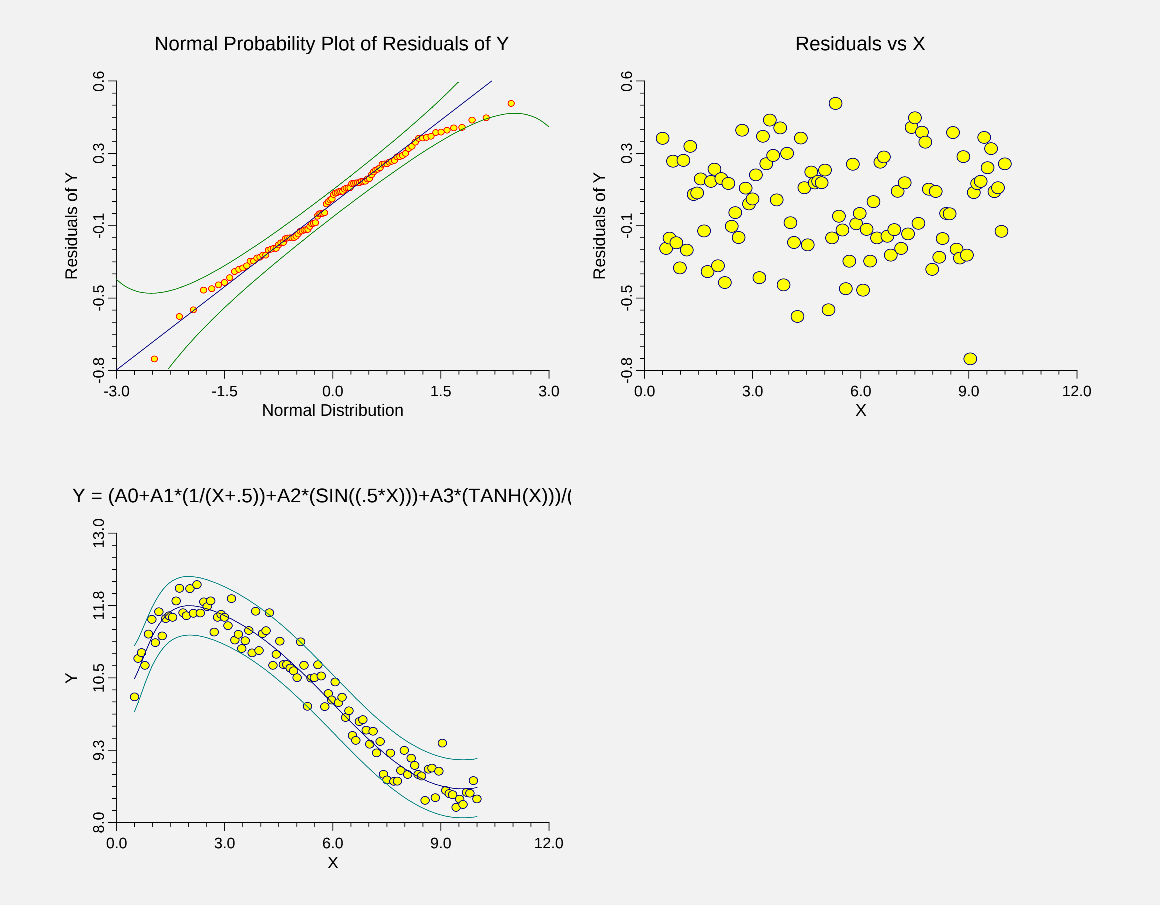
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section

Row                                Predicted     Lower 95.0%   Upper 95.0%
No.   X              Y             Value         Value         Value        Residual
1     0.5            10.16989      10.49218      9.92255       11.06181     -0.3222909
2     0.5959596      10.83415      10.62378      10.07584      11.17172     0.2103729
3     0.6919192      10.93412      10.77391      10.24289      11.30493     0.1602088
4     0.7878788      10.71519      10.92673      10.40745      11.44602     -0.2115456
.     .              .             .             .             .            .
.     .              .             .             .             .            .
.     .              .             .             .             .            .
```

The section shows the values of the residuals and predicted values. If you have observations in which the independent variable is given, but the dependent (Y) variable was left blank, a predicted value and prediction limits will be generated and displayed in this report.

Residual Plots



Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack

of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

Function Plot

This plot displays the data along with the estimated function and prediction limits. It is useful in deciding if the fit is adequate and the prediction limits are appropriate.

In poorly fit models, we have found that it is often necessary to disable the prediction limits so that the data will show up. In these cases, the prediction limits may be so wide that the scale of the plot does not allow the data values to be separated.

Predicting for New Values

You can use your model to predict Y for new values of X. Here’s how. Add new rows to the bottom of your database containing the values of the independent variable that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.

Chapter 385

User-Written Models

Introduction

This procedure is a special case of the Nonlinear Regression procedure in which there is only a single variable in the model. In this model, there are one or more parameters to be estimated from the data. An example of such a model is

$$Y = A + B\ EXP(-CX)$$

This program estimates the parameters in nonlinear models using the Levenberg-Marquardt nonlinear least-squares algorithm as presented in Nash (1987). We have implemented Nash's MRT algorithm with numerical derivatives. This has been a popular algorithm for solving nonlinear least squares problems, since the use of numerical derivatives means you do not have to supply program code for the derivatives.

Starting Values

Starting values must be provided. Instructions for calculating reasonable starting values are given in the chapter on Nonlinear Regression.

Assumptions and Limitations

Usually, nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. In this case, the usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model.

Data Structure

The data are entered in one dependent variable and one independent variable. An example of data appropriate for this procedure, taken from page 476 of Draper and Smith (1981), is shown below.

In this example, the dependent variable (Y) is the proportion of available chlorine in a certain quantity of chlorine solution and the independent variable (X) is the length of time in weeks since the product was produced. When the product is produced, the proportion of chlorine is 0.50.

During the 8 weeks that it takes to reach the consumer, the proportion declines to 0.49. The hypothesized model for predicting Y from X is

$$Y = A + (0.49 - A) EXP(- B(X-8)) + e.$$

Here, A and B are the parameters and e is the error or residual. Note that only 8 of the 44 observations contained in the DS476 database are displayed here.

DS476 dataset (subset)

X	Y
8	0.49
8	0.49
10	0.48
10	0.47
10	0.48
10	0.47
12	0.46
12	0.46

Missing Values

Rows with missing values in the variables being analyzed are ignored in the calculations. When only the value of the dependent variable is missing, predicted values are generated.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These panels specify the variables and model used in the analysis.

Variables

Y (Dependent) Variable

Specifies a single dependent (*Y*) variable from the current database.

Y Transformation

Specifies a power transformation of the dependent variable. Available transformations are

Y'=1/(Y*Y), ***Y'=1/Y***, ***Y'=1/SQRT(Y)***, ***Y'=LN(Y)***, ***Y'=SQRT(Y)***, ***Y'=Y (none)***, and ***Y'=Y*Y***

Model

Bias Correction

This option controls whether a bias-correction factor is applied when the dependent variable has been transformed. Check it to correct the predicted values for the transformation bias. Uncheck it to leave the predicted values unchanged. See the Introduction to Curve Fitting chapter for a discussion of the amount of bias and the bias correction procedures used.

Model

This box contains the nonlinear equation. Note that you do not include the "Y=" portion of the expression--it is assumed.

This expression is made up of

1. Symbols: +, -, *, /, ^, <, >, =, (, and).
2. Functions:

ABS(X)	Absolute value of X	ASN(X)	Arc sine of X
ATN (X)	Arc tangent of X	COS(X)	Cosine of X
EXP(X)	Exponential of X	INT(X)	Integer part of X
LN(X)	Log base e of X	LOG(X)	Log base 10 of X
SGN(X)	Signature of X	SIN(X)	Sine of X
SQR(X)	Square root of X	TAN(X)	Tangent of X
TNH(X)	Hyperbolic tangent of X		

3. One variable referenced by name. For example, if your independent variable is called DOSE, you would use the word DOSE in the equation.
4. Parameters which are defined below.
5. Constants.

The syntax of the model expression follows that of the variable transformations, so we will not go into syntax here, but refer you to the Variable Transformations chapter. Note that only a subset of the functions available as transformations are also available here.

Examples of valid models are

A+B*X^C

A+B*EXP(-C*X)

(A0 + A1*X + A2*X) / (1 + B1*X + B2*X)

Parameter

This is the name of a parameter to be estimated as it appears in the model statement above. The parameter name is any combination of letters and numbers, except that the name must begin with a letter. You should not use symbols in the parameter name. All letters are converted to upper case internally, so it does not matter whether you use upper or lower case. The name cannot be one of the internal mathematical functions like SIN or TAN, as this will confuse the function parser.

The name may be as long as you want, but, for readability, you should keep it short.

Minimum, Starting Value, Maximum

This box contains the minimum value, staring value, and maximum value for the parameter. The three numbers are separated by blanks or commas.

- **Minimum Value**

  This is the smallest value that the parameter can take on. The algorithm searches for a value between this and the Maximum Value. If you want to search in an unlimited range, enter a large negative number such as **-1E9**, which is -1000000000.

  Since this is a search algorithm, the narrower the range that you search in, the quicker the process will converge.

  Care should be taken to specify minima and maxima that keep calculations in check. Suppose, for example, that your equation includes the expression LOG(B*X) and that values of X are positive. Since you cannot take the logarithm of zero or a negative number, you should set the minimum of B as a positive number. This will insure that the estimation procedure will not fail because of impossible calculations.

- **Starting Value**

  This is the beginning value of the parameter. The algorithm searches for a value between the Minimum Value and the Maximum Value, beginning with this number. The closer this value is to the final value, the quicker the algorithm will converge.

  Although specific instructions for finding starting values are given at the first of this chapter, we would like to make the following suggestions here.

  1. Make sure that the starting values you supply are legitimate. For example, if the model you were estimating included the phrase 1/B, you would not want to start with B=0.

  2. Before you go to a lot of effort, make a few trial runs using starting values of 0.0, 0.5, and 1.0. Often, one of these values will converge.

  3. If you have a large number of observations, take a small sample of observations from your original database and work with this subset database. When you find a set of starting values that converges on this subset database, use the resulting parameter estimates as starting values with the complete database. Since nonlinear regression is iterative and each iteration must pass through the complete database, this can save a considerable amount of time while you are searching for starting values.

- **Maximum Value**

  This is the largest value that the parameter can take on. The algorithm searches for a value between the Minimum Value and this value, beginning at the Starting Value. If you want to search in an unlimited range, enter a large positive number such as **1E9**, which is 1000000000.

  Since this is a search algorithm, the narrower the range that you search in, the quicker the process will converge.

  Care should be taken to specify minima and maxima that keep calculations in check. Suppose, for example, that your equation includes the expression LOG(B*X) and that values of X are negative. Since you cannot take the logarithm of zero or a negative number, you should set the maximum of B as a negative number near zero. This will insure that the estimation procedure will not fail because of impossible calculations.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 50) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Iteration Report ... Residual Report

These options specify which reports are displayed.

Select Plots

Function Plot with Actual Y ... Probability Plot with Transformed Y

These options specify which plots are displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic confidence limits of the parameter estimates and predicted values. Usually, this number will range from 0.1 to 0.001. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Function Plot and Residual Plot Tabs

This section controls the plot(s) showing the data with the fitted function line overlain on top and the residual plots.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Line (Function Plot)

This option controls the attributes of the line representing the fitted function. Click this box to bring up the line specification dialog box. This window will let you set the line pattern, size, and color.

Display Prediction Limits (Function Plot)

This option controls whether the prediction limits (confidence limits on the predicted values) are displayed.

Number of Points (Function Plot)

This option specifies at how many points the estimated function is calculated to create the overlay function that is displayed on the Function Plots. A value between 50 and 150 is usually sufficient.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{X}* are replaced by appropriate names. *{M}* is replaced by the model expression. Press the button on the right of the field to specify the font of the text.

Probability Plot Tab

The options on this panel control the appearance of the probability plot of the residuals.

Vertical and Horizontal Axis

Label

This is the text of the axis labels. The characters *{Y}* and *{X}* are replaced by appropriate names. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a probability plot style file. This file sets all probability plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Probability Plot procedure.

Symbol

Click this box to bring up the symbol specification dialog box. This window will let you set the symbol type, size, and color.

Titles

Plot Title

This is the text of the title. The characters *{Y}* and *{M}* are replaced by the name of the variable and the model expression, respectively. Press the button on the right of the field to specify the font of the text.

Storage Tab

The predicted values and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the variables specified here.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id's

A list of the Template Id's of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Fitting a User-Written Model

This section presents an example of how to run a nonlinear regression analysis of the data that was presented above in the Data Structure section. In this example, we will fit the model

$$Y = A + (0.49 - A)\ EXP(- B(X-8))$$

to the data contained in the variables Y and X on the database DS476.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the User-Written Models window.

1 Open the DS476 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **DS476.S0**.
- Click **Open**.

2 Open the User-Written Models window.

- On the menus, select **Analysis**, then **Curve Fitting**, then **One Independent Variable**, then **User-Written Models**. The User-Written Models procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the User-Written Models window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Y** from the list of variables and then click **Ok**.
- Enter **A+(0.49-A)*EXP(-B*(X-8))** in the **Model** box.
- Enter **A** in the **first Parameter** box.
- Enter **0 0.1 1** in the **first Minimum, Starting Value, Maximum** box.
- Enter **B** in the **second Parameter** box.
- Enter **0 0.013 1** in the **second Minimum, Starting Value, Maximum** box.

4 Specify the reports.

- On the User-Written Models window, select the **Reports tab**.
- Check the **Residual Report** box. Leave all other reports and plots checked.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Minimization Phase Section

```
Minimization Phase Section

Itn      Error Sum
No.      of Squares          Lambda         A              B
0        1.643321E-02        0.00004        0.1            0.013
1        0.0147339           0.016          0.1464944      1.375224E-02
.        .                   .              .              .
.        .                   .              .              .
13       5.00168E-03         0.2684354      0.3901402      0.101633
Convergence criterion met.
```

This report displays the error (residual) sum of squares, lambda, and parameter estimates for each iteration. It allows you to observe the algorithm's progress toward the solution.

Model Estimation Section

```
Model Estimation Section

Parameter            Parameter      Asymptotic       Lower          Upper
Name                 Estimate       Standard Error   95% C.L.       95% C.L.
A                    0.3901402      5.033759E-03     0.3799816      0.4002987
B                    0.101633       1.336168E-02     7.466801E-02   0.1285979

Dependent            Y
Independent          X
Model                Y = A+(0.49-A)*EXP(-B*(X-8))
R-Squared            0.873375
Iterations           13
Estimated Model
(.3901402)+(0.49-(.3901402))*EXP(-(.101633)*((X)-8))
```

Parameter Name

The name of the parameter whose results are shown on this line.

Parameter Estimate

The estimated value of this parameter.

Asymptotic Standard Error

An estimate of the standard error of the parameter based on asymptotic (large sample) results.

Lower 95% C.L.

The lower value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Upper 95% C.L.

The upper value of a 95% confidence limit for this parameter. This is a large sample (at least 25 observations for each parameter) confidence limit.

Model

The model that was estimated. Use this to double check that the model estimated was what you wanted.

R-Squared

There is no direct R-Squared defined for nonlinear regression. This is a pseudo R-Squared constructed to approximate the usual R-Squared value used in multiple regression. We use the following generalization of the usual R-Squared formula:

R-Squared = (ModelSS - MeanSS)/(TotalSS-MeanSS)

where *MeanSS* is the sum of squares due to the mean, *ModelSS* is the sum of squares due to the model, and *TotalSS* is the total (uncorrected) sum of squares of Y (the dependent variable).

This version of R-Squared tells you how well the model performs after removing the influence of the mean of Y. Since many nonlinear models do not explicitly include a parameter for the mean of Y, this R-Squared may be negative (in which case we set it to zero) or difficult to interpret. However, if you think of it as a direct extension of the R-Squared that you use in multiple regression, it will serve well for comparative purposes.

Iterations

The number of iterations that were completed before the nonlinear algorithm terminated. If the number of iterations is equal to the Maximum Iterations that you set, the algorithm did not converge, but was aborted.

Estimated Model

The model that was estimated with the parameters replaced with their estimated values. This expression may be copied and pasted as a variable transformation in the spreadsheet. This will allow you to predict for additional values of X.

Analysis of Variance Table

```
Analysis of Variance Table
                                            Sum of          Mean
Source                   DF                 Squares         Square
Mean                     1                  7.9475          7.9475
Model                    2                  7.981998        3.990999
Model (Adjusted)         1                  3.449832E-02    3.449832E-02
Error                    42                 5.00168E-03     1.190876E-04
Total (Adjusted)         43                 0.0395
Total                    44                 7.987
```

Source

The labels of the various sources of variation.

DF

The degrees of freedom.

Sum of Squares

The sum of squares associated with this term. Note that these sums of squares are based on Y, the dependent variable. Individual terms are defined as follows:

Mean The sum of squares associated with the mean of Y. This may or may not be a part of the model. It is presented since it is the amount used to adjust the other sums of squares.

Model The sum of squares associated with the model.

Model (Adjusted) The model sum of squares minus the mean sum of squares.

Error The sum of the squared residuals. This is often called the sum of squares error or just “SSE.”

Total The sum of the squared Y values.

Total (Adjusted) The sum of the squared Y values minus the mean sum of squares.

Mean Square

The sum of squares divided by the degrees of freedom. The Mean Square for Error is an estimate of the underlying variation in the data.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters
                        A               B
A                       1.000000        0.887330
B                       0.887330        1.000000
```

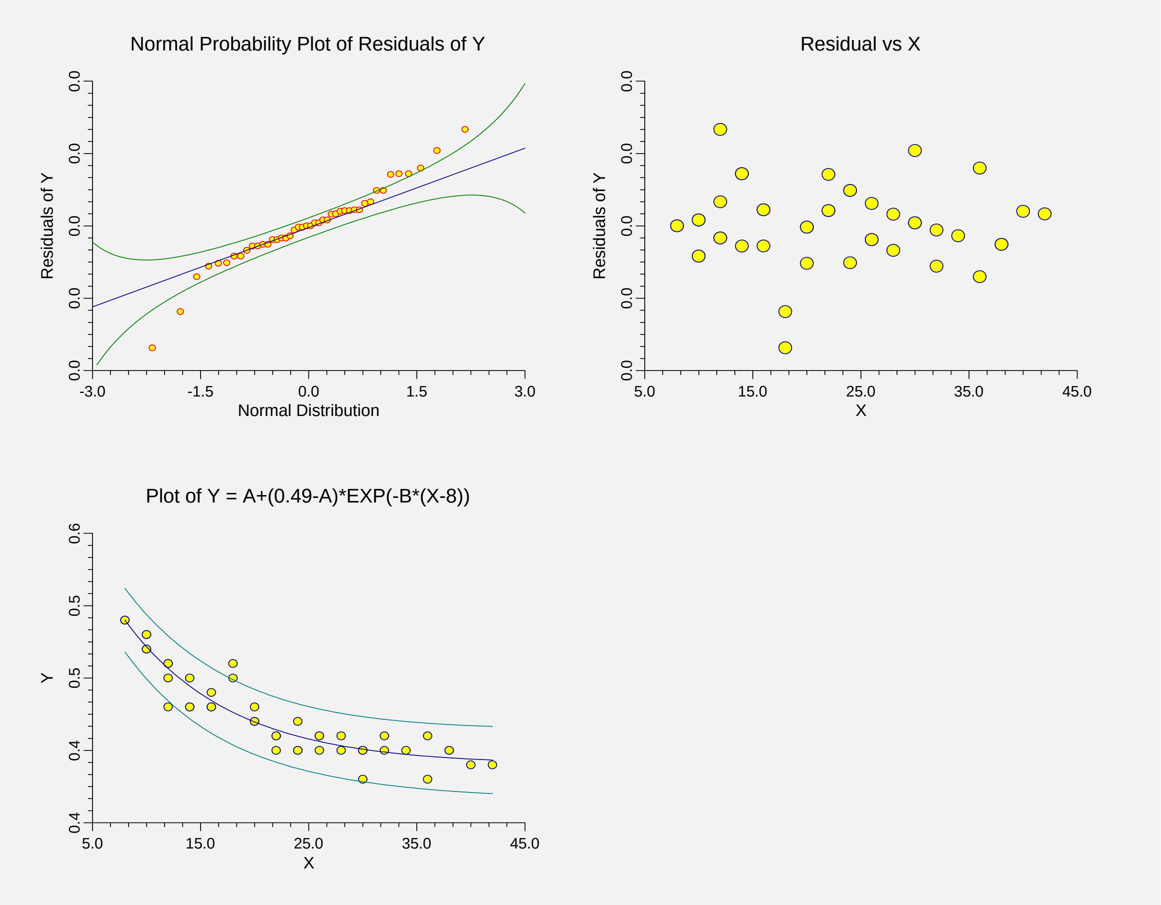
This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

Predicted Values and Residuals Section

```
Predicted Values and Residuals Section

Row                     Predicted    Lower 95.0%     Upper 95.0%
No.    X       Y        Value        Value           Value           Residual
1      8       0.49     0.49         0.4679772       0.5120228       0
2      8       0.49     0.49         0.4679772       0.5120228       0
3      10      0.48     0.4716319    0.4494232       0.4938406       0.0083681
4      10      0.47     0.4716319    0.4494232       0.4938406       -0.0016319
.      .       .        .            .
.      .       .        .            .
.      .       .        .            .
```

The section shows the values of the residuals and predicted values. If you have observations in which the independent variable is available, but the dependent (Y) variable was missing, a predicted value and prediction limits will be generated and displayed in this report.

Residual Plots



Normal Probability Plot

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. We do not recommend that you use this diagnostic with small sample sizes.

Residual versus X Plot

This is a scatter plot of the residuals versus the independent variable, X. The preferred pattern is a rectangular shape or point cloud. Any nonrandom pattern may require a redefining of the model.

Function Plot

This plot displays the data along with the estimated function and prediction limits. It is useful in deciding if the fit is adequate and the prediction limits are appropriate.

In poorly fit models, we have found that it is often necessary to disable the prediction limits so that the data will show up. In these cases, the prediction limits may be so wide that the scale of the plot does not allow the data values to be separated.

Predicting for New Values

You can use your model to predict Y for new values of X. Here’s how. Add new rows to the bottom of your database containing the values of the independent variable that you want to create predictions for. Leave the dependent variable blank. When the program analyzes your data, it will skip these rows during the estimation phase, but it will generate predicted values for all rows, regardless of whether the Y variable is missing or not.

Chapter 390

Area Under Curve

Introduction

Suppose you are studying a drug that influences plasma concentration. A popular method of looking at the absorption and elimination properties of the drug is to follow the plasma concentration across time. When the concentration values are plotted on the vertical axis, the time values are plotted on the horizontal axis, and the points are joined with a line, a curve results. One method of making comparisons among different types of drugs and different doses of the same drug is to compute the area under the curve (AUC).

AUC is computed by the trapezoidal rule as follows:

$$AUC = \frac{1}{2}\sum_{i=1}^{n-1}\left(T_{i+1} - T_i\right)\left(C_{i+1} + C_i - 2B\right)$$

where T_i is the ith time value, C_i is the ith concentration value, n is the number of time values, and B is the baseline value. The area between the baseline and the curve is computed by this formula.

The AUC should be calculated from zero to a time at which the concentration has returned to its regular levels. Also, when making comparisons, you should insure that all AUC's are computed for the same time intervals.

Data Structure

Two different data structures may be used to store data for analysis by this procedure. In the first format, the X (time) values are stored in one variable and the Y (concentration) values are stored in other variables, one variable per group. The AUC dataset is in this format.

In the other format, the X values are stored in one variable, group values are stored in a second variable, and the measurements are stored in a third variable. The data in the AUC dataset was reorganized to be in this format in the AUC1 dataset.

Format Type 1

AUC dataset

Time	P1	P2	P3
0	5	4	6
1	15	14	17
2	20	16	22
3	21	18	23
4	21	17	25
5	19	15	22
10	15	12	18
50	6	3	7

Format Type 2

AUC1 dataset

Time	Subject	Concentration
0	P1	5
1	P1	15
2	P1	20
3	P1	21
4	P1	21
5	P1	19
10	P1	15
50	P1	6
0	P2	4
1	P2	14
2	P2	16
3	P2	18
4	P2	17
5	P2	15
10	P2	12
50	P2	3
0	P3	6
1	P3	17
2	P3	22
3	P3	23
4	P3	25
5	P3	22
10	P3	18
50	P3	7

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Y Variables

Specify one or more vertical (Y) variables. These variables contain the measurements.

When multiple Y variables are specified, a separate AUC calculation is made for each variable. When only one Y variable is given, you must also specify a Break Variable. You cannot specify both a break variable and multiple Y variables.

X Variable

This option specifies the horizontal (X) variable. Usually, X represents the time at which the measurement was made.

Break Variable

Specify an optional break variable used to separate the X and Y variables into groups. A separate AUC calculation is made for each unique value of this variable. When this variable is specified, only one Y variable may be used.

Model

Baseline

This option specifies a baseline value for the area under the curve. Usually, this value is zero.

The area between a horizontal line at this value and the data points is calculated. Points below this amount count as negative area (they are subtracted). Points above this value are added. The trapezoidal formula is used to calculate the area.

Reports Tab

Select Reports

AUC Report and Data Report

These options specify whether the corresponding reports are displayed.

Select Plots

Data Plot

This option specifies whether to display the data plot.

Report Options

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, whereas the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

Value Labels

This option lets you select whether to display only values, value labels, or both. Use this option if you want the table to automatically attach labels to the values (like 1=Yes, 2=No, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Data Plot Tab

These options specify the data plot.

Vertical and Horizontal Axis

Label

This is the text of the label. The characters *{Y}* are replaced by the name of the Y variable and the characters *{X}* are replaced by the name of the X variable. Press the button on the right of the field to specify the font of the text.

Minimum and Maximum

These options specify the minimum and maximum values to be displayed on the vertical (Y) and horizontal (X) axis. If left blank, these values are calculated from the data.

Tick Label Settings...

Pressing these buttons brings up a window that sets the font, rotation, and number of decimal places displayed in the reference numbers along the vertical and horizontal axes.

Ticks: Major and Minor

These options set the number of major and minor tickmarks displayed on each axis.

Show Grid Lines

These check boxes indicate whether the grid lines should be displayed.

Plot Settings

Plot Style File

Designate a scatter plot style file. This file sets all scatter plot options that are not set directly on this panel. Unless you choose otherwise, the default style file (Default) is used. These files are created in the Scatter Plot procedure.

Baseline

Check this box to cause a horizontal line representing the baseline to be displayed. The Baseline value is specified on the Variables tab. The color and thickness of the line can be changed using the button just below this check box.

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies legend label. The characters {G} are replaced by the name of the break variable.

Titles

Plot Title

This is the text of the title. Press the button on the right of the field to specify the font of the text.

Symbols

These options specify the attributes of the symbols used for each appraiser in the plots.

Symbol 1 - 15

These options specify the symbols used in the plot of each appraiser. The first symbol is used by the first appraiser, the second symbol by the second appraiser, and so on. These symbols are provided to allow the various appraisers to be easily identified, even on black and white printers.

Clicking on a symbol box (or the small button to the right of the symbol box) will bring up a window that allows the color, width, and pattern of the line to be changed.

Template Tab

The options on this panel allow various sets of options to be loaded (File menu: Load Template) or stored (File menu: Save Template). A template file contains all the settings for this procedure.

Specify the Template File Name

File Name

Designate the name of the template file either to be loaded or stored.

Select a Template to Load or Save

Template Files

A list of previously stored template files for this procedure.

Template Id’s

A list of the Template Id’s of the corresponding files. This id value is loaded in the box at the bottom of the panel.

Example 1 – Calculating Area Under the Curve

This section presents a tutorial of calculating the AUC for three individuals labeled P1, P2, and P3. The data are contained on the AUC database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Area Under Curve window. The corresponding setup for the AUC1 dataset is given as the **Example2** template.

1 Open the AUC dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **AUC.S0**.
- Click **Open**.

2 Open the Area Under Curve window.

- On the menus, select **Analysis**, then **Other**, then **Area Under Curve**. The Area Under Curve procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Area Under Curve window, select the **Variables tab**.
- Double-click in the **Y Variables** text box. This will bring up the variable selection window.
- Select variables **P1, P2, P3** from the list of variables and then click **Ok**. “P1-P3” will appear in this box.
- Double-click in the **X Variable** text box. This will bring up the variable selection window.
- Select **Time** and then click **Ok**. “Time” will appear in this box.

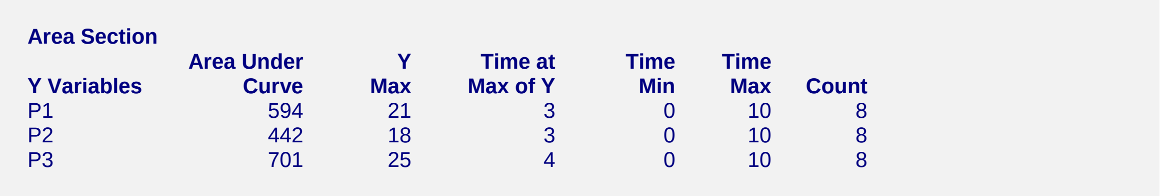
4 Specify the reports.

- On the Area Under Curve window, select the **Reports tab**.
- Check the **Data Report**. The other reports should be checked already.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Area Section



```
Area Section
                   Area Under       Y      Time at     Time     Time
Y Variables             Curve     Max     Max of Y      Min      Max    Count
P1                        594      21            3        0       10        8
P2                        442      18            3        0       10        8
P3                        701      25            4        0       10        8
```

This report shows the area under curve as well as supporting information.

Y Variables

This gives the value of the Y variable (if multiple Y variables were specified) or the value of the break variable (if a break variable was specified).

Area Under Curve

The is the area under the curve calculated in the interval between Time Min and Time Max. It is calculated using the trapezoidal formula presented earlier.

Y Max

This is the maximum value of the vertical (concentration) variable. This is an alternative measure of absorption that is sometimes used.

Time at Max of Y

This is the value of the time (X) variable at which the maximum Y value (Y Max) was found.

Time Min and Max

These give the horizontal range over which the area was calculated. These values are given so that you can compare them. Under normal circumstances, these values should be equal across groups.

Count

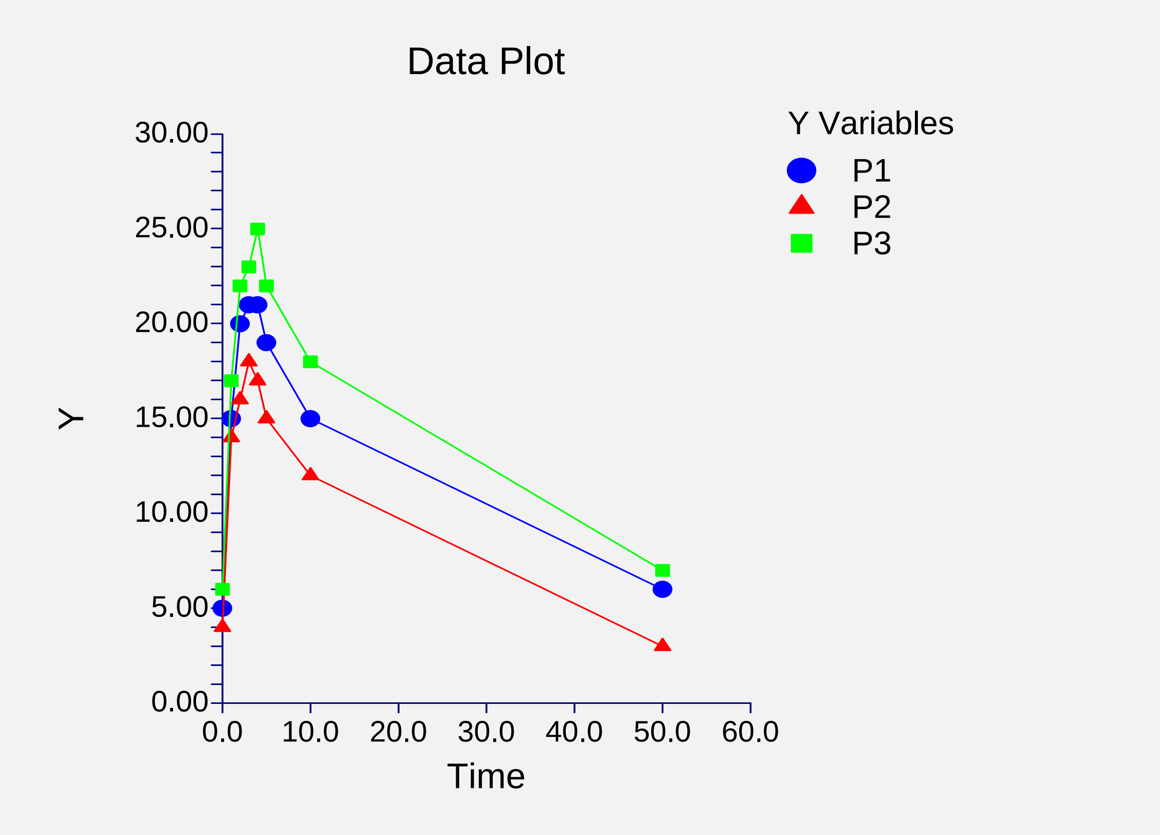
This is the number of X values that were found on the database for this group. This report helps you find missing values.

Data Section

```
Data Section

Y Variables     Time       Y
P1                 0       5
P1                 1      15
P1                 2      20
P1                 3      21
P1                 4      21
P1                 5      19
P1                10      15
P1                50       6
P2                 0       4
P2                 1      14
P2                 2      16
P2                 3      18
P2                 4      17
P2                 5      15
P2                10      12
P2                50       3
P3                 0       6
P3                 1      17
P3                 2      22
P3                 3      23
P3                 4      25
P3                 5      22
P3                10      18
P3                50       7
```

This report displays the data that are plotted. Note that if you have several measurements at the same time value, only their average is plotted.

Plots Section



This is a plot of the three curves under which the area was calculated. These plots let you spot any data inadequacies.

References

A

Agresti, A. and Coull, B. 1998. “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” *American Statistician*, Volume 52 Number 2, pages 119-126.

A’Hern, R. P. A. 2001. "Sample size tables for exact single-stage phase II designs." *Statistics in Medicine*, Volume 20, pages 859-866.

AIAG (Automotive Industry Action Group). 1995. *Measurement Systems Analysis*. This booklet was developed by Chrysler/Ford/GM Supplier Quality Requirements Task Force. It gives a detailed discussion of how to design and analyze an R&R study. The book may be obtained from ASQC or directly from AIAG by calling 801-358-3570.

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle," In B. N. Petrov & F. Csaki (Eds.), *The second international symposium on information theory*. Budapest, Hungary: Akademiai Kiado.

Akaike, H. 1974. "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19, (6): pages 716-723.

Albert, A. and Harris, E. 1987. *Multivariate Interpretation of Clinical Laboratory Data*. Marcel Dekker, New York, New York. This book is devoted to a discussion of how to apply multinomial logistic regression to medical diagnosis. It contains the algorithm that is the basis of our multinomial logistic regression routine.

Allen, D. and Cady, F.. 1982. *Analyzing Experimental Data by Regression*. Wadsworth. Belmont, Calif. This book works completely through several examples. It is very useful to those who want to see complete analyses of complex data.

Al-Sunduqchi, Mahdi S. 1990. *Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics*. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.

Altman, Douglas. 1991. *Practical Statistics for Medical Research*. Chapman & Hall. New York, NY. This book provides an introductory discussion of many statistical techniques that are used in medical research. It is the only book we found that discussed ROC curves.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. 1997. *Statistical Models Based on Counting Processess*. Springer-Verlag, New York. This is an advanced book giving many of the theoretically developments of survival analysis.

Anderson, R.L. and Hauck, W.W. 1983. "A new Procedure for testing equivalence in comparative bioavailability and other clinical trials." *Commun. Stat. Theory Methods.*, Volume 12, pages 2663-2692.

Anderson, T.W. and Darling, D.A. 1954. "A test of goodness-of-fit." *J. Amer. Statist. Assoc*, Volume 49, pages 765-769.

Andrews, D.F., and Herzberg, A.M. 1985. *Data*. Springer-Verlag, New York. This book is a collection of many different data sets. It gives a complete description of each.

Armitage. 1955. “Tests for linear trends in proportions and frequencies.” *Biometrics*, Volume 11, pages 375-386.

Armitage, P., and Colton, T. 1998. *Encyclopedia of Biostatistics*. John Wiley, New York.

Armitage,P., McPherson, C.K., and Rowe, B.C. 1969. “Repeated significance tests on accumulating data.” *Journal of the Royal Statistical Society, Series A*, 132, pages 235-244.

Atkinson, A.C. 1985. *Plots, Transformations, and Regression*. Oxford University Press, Oxford (also in New York). This book goes into the details of regression diagnostics and plotting. It puts together much of the recent work in this area.

Atkinson, A.C., and Donev, A.N. 1992. *Optimum Experimental Designs*. Oxford University Press, Oxford. This book discusses D-Optimal designs.

Austin, P.C., Grootendorst, P., and Anderson, G.M. 2007. “A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study,” *Statistics in Medicine*, Volume 26, pages 734-753.

B

Bain, L.J. and Engelhardt, M. 1991. *Statistical Analysis of Reliability and Life-Testing Models*. Marcel Dekker. New York. This book contains details for testing data that follow the exponential and Weibull distributions.

Baker, Frank. 1992. *Item Response Theory*. Marcel Dekker. New York. This book contains a current overview of IRT. It goes through the details, providing both formulas and computer code. It is not light reading, but it will provide you with much of what you need if you are attempting to use this technique.

Barnard, G.A. 1947. "Significance tests for 2 x 2 tables." *Biometrika* 34:123-138.

Barrentine, Larry B. 1991. *Concepts for R&R Studies*. ASQC Press. Milwaukee, Wisconsin. This is a very good applied work book on the subject of repeatability and reproducibility studies. The ISBN is 0-87389-108-2. ASQC Press may be contacted at 800-248-1946.

Bartholomew, D.J. 1963. "The Sampling Distribution of an Estimate Arising in Life Testing." *Technometrics*, Volume 5 No. 3, 361-374.

Bartlett, M.S. 1950. "Tests of significance in factor analysis." *British Journal of Psychology (Statistical Section)*, 3, 77-85.

Bates, D. M. and Watts, D. G. 1981. "A relative offset orthogonality convergence criterion for nonlinear least squares," *Technometrics*, Volume 23, 179-183.

Beal, S. L. 1987. "Asymptotic Confidence Intervals for the Difference between Two Binomial Parameters for Use with Small Samples." *Biometrics*, Volume 43, Issue 4, 941-950.

Belsley, Kuh, and Welsch. 1980. *Regression Diagnostics*. John Wiley & Sons. New York. This is the book that brought regression diagnostics into the main-stream of statistics. It is a graduate level treatise on the subject.

Benjamini, Y. and Hochberg, Y. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57, No. 1, 289-300.

Bertsekas, D.P. 1991. *Linear Network Optimization: Algorithms and Codes*. MIT Press. Cambridge, MA.

Blackwelder, W.C. 1993. "Sample size and power in prospective analysis of relative risk." *Statistics in Medicine*, Volume 12, 691-698.

Blackwelder, W.C. 1998. "Equivalence Trials." In *Encyclopedia of Biostatistics*, John Wiley and Sons. New York. Volume 2, 1367-1372.

Bloomfield, P. 1976. *Fourier Analysis of Time Series*. John Wiley and Sons. New York. This provides a technical introduction to fourier analysis techniques.

Bock, R.D., Aiken, M. 1981. "Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, 46, 443-459.

Bolstad, B.M., et al. 2003. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19, 185-193.

Bonett, Douglas. 2002. “Sample Size Requirements for Testing and Estimating Coefficient Alpha.” *Journal of Educational and Behavioral Statistics*, Vol. 27, pages 335-340.

Box, G.E.P. and Jenkins, G.M. 1976. *Time Series Analysis - Forecasting and Control*. Holden-Day.: San Francisco, California. This is the landmark book on ARIMA time series analysis. Most of the material in chapters 6 - 9 of this manual comes from this work.

Box, G.E.P. 1949. "A general distribution theory for a class of likelihood criteria." *Biometrika,* 1949, **36**, 317-346.

Box, G.E.P. 1954a. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: I." *Annals of Mathematical Statistics*, **25**, 290-302.

Box, G.E.P. 1954b. "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variable Problems: II." *Annals of Mathematical Statistics*, **25**, 484-498.

Box, G.E.P., Hunter, S. and Hunter. 1978. *Statistics for Experimenters*. John Wiley & Sons, New York. This is probably the leading book in the area experimental design in industrial experiments. You definitely should acquire and study this book if you plan anything but a casual acquaintance with experimental design. The book is loaded with examples and explanations.

Breslow, N. E. and **Day, N. E.** 1980. *Statistical Methods in Cancer Research: Volume 1. The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Brown, H., and Prescott, R. 2006. *Applied Mixed Models in Medicine*. 2nd ed. John Wiley & Sons Ltd. Chichester, West Sussex, England.

Brush, Gary G. 1988. *Volume 12: How to Choose the Proper Sample Size*, American Society for Quality Control, 310 West Wisconsin Ave, Milwaukee, Wisconsin, 53203. This is a small workbook for quality control workers.

Burdick, R.K. and Larsen, G.A. 1997. “Confidence Intervals on Measures of Variability in R&R Studies.” *Journal of Quality Technology, Vol. 29, No. 3, Pages 261-273*. This article presents the formulas used to construct confidence intervals in an R&R study.

Bury, Karl. 1999. *Statistical Distributions in Engineering.*. Cambridge University Press. New York, NY. (www.cup.org).

C

Cameron, A.C. and Trivedi, P.K. 1998. *Regression Analysis of Count Data*. Cambridge University Press. New York, NY. (www.cup.org).

Carmines, E.G. and Zeller, R.A. 1990. *Reliability and Validity Assessment*. Sage University Paper. 07-017. Newbury Park, CA.

Casagrande, J. T., Pike, M.C., and Smith, P. G. 1978. “The Power Function of the “Exact” Test for Comparing Two Binomial Distributions,” *Applied Statistics*, Volume 27, No. 2, pages 176-180. This article presents the algorithm upon which our Fisher’s exact test is based.

Cattell, R.B. 1966. "The scree test for the number of factors." *Mult. Behav. Res.* 1, 245-276.

Cattell, R.B. and Jaspers, J. 1967. "A general plasmode (No. 30-10-5-2) for factor analytic exercises and research." *Mult. Behav. Res. Monographs*. 67-3, 1-212.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. 1983. *Graphicals Methods for Data Analysis.* Duxbury Press, Boston, Mass. This wonderful little book is full of examples of ways to analyze data graphically. It gives complete (and readable) coverage to such topics as scatter plots, probability plots, and box plots. It is strongly recommended.

Chatfield, C. 1984. *The Analysis of Time Series*. Chapman and Hall. New York. This book gives a very readable account of both ARMA modeling and spectral analysis. We recommend it to those who wish to get to the bottom of these methods.

Chatterjee and Price. 1979. *Regression Analysis by Example*. John Wiley & Sons. New York. A great hands-on book for those who learn best from examples. A newer edition is now available.

Chen, K.W.; Chow, S.C.; and Li, G. 1997. "A Note on Sample Size Determination for Bioequivalence Studies with Higher-Order Crossover Designs" *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 25, No. 6, pages 753-765.

Chen, T. T. 1997. "Optimal Three-Stage Designs for Phase II Cancer Clinical Trials." *Statistics in Medicine*, Volume 16, pages 2701-2711.

Chen, Xun. 2002. "A quasi-exact method for the confidence intervals of the difference of two independent binomial proportions in small sample cases." *Statistics in Medicine*, Volume 21, pages 943-956.

Chow, S.C. and Liu, J.P. 1999. *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker. New York.

Chow, S.C.; Shao, J.; Wang, H. 2003. *Sample Size Calculations in Clinical Research*. Marcel Dekker. New York.

Chow, S.-C.; Shao, J.; Wang, H. 2008. *Sample Size Calculations in Clinical Research, Second Edition*. Chapman & Hall/CRC. Boca Raton, Florida.

Cochran and Cox. 1992. *Experimental Designs. Second Edition*. John Wiley & Sons. New York. This is one of the classic books on experimental design, first published in 1957.

Cochran, W.G. and Rubin, D.B. 1973. "Controlling bias in observational studies," *Sankhya, Ser. A*, Volume 35, Pages 417-446.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey. This is a very nice, clearly written book. There are MANY examples. It is the largest of the sample size books. It does not deal with clinical trials.

Cohen, Jacob. 1990. "Things I Have Learned So Far." *American Psychologist*, December, 1990, pages 1304-1312. This is must reading for anyone still skeptical about the need for power analysis.

Collett, D. 1991. *Modelling Binary Data.* Chapman & Hall, New York, New York. This book covers such topics as logistic regression, tests of proportions, matched case-control studies, and so on.

Collett, D. 1994. *Modelling Survival Data in Medical Research.* Chapman & Hall, New York, New York. This book covers such survival analysis topics as Cox regression and log rank tests.

Conlon, M. and Thomas, R. 1993. "The Power Function for Fisher's Exact Test." *Applied Statistics*, Volume 42, No. 1, pages 258-260. This article was used to validate the power calculations of Fisher's Exact Test in PASS. Unfortunately, we could not use the algorithm to improve the speed because the algorithm requires equal sample sizes.

Conover, W.J. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons, Inc. New York.

Conover, W.J., Johnson, M.E., and **Johnson, M.M.** 1981. *Technometrics***, 23,** 351-361**.**

Cook, D. and Weisberg, S. 1982. *Residuals and Influence in Regression*. Chapman and Hall. New York. This is an advanced text in the subject of regression diagnostics.

Cooley, W.W. and Lohnes, P.R. 1985. *Multivariate Data Analysis*. Robert F. Krieger Publishing Co. Malabar, Florida.

Cox, D. R. 1972. "Regression Models and life tables." *Journal of the Royal Statistical Society, Series B*, Volume 34, Pages 187-220. This article presents the proportional hazards regression model.

Cox, D. R. 1975. “Contribution to discussion of Mardia (1975a).” *Journal of the Royal Statistical Society, Series B*, Volume 37, Pages 380-381.

Cox, D.R. and Snell, E.J. 1981. *Applied Statistics: Principles and Examples*. Chapman & Hall. London, England.

Cureton, E.E. and D'Agostino, R.B. 1983. *Factor Analysis - An Applied Approach*. Lawrence Erlbaum Associates. Hillsdale, New Jersey. (This is a wonderful book for those who want to learn the details of what factor analysis does. It has both the theoretical formulas and simple worked examples to make following along very easy.)

D

D'Agostino, R.B., Belanger, A., D'Agostino, R.B. Jr. 1990."A Suggestion for Using Powerful and Informative Tests of Normality.", *The American Statistician*, November 1990, Volume 44 Number 4, pages 316-321. This tutorial style article discusses D'Agostino’s tests and tells how to interpret normal probability plots.

D'Agostino, R.B., Chase, W., Belanger, A. 1988."The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations.", *The American Statistician,* August 1988, Volume 42 Number 3, pages 198-202.

D'Agostino, R.B. Jr. 2004. *Tutorials in Biostatistics*. Volume 1. John Wiley & Sons. Chichester, England.

Dallal, G. 1986. "An Analytic Approximation to the Distribution of Lilliefors's Test Statistic for Normality," *The American Statistician*, Volume 40, Number 4, pages 294-296.

Daniel, C. and Wood, F. 1980. *Fitting Equations to Data*. John Wiley & Sons. New York. This book gives several in depth examples of analyzing regression problems by computer.

Daniel, W. 1990. *Applied Nonparametric Statistics.* 2nd ed. PWS-KENT Publishing Company. Boston.

Davies, Owen L. 1971. *The Design and Analysis of Industrial Experiments*. Hafner Publishing Company, New York. This was one of the first books on experimental design and analysis. It has many examples and is highly recommended.

Davis, J. C. 1985. *Statistics and Data Analysis in Geology*. John Wiley. New York. (A great layman's discussion of many statistical procedures, including factor analysis.)

Davison, A.C. and Hinkley, D.V. 1999. *Bootstrap Methods and their Applications*. Cambridge University Press. NY, NY. This book provides and detailed account of bootstrapping.

Davison, Mark. 1983. *Multidimensional Scaling*. John Wiley & Sons. NY, NY. This book provides a very good, although somewhat advanced, introduction to the subject.

DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. 1988. “Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.” *Biometrics*, 44, pages 837-845.

DeMets, D.L. and Lan, K.K.G. 1984. “An overview of sequential methods and their applications in clinical trials.” *Communications in Statistics, Theory and Methods,* 13, pages 2315-2338.

DeMets, D.L. and Lan, K.K.G. 1994. “Interim analysis: The alpha spending function approach.” *Statistics in Medicine*, 13, pages 1341-1352.

Demidenko, E. 2004. *Mixed Models – Theory and Applications*. John Wiley & Sons. Hoboken, New Jersey.

Desu, M. M. and Raghavarao, D. 1990. *Sample Size Methodology*. Academic Press. New York. (Presents many useful results for determining sample sizes.)

DeVor, Chang, and Sutherland. 1992. *Statistical Quality Design and Control*. Macmillan Publishing. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 800 pages.

Devroye, Luc. 1986. *Non-Uniform Random Variate Generation.* Springer-Verlag. New York. This book is currently available online at http://jeff.cs.mcgill.ca/~luc/rnbookindex.html.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. 1994. *Analysis of Longitudinal Data*. Oxford University Press. New York, New York.

Dillon, W. and Goldstein, M. 1984. *Multivariate Analysis - Methods and Applications*. John Wiley. NY, NY. This book devotes a complete chapter to loglinear models. It follows Fienberg's book, providing additional discussion and examples.

Dixon, W. J. and Tukey, J. W. 1968. “Approximate behavior of the distribution of Winsorized t,” *Technometrics*, Volume 10, pages 83-98.

Dodson, B. 1994. *Weibull Analysis*. ASQC Quality Press. Milwaukee, Wisconsin. This paperback book provides the basics of Weibull fitting. It contains many of the formulas used in our Weibull procedure.

Donnelly, Thomas G. 1980. “ACM Algorithm 462: Bivariate Normal Distribution,” *Collected Algorithms from ACM*, Volume II, New York, New York.

Donner, Allan. 1984. “Approaches to Sample Size Estimation in the Design of Clinical Trials--A Review,” *Statistics in Medicine*, Volume 3, pages 199-214. This is a well done review of the clinical trial literature. Although it is becoming out of date, it is still a good place to start.

Donner, A. and Klar, N. 1996. “Statistical Considerations in the Design and Analysis of Community Intervention Trials.” *The Journal of Clinical Epidemiology*, Vol. 49, No. 4, 1996, pages 435-439.

Donner, A. and Klar, N. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold. London.

Draghici, S. 2003. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC. London. This is an excellent overview of most areas of Microarray analysis.

Draper, N.R. and Smith, H. 1966. *Applied Regression Analysis*. John Wiley & Sons. New York. This is a classic text in regression analysis. It contains both in depth theory and applications. This text is often used in graduate courses in regression analysis.

Draper, N.R. and Smith, H. 1981. *Applied Regression Analysis - Second Edition*. John Wiley & Sons. New York, NY. This is a classic text in regression analysis. It contains both in-depth theory and applications. It is often used in graduate courses in regression analysis.

Dudoit, S., Shaffer, J.P., and Boldrick, J.C. 2003. “Multiple Hypothesis Testing in Microarray Experiments,” *Statistical Science*, Volume 18, No. 1, pages 71-103.

Dudoit, S., Yang, Y.H., Callow, M.J., and **Speed, T.P.** 2002. “Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Experiments,” *Statistica Sinica*, Volume 12, pages 111-139.

du Toit, S.H.C., Steyn, A.G.W., and Stumpf, R.H. 1986. *Graphical Exploratory Data Analysis.* Springer-Verlag. New York. This book contains examples of graphical analysis for a broad range of topics.

Dunn, O. J. 1964. “Multiple comparisons using rank sums,” *Technometrics*, Volume 6, pages 241-252.

Dunnett, C. W. 1955. “A Multiple comparison procedure for Comparing Several Treatments with a Control,” *Journal of the American Statistical Association*, Volume 50, pages 1096-1121.

Dunteman, G.H. 1989. *Principal Components Analysis*. Sage University Papers, 07-069. Newbury Park, California. Telephone (805) 499-0721. This monograph costs only $7. It gives a very good introduction to PCA.

Dupont, William. 1988. “Power Calculations for Matched Case-Control Studies,” *Biometrics*, Volume 44, pages 1157-1168.

Dupont, William and **Plummer, Walton D.** 1990. “Power and Sample Size Calculations--A Review and Computer Program,” *Controlled Clinical Trials*, Volume 11, pages 116-128. Documents a nice public-domain program on sample size and power analysis.

Durbin, J. and Watson, G. S. 1950. “Testing for Serial Correlation in Least Squares Regression - I,” *Biometrika*, Volume 37, pages 409-428.

Durbin, J. and Watson, G. S. 1951. “Testing for Serial Correlation in Least Squares Regression - II,” *Biometrika*, Volume 38, pages 159-177.

Dyke, G.V. and Patterson, H.D. 1952. “Analysis of factorial arrangements when the data are proportions.” *Biometrics*. Volume 8, pages 1-12. This is the source of the data used in the LLM tutorial.

E

Eckert, Joseph K. 1990. *Property Appraisal and Assessment Administration*. International Association of Assessing Officers. 1313 East 60th Street. Chicago, IL 60637-2892. Phone: (312) 947-2044. This is a how-to manual published by the IAAO that describes how to apply many statistical procedures to real estate appraisal and tax assessment. We strongly recommend it to those using our *Assessment Model* procedure.

Edgington, E. 1987. *Randomization Tests*. Marcel Dekker. New York. A comprehensive discussion of randomization tests with many examples.

Edwards, L.K. 1993. *Applied Analysis of Variance in the Behavior Sciences*. Marcel Dekker. New York. Chapter 8 of this book is used to validate the repeated measures module of PASS.

Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall. New York.

Elandt-Johnson, R.C. and Johnson, N.L. 1980. *Survival Models and Data Analysis*. John Wiley. NY, NY. This book devotes several chapters to population and clinical life-table analysis.

Epstein, Benjamin. 1960. “Statistical Life Test Acceptance Procedures.” *Technometrics*. Volume 2.4, pages 435-446.

Everitt, B.S. and Dunn, G. 1992. *Applied Multivariate Data Analysis*. Oxford University Press. New York. This book provides a very good introduction to several multivariate techniques. It helps you understand how to interpret the results.

F

Farrington, C. P. and Manning, G. 1990. “Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.” *Statistics in Medicine*, Vol. 9, pages 1447-1454. This article contains the formulas used for the Equivalence of Proportions module in PASS.

Feldt, L.S.; Woodruff, D.J.; & Salih, F.A. 1987. “Statistical inference for coefficient alpha.” *Applied Psychological Measurement*, Vol. 11, pages 93-103.

Feldt, L.S.; Ankenmann, R.D. 1999. “Determining Sample Size for a Test of the Equality of Alpha Coefficients When the Number of Part-Tests is Small.” *Psychological Methods*, Vol. 4(4), pages 366-377.

Fienberg, S. 1985. *The Analysis of Cross-Classified Categorical Data*. MIT Press. Cambridge, Massachusetts. This book provides a very good introduction to the subject. It is a must for any serious student of the subject.

Finney, D. 1971. *Probit Analysis*. Cambridge University Press. New York, N.Y.

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge University Press. New York, New York.

Fisher, R.A. 1936. “The use of multiple measurements in taxonomic problems.” *Annuals of Eugenics*, Volume 7, Part II, 179-188. This article is famous because in it Fisher included the ‘iris data’ that is always presented when discussing discriminant analysis.

Fleiss, Joseph L. 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, J. L., Levin, B., Paik, M.C. 2003. *Statistical Methods for Rates and Proportions. Third Edition.* John Wiley & Sons. New York. This book provides a very good introduction to the subject.

Fleiss, Joseph L. 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons. New York. This book provides a very good introduction to clinical trials. It may be a bit out of date now, but it is still very useful.

Fleming, T. R. 1982. "One-sample multiple testing procedure for Phase II clinical trials." *Biometrics*, Volume 38, pages 143-151.

Flury, B. and Riedwyl, H. 1988. *Multivariate Statistics: A Practical Approach*. Chapman and Hall. New York. This is a short, paperback text that provides lots of examples.

Flury, B. 1988. *Common Principal Components and Related Multivariate Models*. John Wiley & Sons. New York. This reference describes several advanced PCA procedures.

G

Gans. 1984. “The Search for Significance: Different Tests on the Same Data.” *The Journal of Statistical Computation and Simulation*, 1984, pages 1-21.

Gart, John J. and Nam, Jun-mo. 1988. "Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness." *Biometrics*, Volume 44, Issue 2, 323-338.

Gart, John J. and Nam, Jun-mo. 1990. "Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables." *Biometrics*, Volume 46, Issue 3, 637-643.

Gehlback, Stephen. 1988. *Interpreting the Medical Literature: Practical Epidemiology for Clinicians*. Second Edition. McGraw-Hill. New York. Telephone: (800)722-4726. The preface of this book states that its purpose is to provide the reader with a useful approach to interpreting the quantitative results given in medical literature. We reference it specifically because of its discussion of ROC curves.

Gentle, James E. 1998. *Random Number Generation and Monte Carlo Methods*. Springer. New York.

Gibbons, J. 1976. *Nonparametric Methods for Quantitative Analysis*. Holt, Rinehart and Winston. New York.

Gleason, T.C. and Staelin, R. 1975. "A proposal for handling missing data." *Psychometrika*, 40, 229-252.

Goldstein, Richard. 1989. “Power and Sample Size via MS/PC-DOS Computers,” *The American Statistician*, Volume 43, Number 4, pages 253-260. A comparative review of power analysis software that was available at that time.

Gomez, K.A. and Gomez, A. A. 1984. *Statistical Procedures for Agricultural Research*. John Wiley & Sons. New York. This reference contains worked-out examples of many complex ANOVA designs. It includes split-plot designs. We recommend it.

Graybill, Franklin. 1961. *An Introduction to Linear Statistical Models*. McGraw-Hill. New York, New York. This is an older book on the theory of linear models. It contains a few worked examples of power analysis.

Greenacre, M. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press. Orlando, Florida. This book goes through several examples. It is probably the most complete book in English on the subject.

Greenacre, Michael J. 1993. *Correspondence Analysis in Practice*. Academic Press. San Diego, CA. This book provides a self-teaching course in correspondence analysis. It is the clearest exposition on the subject that I have every seen. If you want to gain an understanding of CA, you must obtain this (paperback) book.

Griffiths, P. and Hill, I.D. 1985. *Applied Statistics Algorithms*, The Royal Statistical Society, London, England. See page 243 for ACM algorithm 291.

Gross and Clark 1975. *Survival Distributions*: Reliability Applications in Biomedical Sciences. John Wiley, New York.

Gu, X.S., and Rosenbaum, P.R. 1993. “Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms,” *Journal of Computational and Graphical Statistics*, Vol. 2, No. 4, pages 405-420.

Guenther, William C. 1977. “Desk Calculation of Probabilities for the Distribution of the Sample Correlation Coefficient,” *The American Statistician*, Volume 31, Number 1, pages 45-48.

Guenther, William C. 1977. *Sampling Inspection in Statistical Quality Control*. Griffin’s Statistical Monographs, Number 37. London.

H

Haberman, S.J. 1972. “Loglinear Fit of Contingency Tables.” *Applied Statistics*. Volume 21, pages 218-225. This lists the fortran program that is used to create our LLM algorithm.

Hahn, G. J. and Meeker, W.Q. 1991. *Statistical Intervals*. John Wiley & Sons. New York.

Hambleton, R.K; Swaminathan, H; Rogers, H.J. 1991. *Fundamentals of Item Response Theory*. Sage Publications. Newbury Park, California. Phone: (805)499-0721. Provides an inexpensive, readable introduction to IRT. A good place to start.

Hamilton, L. 1991. *Regression with Graphics: A Second Course in Applied Statistics*. Brooks/Cole Publishing Company. Pacific Grove, California. This book gives a great introduction to the use of graphical analysis with regression. It is a must for any serious user of regression. It is written at an introductory level.

Hand, D.J. and Taylor, C.C. 1987. *Multivariate Analysis of Variance and Repeated Measures*. Chapman and Hall. London, England.

Hanley, J. A. and McNeil, B. J. 1982. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, 143, 29-36. April, 1982.

Hanley, J. A. and McNeil, B. J. 1983. "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases." *Radiology*, 148, 839-843. September, 1983.

Hartigan, J. 1975. *Clustering Algorithms*. John Wiley. New York. (This is the “bible” of cluster algorithms. Hartigan developed the K-means algorithm used in **NCSS**.)

Haupt, R.L. and Haupt, S.E. 1998. *Practical Genetic Algorithms*. John Wiley. New York.

Hernandez-Bermejo, B. and Sorribas, A. 2001. “Analytical Quantile Solution for the S-distribution, Random Number Generation and Statistical Data Modeling.” *Biometrical Journal* 43, 1007-1025.

Hintze, J. L. and Nelson, R.D. 1998. "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician* 52, 181-184.

Hoaglin, Mosteller, and Tukey. 1985. *Exploring Data Tables, Trends, and Shapes*. John Wiley. New York.

Hoaglin, Mosteller, and Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons. New York.

Hochberg, Y. and Tamhane, A. C. 1987. *Multiple Comparison Procedures*. John Wiley & Sons. New York.

Hoerl, A.E. and Kennard, R.W. 1970. "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 55-82.

Hoerl, A.E. and Kennard R.W. 1976. "Ridge regression: Iterative estimation of the biasing parameter." *Communications in Statistics* A5, 77-88.

Howe, W.G. 1969. “Two-Sided Tolerance Limits for Normal Populations—Some Improvements.” *Journal of the American Statistical Association*, 64, 610-620.

Hosmer, D. and Lemeshow, S. 1989. *Applied Logistic Regression*. John Wiley & Sons. New York. This book gives an advanced, in depth look at logistic regression.

Hosmer, D. and Lemeshow, S. 1999. *Applied Survival Analysis*. John Wiley & Sons. New York.

Hotelling, H. 1933. "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology* 24, 417-441, 498-520.

Hsieh, F.Y. 1989. “Sample Size Tables for Logistic Regression,” *Statistics in Medicine*, Volume 8, pages 795-802. This is the article that was the basis for the sample size calculations in logistic regression in PASS 6.0. It has been superceded by the 1998 article.

Hsieh, F.Y., Block, D.A., and Larsen, M.D. 1998. “A Simple Method of Sample Size Calculation for Linear and Logistic Regression,” *Statistics in Medicine*, Volume 17, pages 1623-1634. The sample size calculation for logistic regression in PASS are based on this article.

Hsieh, F.Y. and Lavori, P.W. 2000. “Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates,” *Controlled Clinical Trials*, Volume 21, pages 552-560. The sample size calculation for Cox regression in PASS are based on this article.

Hsu, Jason. 1996. *Multiple Comparisons: Theory and Methods.* Chapman & Hall. London. This book gives a beginning to intermediate discussion of multiple comparisons, stressing the interpretation of the various MC tests. It provides details of three popular MC situations: all pairs, versus the best, and versus a control. The power calculations used in the MC module of PASS came from this book.

I

Irizarry, R.A., et al. 2003a. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4, 249-264.

Irizarry, R.A., et al. 2003b. Summaries of Affymetrix GeneChip Probe Level Data. *Nucleic Acids Research*, 31, e15.

J

Jackson, J.E. 1991. *A User's Guide To Principal Components.* John Wiley & Sons. New York. This is a great book to learn about PCA from. It provides several examples and treats everything at a level that is easy to understand.

James, Mike. 1985. *Classification Algorithms*. John Wiley & Sons. New York. This is a great text on the application of discriminant analysis. It includes a simple, easy-to-understand, theoretical development as well as discussions of the application of discriminant analysis.

Jammalamadaka, S.R. and SenGupta, A. 2001. *Topics in Circular Statistics*. World Scientific. River Edge, New Jersey.

Jobson, J.D. 1992. *Applied Multivariate Data Analysis - Volume II: Categorical and Multivariate Methods*. Springer-Verlag. New York. This book is a useful reference for loglinear models and other multivariate methods. It is easy to follows and provides lots of examples.

Jolliffe, I.T. 1972. "Discarding variables in a principal component analysis, I: Artifical data." *Applied Statistics*, 21:160-173.

Johnson, N.L., Kotz, S., and Kemp, A.W. 1992. *Univariate Discrete Distributions, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions Volume 1, Second Edition*. John Wiley & Sons. New York.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1995. *Continuous Univariate Distributions Volume 2, Second Edition*. John Wiley & Sons. New York.

Jolliffe, I.T. 1986. *Principal Component Analysis*. Springer-Verlag. New York. This book provides an easy-reading introduction to PCA. It goes through several examples.

Julious, Steven A. 2004. "Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data." *Statistics in Medicine*, 23:1921-1986.

Jung, S.-H. 2005. "Sample size for FDR-control in microarray data analysis" *Bioinformatics*, 21(14):3097-3104.

Juran, J.M. 1979. *Quality Control Handbook*. McGraw-Hill. New York.

K

Kaiser, H.F. 1960. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement*. 20:141-151.

Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Karian, Z.A and Dudewicz, E.J. 2000. *Fitting Statistical Distributions.* CRC Press, New York.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data*. John Wiley. New York. This book gives an excellent introduction to cluster analysis. It treats the forming of the distance matrix and several different types of cluster methods, including fuzzy. All this is done at an elementary level so that users at all levels can gain from it.

Kay, S.M. 1988. *Modern Spectral Estimation*. Prentice-Hall: Englewood Cliffs, New Jersey. A very technical book on spectral theory.

Kendall,M. and Ord, J.K. 1990. *Time Series.* Oxford University Press. New York. This is theoretical introduction to time series analysis that is very readable.

Kendall,M. and Stuart, A. 1987. *Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory.* Oxford University Press. New York. This is a fine math-stat book for graduate students in statistics. We reference it because it includes formulas that are used in the program.

Kenward, M. G. and Roger, J. H. 1997. “Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood,” *Biometrics*, 53, pages 983-997.

Keppel, Geoffrey. 1991. *Design and Analysis - A Researcher’s Handbook.* Prentice Hall. Englewood Cliffs, New Jersey. This is a very readable primer on the topic of analysis of variance. Recommended for those who want the straight scoop with a few, well-chosen examples.

Kirk, Roger E. 1982. *Experimental Design: Procedures for the Behavioral Sciences.* Brooks/Cole. Pacific Grove, California. This is a respected reference on experimental design and analysis of variance.

Klein, J.P. and Moeschberger, M.L.. 1997. *Survival Analysis.* Springer-Verlag. New York. This book provides a comprehensive look at the subject complete with formulas, examples, and lots of useful comments. It includes all the more recent developments in this field. I recommend it.

Koch, G.G.; Atkinson, S.S.; Stokes, M.E. 1986. *Encyclopedia of Statistical Sciences.* Volume 7. John Wiley. New York. Edited by Samuel Kotz and Norman Johnson. The article on Poisson Regression provides a very good summary of the subject.

Kotz and Johnson. 1993. *Process Capability Indices.* Chapman & Hall. New York. This book gives a detailed account of the capability indices used in SPC work. 207 pages.

Kraemer, H. C. and **Thiemann, S.** 1987. *How Many Subjects*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Kruskal, J. 1964. “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.” *Psychometrika* 29, pages 1-27, 115-129. This article presents the algorithm on which the non-metric algorithm used in NCSS is based.

Kruskal, J. and Wish, M. 1978. *Multidimensional Scaling*. Sage Publications. Beverly Hills, CA. This is a well-written monograph by two of the early pioneers of MDS. We suggest it to all serious students of MDS.

Kuehl, R.O. 2000. *Design of Experiment: Statistical Principles of Research Design and Analysis, 2nd Edition.* Brooks/Cole. Pacific Grove, California. This is a good graduate level text on experimental design with many examples.

L

Lachenbruch, P.A. 1975. *Discriminant Analysis.* Hafner Press. New York. This is an in-depth treatment of the subject. It covers a lot of territory, but has few examples.

Lachin, John M. 2000. *Biostatistical Methods.* John Wiley & Sons. New York. This is a graduate-level methods book that deals with statistical methods that are of interest to biostatisticians such as odds ratios, relative risks, regression analysis, case-control studies, and so on.

Lachin, John M. and **Foulkes, Mary A. 1986.** “Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification,” *Biometrics*, Volume 42, September, pages 507-516.

Lan, K.K.G. and DeMets, D.L. 1983. “Discrete sequential boundaries for clinical trials.” *Biometrika*, 70, pages 659-663.

Lan, K.K.G. and Zucker, D.M. 1993. “Sequential monitoring of clinical trials: the role of information and Brownian motion.” *Statistics in Medicine*, 12, pages 753-765.

Lance, G.N. and Williams, W.T. 1967. “A general theory of classificatory sorting strategies. I. Hierarchical systems.” *Comput. J.* 9, pages 373-380.

Lance, G.N. and Williams, W.T. 1967. “Mixed-data classificatory programs I. Agglomerative systems.” *Aust.Comput. J.* 1, pages 15-20.

Lawless, J.F. 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley, New York.

Lawson, John. 1987. *Basic Industrial Experimental Design Strategies*. Center for Statistical Research at Brigham Young University. Provo, Utah. 84602. This is a manuscript used by Dr. Lawson in courses and workshops that he provides to industrial engineers. It is the basis for many of our experimental design procedures.

Lebart, Morineau, and Warwick. 1984. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons. This book devotes a large percentage of its discussion to correspondence analysis.

Lee, E.T. 1974. "A Computer Program for Linear Logistic Regression Analysis" in *Computer Programs in Biomedicine*, Volume 4, pages 80-92.

Lee, E.T. 1980. *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications. Belmont, California.

Lee, E.T. 1992. *Statistical Methods for Survival Data Analysis*. Second Edition. John Wiley & Sons. New York. This book provides a very readable introduction to survival analysis techniques.

Lee, M.-L. T. 2004. *Analysis of Microarray Gene Expression Data*. Kluwer Academic Publishers. Norwell, Massachusetts.

Lee, S. K. 1977. “On the Asymptotic Variances of u Terms in Loglinear Models of Multidimensional Contingency Tables.” *Journal of the American Statistical Association*. Volume 72 (June, 1977), page 412. This article describes methods for computing standard errors that are used in the LLM section of this program.

Lenth, Russell V. 1987. “Algorithm AS 226: Computing Noncentral Beta Probabilities,” *Applied Statistics*, Volume 36, pages 241-244.

Lenth, Russell V. 1989. “Algorithm AS 243: Cumulative Distribution Function of the Non-central t Distribution,” *Applied Statistics*, Volume 38, pages 185-189.

Lesaffre, E. and Albert, A. 1989. “Multiple-group Logistic Regression Diagnostics” *Applied Statistics*, Volume 38, pages 425-440. See also Pregibon 1981.

Levene, H. 1960. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al., eds. Stanford University Press, Stanford Calif., pp. 278-292.

Lewis, J.A. 1999. “Statistical principles for clinical trials (ICH E9) an introductory note on an international guideline.” *Statistics in Medicine*, 18, pages 1903-1942.

Lipsey, Mark W. 1990. *Design Sensitivity Statistical Power for Experimental Research*, Sage Publications, 2111 West Hillcrest Drive, Newbury Park, CA. 91320. This is an excellent introduction to power analysis.

Little, R. and Rubin, D. 1987. *Statistical Analysis with Missing Data*. John Wiley & Sons. New York. This book is completely devoted to dealing with missing values. It gives a complete treatment of using the EM algorithm to estimate the covariance matrix.

Little, R. C. et al. 2006. *SAS for Mixed Models – Second Edition*. SAS Institute Inc., Cary, North Carolina.

Liu, H. and Wu, T. 2005. "Sample Size Calculation and Power Analysis of Time-Averaged Difference," *Journal of Modern Applied Statistical Methods*, Vol. 4, No. 2, pages 434-445.

Lu, Y. and Bean, J.A. 1995. “On the sample size for one-sided equivalence of sensitivities based upon McNemar’s test,” *Statistics in Medicine*, Volume 14, pages 1831-1839.

Lui, J., Hsueh, H., Hsieh, E., and Chen, J.J. 2002. “Tests for equivalence or non-inferiority for paired binary data,” *Statistics in Medicine*, Volume 21, pages 231-245.

Lloyd, D.K. and Lipow, M. 1991. *Reliability: Management, Methods, and Mathematics*. ASQC Quality Press. Milwaukee, Wisconsin.

Locke, C.S. 1984. “An exact confidence interval for untransformed data for the ratio of two formulation means,” *J. Pharmacokinet. Biopharm.*, Volume 12, pages 649-655.

Lockhart, R. A. & Stephens, M. A. 1985. “Tests of fit for the von Mises distribution.” *Biometrika* 72, pages 647-652.

M

Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. *Sample Size Tables for Clinical Studies, 2nd Edition*. Blackwell Science. Malden, Mass. A very good & easy to read book on determining appropriate sample sizes in many situations.

Makridakis, S. and Wheelwright, S.C. 1978. *Iterative Forecasting*. Holden-Day.: San Francisco, California. This is a very good book for the layman since it includes several detailed examples. It is written for a person with a minimum amount of mathematical background.

Manly, B.F.J. 1986. *Multivariate Statistical Methods - A Primer*. Chapman and Hall. New York. This nice little paperback provides a simplified introduction to many multivariate techniques, including MDS.

Mardia, K.V. and Jupp, P.E. 2000. *Directional Statistics*. John Wiley & Sons. New York.

Marple, S.L. 1987. *Digital Spectral Analysis with Applications*. Prentice-Hall: Englewood Cliffs, New Jersey. A technical book about spectral analysis.

Martinez and Iglewicz. 1981. "A test for departure from normality based on a biweight estimator of scale." *Biometrika*, 68, 331-333).

Marubini, E. and **Valsecchi, M.G.** 1996. *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley: New York, New York.

Mather, Paul. 1976. *Computational Methods of Multivariate Analysis in Physical Geography*. John Wiley & Sons. This is a great book for getting the details on several multivariate procedures. It was written for non-statisticians. It is especially useful in its presentation of cluster analysis. Unfortunately, it is out-of-print. You will have to look for it in a university library (it is worth the hunt).

Matsumoto, M. and Nishimura,T. 1998. "Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator" *ACM Trans. On Modeling and Computer Simulations*.

Mauchly, J.W. 1940. "Significance test for sphericity of a normal n-variate distribution." *Annals of Mathematical Statistics*, 11: 204-209

McCabe, G.P. 1984. "Principal variables." *Technometrics*, 26, 137-144.

McClish, D.K. 1989. "Analyzing a Portion of the ROC Curve." *Medical Decision Making*, 9: 190-195

McHenry, Claude. 1978. "Multivariate subset selection." *Journal of the Royal Statistical Society, Series C*. Volume 27, No. 23, pages 291-296.

McNeil, D.R. 1977. *Interactive Data Analysis*. John Wiley & Sons. New York.

Mendenhall, W. 1968. *Introduction to Linear Models and the Design and Analysis of Experiments*. Wadsworth. Belmont, Calif.

Metz, C.E. 1978. "Basic principles of ROC analysis." *Seminars in Nuclear Medicine*, Volume 8, No. 4, pages 283-298.

Miettinen, O.S. and Nurminen, M. 1985. "Comparative analysis of two rates." *Statistics in Medicine* 4: 213-226.

Milliken, G.A. and Johnson, D.E. 1984. *Analysis of Messy Data, Volume 1*. Van Nostrand Rienhold. New York, NY.

Milne, P. 1987. *Computer Graphics for Surveying.* E. & F. N. Spon, 29 West 35th St., NY, NY 10001

Montgomery, Douglas. 1984. *Design and Analysis of Experiments*. John Wiley & Sons, New York. A textbook covering a broad range of experimental design methods. The book is not limited to industrial investigations, but gives a much more general overview of experimental design methodology.

Montgomery, Douglas and Peck. 1992. *Introduction to Linear Regression Analysis*. A very good book on this topic.

Montgomery, Douglas C. 1991. *Introduction to Statistical Quality Control.* Second edition. John Wiley & Sons. New York. This is a comprehensive textbook of SPC including control charts, process capability, and experimental design. It has many examples. 700 pages.

Moore, D. S. and McCabe, G. P. 1999. *Introduction to the Practice of Statistics*. W. H. Freeman and Company. New York.

Mosteller, F. and Tukey, J.W. 1977. *Data Analysis and Regression*. Addison-Wesley. Menlo Park, California. This book should be read by all serious users of regression analysis. Although the terminology is a little different, this book will give you a fresh look at the whole subject.

Motulsky, Harvey. 1995. *Intuitive Biostatistics.* Oxford University Press. New York, New York. This is a wonderful book for those who want to understand the basic concepts of statistical testing. The author presents a very readable coverage of the most popular biostatistics tests. If you have forgotten how to interpret the various statistical tests, get this book!

Moura, Eduardo C. 1991. *How To Determine Sample Size And Estimate Failure Rate in Life Testing*. ASQC Quality Press. Milwaukee, Wisconsin.

Mueller, K. E., and Barton, C. N. 1989. "Approximate Power for Repeated-Measures ANOVA Lacking Sphericity." *Journal of the American Statistical Association*, Volume 84, No. 406, pages 549-555.

Mueller, K. E., LaVange, L.E., Ramey, S.L., and Ramey, C.T. 1992. "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications." *Journal of the American Statistical Association*, Volume 87, No. 420, pages 1209-1226.

Mukerjee, H., Robertson, T., and Wright, F.T. 1987. "Comparison of Several Treatments With a Control Using Multiple Contrasts." *Journal of the American Statistical Association*, Volume 82, No. 399, pages 902-910.

Muller, K. E. and Stewart, P.W. 2006. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley & Sons Inc. Hoboken, New Jersey.

Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, Massachusetts. This is one of the bibles on the topic of regression analysis.

N

Naef, F. et al. 2002. "Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays," *Genome Biol.*, 3, RESEARCH0018.

Nam, Jun-mo. 1992. "Sample Size Determination for Case-Control Studies and the Comparison of Stratified and Unstratified Analyses," *Biometrics*, Volume 48, pages 389-395.

Nam, Jun-mo. 1997. "Establishing equivalence of two treatments and sample size requirements in matched-pairs design," *Biometrics*, Volume 53, pages 1422-1430.

Nam, J-m. and Blackwelder, W.C. 2002. "Analysis of the ratio of marginal probabilities in a matched-pair setting," *Statistics in Medicine*, Volume 21, pages 689-699.

Nash, J. C. 1987. *Nonlinear Parameter Estimation*. Marcel Dekker, Inc. New York, NY.

Nash, J.C. 1979. *Compact Numerical Methods for Computers*. John Wiley & Sons. New York, NY.

Nel, D.G. and van der Merwe, C.A. 1986. "A solution to the multivariate Behrens-Fisher problem." *Communications in Statistics—Series A, Theory and Methods*, 15, pages 3719-3735.

Nelson, W.B. 1982. *Applied Life Data Analysis*. John Wiley, New York.

Nelson, W.B. 1990. *Accelerated Testing*. John Wiley, New York.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. 1996. *Applied Linear Statistical Models*. Richard D. Irwin, Inc. Chicago, Illinois. This mammoth book covers regression analysis and analysis of variance thoroughly and in great detail. We recommend it.

Neter, J., Wasserman, W., and Kutner, M. 1983. *Applied Linear Regression Models*. Richard D. Irwin, Inc. Chicago, Illinois. This book provides you with a complete introduction to the methods of regression analysis. We suggest it to non-statisticians as a great reference tool.

Newcombe, Robert G. 1998a. "Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods." *Statistics in Medicine*, Volume 17, 857-872.

Newcombe, Robert G. 1998b. "Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods." *Statistics in Medicine*, Volume 17, 873-890.

Newcombe, Robert G. 1998c. "Improved Confidence Intervals for the Difference Between Binomial Proportions Based on Paired Data." *Statistics in Medicine*, Volume 17, 2635-2650.

Newton, H.J. 1988. *TIMESLAB: A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole: Pacific Grove, California. This book is loaded with theoretical information about time series analysis. It includes software designed by Dr. Newton for performing advanced time series and spectral analysis. The book requires a strong math and statistical background.

O

O'Brien, P.C. and Fleming, T.R. 1979. "A multiple testing procedure for clinical trials." *Biometrics,* 35, pages 549-556.

O'Brien, R.G. and Kaiser, M.K. 1985. "MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer." *Psychological Bulletin,* 97, pages 316-333.

Obuchowski, N. 1998. "Sample Size Calculations in Studies of Test Accuracy." *Statistical Methods in Medical Research,* 7, pages 371-392.

Obuchowski, N. and McClish, D. 1997. "Sample Size Determination for Diagnostic Accuracy Studies Involving Binormal ROC Curve Indices." *Statistics in Medicine,* 16, pages 1529-1542.

Odeh, R.E. and Fox, M. 1991. *Sample Size Choice*. Marcel Dekker, Inc. New York, NY.

O'Neill and Wetherill. 1971 "The Present State of Multiple Comparison Methods,*" The Journal of the Royal Statistical Society*, Series B, vol.33, 218-250).

Orloci, L. & Kenkel, N. 1985. *Introduction to Data Analysis*. International Co-operative Publishing House. Fairland, Maryland. This book was written for ecologists. It contains samples and BASIC programs of many statistical procedures. It has one brief chapter on MDS, and it includes a non-metric MDS algorithm.

Ostle, B. 1988. *Statistics in Research. Fourth Edition*. Iowa State Press. Ames, Iowa. A comprehension book on statistical methods.

Ott, L. 1977. *An Introduction to Statistical Methods and Data Analysis*. Wadsworth. Belmont, Calif. Use the second edition.

Ott, L. 1984. *An Introduction to Statistical Methods and Data Analysis, Second Edition*. Wadsworth. Belmont, Calif. This is a complete methods text. Regression analysis is the focus of five or six chapters. It stresses the interpretation of the statistics rather than the calculation, hence it provides a good companion to a statistical program like ours.

Owen, Donald B. 1956. "Tables for Computing Bivariate Normal Probabilities," *Annals of Mathematical Statistics*, Volume 27, pages 1075-1090.

Owen, Donald B. 1965. "A Special Case of a Bivariate Non-Central t-Distribution," *Biometrika*, Volume 52, pages 437-446.

P

Pandit, S.M. and Wu, S.M. 1983. *Time Series and System Analysis with Applications*. John Wiley and Sons. New York. This book provides an alternative to the Box-Jenkins approach for dealing with ARMA models. We used this approach in developing our automatic ARMA module.

Parmar, M.K.B. and Machin, D. 1995. *Survival Analysis*. John Wiley and Sons. New York.

Parmar, M.K.B., Torri, V., and Steart, L. 1998. "Extracting Summary Statistics to Perform Meta-Analyses of the Published Literature for Survival Endpoints." *Statistics in Medicine* 17, 2815-2834.

Pearson, K. 1901. "On lines and planes of closest fit to a system of points in space." *Philosophical Magazine* 2, 557-572.

Pan, Z. and Kupper, L. 1999. "Sample Size Determination for Multiple Comparison Studies Treating Confidence Interval Width as Random." *Statistics in Medicine* 18, 1475-1488.

Pedhazur, E.L. and Schmelkin, L.P. 1991. *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. This mammoth book (over 800 pages) covers multivariate analysis, regression analysis, experimental design, analysis of variance, and much more. It provides annotated output from SPSS and SAS which is also useful to our users. The text is emphasizes the social sciences. It provides a "how-to," rather than a theoretical, discussion. Its chapters on factor analysis are especially informative.

Phillips, Kem F. 1990. “Power of the Two One-Sided Tests Procedure in Bioequivalence,” *Journal of Pharmacokinetics and Biopharmaceutics*, Volume 18, No. 2, pages 137-144.

Pocock, S.J. 1977. “Group sequential methods in the design and analysis of clinical trials.” *Biometrika*, 64, pages 191-199.

Press, S. J. and Wilson, S. 1978. “Choosing Between Logistic Regression and Discriminant Analysis.” *Journal of the American Statistical Association*, Volume 73, Number 364, Pages 699-705. This article details the reasons why logistic regression should be the preferred technique.

Press, William H. 1986. *Numerical Recipes*, Cambridge University Press, New York, New York.

Pregibon, Daryl. 1981. “Logistic Regression Diagnostics.” *Annals of Statistics*, Volume 9, Pages 705-725. This article details the extensions of the usual regression diagnostics to the case of logistic regression. These results were extended to multiple-group logistic regression in Lesaffre and Albert (1989).

Price, K., Storn R., and Lampinen, J. 2005. *Differential Evolution – A Practical Approach to Global Optimization.* Springer. Berlin, Germany.

Prihoda, Tom. 1983. “Convenient Power Analysis For Complex Analysis of Variance Models.” *Poster Session of the American Statistical Association Joint Statistical Meetings*, August 15-18, 1983, Toronto, Canada. Tom is currently at the University of Texas Health Science Center. This article includes FORTRAN code for performing power analysis.

R

Ramsey, Philip H. 1978 “Power Differences Between Pairwise Multiple Comparisons,” *JASA*, vol. 73, no. 363, pages 479-485.

Rao, C.R. , Mitra, S.K., & Matthai, A. 1966. *Formulae and Tables for Statistical Work.* Statistical Publishing Society, Indian Statistical Institute, Calcutta, India.

Ratkowsky, David A. 1989. *Handbook of Nonlinear Regression Models*. Marcel Dekker. New York. A good, but technical, discussion of various nonlinear regression models.

Rawlings John O. 1988. *Applied Regression Analysis: A Research Tool*. Wadsworth. Belmont, California. This is a readable book on regression analysis. It provides a thorough discourse on the subject.

Reboussin, D.M., DeMets, D.L., Kim, K, and Lan, K.K.G. 1992. “Programs for computing group sequential boundaries using the Lan-DeMets Method.” Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

Rencher, Alvin C. 1998. *Multivariate Statistical Inference and Applications*. John Wiley. New York, New York. This book provides a comprehensive mixture of theoretical and applied results in multivariate analysis. My evaluation may be biased since Al Rencher took me fishing when I was his student.

Robins, Greenland, and Breslow. 1986. “A General Estimator for the Variance of the Mantel-Haenszel Odds Ratio,” *American Journal of Epidemiology*, vol.42, pages 719-723.

Robins, Breslow, and Greenland. 1986. “Estimators of the Mantel-Haenszel variance consisten in both sparse data and large-strata limiting models,” *Biometrics*, vol. 42, pages 311-323.

Rosenbaum, P.R. 1989. “Optimal Matching for Observational Studies,” *Journal of the American Statistical Association*, vol. 84, no. 408, pages 1024-1032.

Rosenbaum, P.R., and Rubin, D.B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, vol. 70, pages 41-55.

Rosenbaum, P.R., and Rubin, D.B. 1984. “Reducing bias in observational studies using subclassification on the propensity score,” *Journal of the American Statistical Association*, vol. 79, pages 516-524.

Rosenbaum, P.R., and Rubin, D.B. 1985a. “Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score,” *American Statistician*, vol. 39, pages 33-38.

Rosenbaum, P.R., and Rubin, D.B. 1985b. “The Bias Due to Incomplete Matching,” *Biometrics*, vol. 41, pages 106-116.

Ryan, Thomas P. 1989. *Statistical Methods for Quality Improvement*. John Wiley & Sons. New York. This is a comprehensive treatment of SPC including control charts, process capability, and experimental design. It provides many rules-of-thumb and discusses many non-standard situations. This is a very good ‘operators manual’ type of book. 446 pages.

Ryan, Thomas P. 1997. *Modern Regression Methods*. John Wiley & Sons. New York. This is a comprehensive treatment of regression analysis. The author often deals with practical issues that are left out of other texts.

S

Sahai, Hardeo & Khurshid, Anwer. 1995. *Statistics in Epidemiology*. CRC Press. Boca Raton, Florida.

Schiffman, Reynolds, & Young. 1981. *Introduction to Multidimensional Scaling*. Academic Press. Orlando, Florida. This book goes through several examples.

Schilling, Edward. 1982. *Acceptance Sampling in Quality Control*. Marcel-Dekker. New York.

Schlesselman, Jim. 1981. *Case-Control Studies*. Oxford University Press. New York. This presents a complete overview of case-control studies. It was our primary source for the Mantel-Haenszel test.

Schmee and Hahn. November, 1979. "A Simple Method for Regression Analysis." *Technometrics*, Volume 21, Number 4, pages 417-432.

Schoenfeld, David A. 1983. “Sample-Size Formula for the Proportional-Hazards Regression Model” *Biometrics*, Volume 39, pages 499-503.

Schoenfeld, David A. and **Richter, Jane R.** 1982**.** “Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint,” *Biometrics,* March 1982, Volume 38, pages 163-170.

Schork, M. and Williams, G. 1980. "Number of Observations Required for the Comparison of Two Correlated Proportions." *Communications in Statistics-Simula. Computa.,* B9(4), 349-357.

Schuirmann, Donald. 1981**.** “On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval,” *Biometrics,* Volume 37, pages 617.

Schuirmann, Donald. 1987**.** “A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability,” *Journal of Pharmacokinetics and Biopharmaceutics,* Volume 15, Number 6, pages 657-680.

Seber, G.A.F. 1984. *Multivariate Observations*. John Wiley & Sons. New York. (This book is an encyclopedia of multivariate techniques. It emphasizes the mathematical details of each technique and provides a complete set of references. It will only be useful to those comfortable with reading mathematical equations based on matrices.)

Seber, G.A.F. and Wild, C.J. 1989. *Nonlinear Regression*. John Wiley & Sons. New York. This book is an encyclopedia of nonlinear regression.

Senn, Stephen. 1993. *Cross-over Trials in Clinical Research*. John Wiley & Sons. New York.

Senn, Stephen. 2002. *Cross-over Trials in Clinical Research*. Second Edition. John Wiley & Sons. New York.

Shapiro, S.S. and Wilk, M.B. 1965 "An analysis of Variance test for normality." *Biometrika,* Volume 52, pages 591-611.

Shuster, Jonathan J. 1990. *CRC Handbook of Sample Size Guidelines for Clinical Trials.* CRC Press, Boca Raton, Florida. This is an expensive book ($300) of tables for running log-rank tests. It is well documented, but at this price it better be.

Signorini, David. 1991. “Sample size for Poisson regression,” *Biometrika,* Volume 78, 2, pages 446-450.

Simon, Richard. “Optimal Two-Stage Designs for Phase II Clinical Trials,” *Controlled Clinical Trials,* 1989, Volume 10, pages 1-10.

Snedecor, G. and Cochran, Wm. 1972. *Statistical Methods*. The Iowa State University Press. Ames, Iowa.

Sorribas, A., March, J., and Trujillano, J. 2002. “A new parametric method based on S-distributions for computing receiver operating characteristic curves for continuous diagnostic tests.” *Statistics in Medicine* 21, 1213-1235.

Spath, H. 1985. *Cluster Dissection and Analysis.* Halsted Press. New York. (This book contains a detailed discussion of clustering techniques for large data sets. It contains some heavy mathematical notation.)

Speed, T.P. (editor). 2003. *Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC. Boca Raton, Florida.

Stekel, D. 2003. *Microarray Bioinformatics.* Cambridge University Press. Cambridge, United Kingdom.

Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., and Song, F. 2000. *Methods for Meta-Analysis in Medical Research.* John Wiley & Sons. New York.

Swets, John A. 1996. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics - Collected Papers.* Lawrence Erlbaum Associates. Mahway, New Jersey.

T

Tabachnick, B. and Fidell, L. 1989. *Using Multivariate Statistics.* Harper Collins. 10 East 53d Street, NY, NY 10022. This is an extremely useful text on multivariate techniques. It presents computer printouts and discussion from several popular programs. It provides checklists for each procedure as well as sample written reports. I strongly encourage you to obtain this book!

Tango, Toshiro. 1998. "Equivalence Test and Confidence Interval for the Difference in Proportions for the Paired-Sample Design." *Statistics in Medicine*, Volume 17, 891-908.

Therneau, T.M. and Grambsch, P.M. 2000. *Modeling Survival Data*. Springer: New York, New York. A the time of the writing of the Cox regression procedure, this book provides a thorough, up-to-date discussion of this procedure as well as many extensions to it. Recommended, especially to those with at least a masters in statistics.

Thomopoulos, N.T. 1980. *Applied Forecasting Methods*. Prentice-Hall: Englewood Cliffs, New Jersey. This book contains a very good presentation of the classical forecasting methods discussed in chapter two.

Thompson, Simon G. 1998. *Encyclopedia of Biostatistics, Volume 4*. John Wiley & Sons. New York. Article on Meta-Analysis on pages 2570-2579.

Tiku, M. L. 1965. "Laguerre Series Forms of Non-Central X^2 and F Distributions," *Biometrika*, Volume 42, pages 415-427.

Torgenson, W.S. 1952. "Multidimensional scaling. I. Theory and method." *Psychometrika* 17, 401-419. This is one of the first articles on MDS. There have been many advances, but this article presents many insights into the application of the technique. It describes the algorithm on which the metric solution used in this program is based.

Tubert-Bitter, P., Manfredi,R., Lellouch, J., Begaud, B. 2000. "Sample size calculations for risk equivalence testing in pharmacoepidemiology." *Journal of Clinical Epidemiology* 53, 1268-1274.

Tukey, J.W. and McLaughlin, D.H. 1963. "Less Vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization." *Sankhya, Series A* 25, 331-352.

Tukey, J.W. 1977. *Exploratory Data Analysis.* Addison-Wesley Publishing Company. Reading, Mass.

U

Upton, G.J.G. 1982."A Comparison of Alternative Tests for the 2 x 2 Comparative Trial.", *Journal of the Royal Statistical Society,* Series A,, Volume 145, pages 86-105.

Upton, G.J.G. and Fingleton, B. 1989. *Spatial Data Analysis by Example: Categorical and Directional Data. Volume 2.* John Wiley & Sons. New York.

V

Velicer, W.F. 1976. "Determining the number of components from the matrix of partial correlations." *Psychometrika*, 41, 321-327.

Velleman, Hoaglin. 1981. *ABC's of Exploratory Data Analysis.* Duxbury Press, Boston, Massachusetts.

Voit, E.O. 1992. "The S-distribution. A tool for approximation and classification of univariate, unimodal probability distributions." *Biometrical J.* 34, 855-878.

Voit, E.O. 2000. "A Maximum Likelihood Estimator for Shape Parameters of S-Distributions." *Biometrical J.* 42, 471-479.

Voit, E.O. and Schwacke, L. 1998. "Scalability properties of the S-distribution." *Biometrical J.* 40, 665-684.

Voit, E.O. and Yu, S. 1994. "The S-distribution. Approximation of discrete distributions." *Biometrical J.* 36, 205-219.

W

Walter, S.D., Eliasziw, M., and Donner, A. 1998. "Sample Size and Optimal Designs For Reliability Studies." *Statistics in Medicine*, 17, 101-110.

Welch, B.L. 1938. "The significance of the difference between two means when the population variances are unequal." *Biometrika*, 29, 350-362.

Welch, B.L. 1947. "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. 1949. "Further Note on Mrs. Aspin's Tables and on Certain Approximations to the Tabled Function," *Biometrika*, 36, 293-296.

Westfall, P. et al. 1999. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Institute Inc. Cary, North Carolina.

Westgard, J.O. 1981. "A Multi-Rule Shewhart Chart for Quality Control in Clinical Chemistry," *Clinical Chemistry*, Volume 27, No. 3, pages 493-501. (This paper is available online at the www.westgard.com).

Westlake, W.J. 1981. "Bioequivalence testing—a need to rethink," *Biometrics*, Volume 37, pages 591-593.

Whittemore, Alice. 1981. "Sample Size for Logistic Regression with Small Response Probability," *Journal of the American Statistical Association*, Volume 76, pages 27-32.

Wickens, T.D. 1989. *Multiway Contingency Tables Analysis for the Social Sciences.* Lawrence Erlbaum Associates. Hillsdale, New Jersey. A thorough book on the subject. Discusses loglinear models in depth.

Wilson, E.B.. 1927. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, Volume 22, pages 209-212. This article discusses the 'score' method that has become popular when dealing with proportions.

Winer, B.J. 1991**.** *Statistical Principles in Experimental Design (Third Edition)*. McGraw-Hill. New York, NY. A very complete analysis of variance book.

Wit, E., and McClure, J. 2004. *Statistics for Microarrays*. John Wiley & Sons Ltd, Chichester, West Sussex, England.

Wolfinger, R., Tobias, R. and Sall, J. 1994. "Computing Gaussian likelihoods and their derivatives for general linear mixed models," *SIAM Journal of Scientific Computing*, 15, no.6, pages 1294-1310.

Woolson, R.F., Bean, J.A., and Rojas, P.B. 1986. "Sample Size for Case-Control Studies Using Cochran's Statistic," *Biometrics*, Volume 42, pages 927-932.

Y

Yuen, K.K. and Dixon, W. J. 1973. “The approximate behavior and performance of the two-sample trimmed t,” *Biometrika*, Volume 60, pages 369-374.

Yuen, K.K. 1974. “The two-sample trimmed t for unequal population variances,” *Biometrika*, Volume 61, pages 165-170.

Z

Zar, Jerrold H. 1984**.** *Biostatistical Analysis (Second Edition)*. Prentice-Hall. Englewood Cliffs, New Jersey. This introductory book presents a nice blend of theory, methods, and examples for a long list of topics of interest in biostatistical work.

Zhou, X., Obuchowski, N., McClish, D. 2002**.** *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc. New York, New York. This is a great book on the designing and analyzing diagnostic tests. It is especially useful for its presentation of ROC curves.

Chapter Index

Index