User's Guide IV

Multivariate Analysis, Clustering, Meta-Analysis, Forecasting / Time Series, Operations Research, Mass Appraisal

NCSS
Statistical System

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

NCSS User's Guide IV

Direct inquiries to:

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

Warning:

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 IV

Table of Contents

Multivariate Analysis

400 Canonical Correlation
401 Correlation Matrix
402 Equality of Covariance
405 Hotelling's One-Sample T2
410 Hotelling's Two-Sample T2
415 Multivariate Analysis of Variance (MANOVA)
420 Factor Analysis
425 Principal Components Analysis
430 Correspondence Analysis
435 Multidimensional Scaling
440 Discriminant Analysis

Clustering

445 Hierarchical (Dendrograms)
446 K-Means
447 Medoid Partitioning
448 Fuzzy
449 Regression
450 Double Dendrograms

Meta-Analysis

455 Means
456 Proportions
457 Correlated Proportions
458 Hazard Ratios

Forecasting / Time Series

Exponential Smoothing

465 Horizontal
466 Trend
467 Trend & Seasonal

Time Series Analysis

468 Spectral Analysis
469 Decomposition Forecasting
470 The Box-Jenkins Method
471 ARIMA (Box-Jenkins)
472 Autocorrelations
473 Cross-Correlations
474 Automatic ARMA
475 Theoretical ARMA

Operations Research

480 Linear Programming

Mass Appraisal

485 Appraisal Ratios
486 Comparables – Sales Price
487 Hybrid Appraisal Models

References and Indices

References
Chapter Index
Index

User's Guide I
Table of Contents

Quick Start & Self Help

1 Installation and Basics
2 Creating / Loading a Database
3 Data Transformation
4 Running Descriptive Statistics
5 Running a Two-Sample T-Test
6 Running a Regression Analysis
7 Data Window
8 Procedure Window
9 Output Window
10 Filters
11 Writing Transformations
12 Importing Data
13 Value Labels
14 Database Subsets
15 Data Simulation
16 Cross Tabs on Summarized Data

Quick Start Index

Introduction

100 Installation
101 Tutorial
102 Databases
103 Spreadsheets
104 Merging Two Databases
105 Procedures
106 Output
107 Navigator and Quick Launch

Data

115 Importing Data
116 Exporting Data
117 Data Report
118 Data Screening
119 Transformations
120 If-Then Transformations
121 Filter
122 Data Simulator
123 Data Matching – Optimal and Greedy
124 Data Stratification

Tools

130 Macros
135 Probability Calculator

Graphics

Introduction

140 Introduction to Graphics

Single-Variable Charts

141 Bar Charts
142 Pie Charts
143 Histograms
144 Probability Plots

Two-Variable Charts (Discrete / Continuous)

150 Dot Plots
151 Histograms – Comparative
152 Box Plots
153 Percentile Plots
154 Violin Plots
155 Error-Bar Charts

Two-Variable Charts (Both Continuous)

160 Function Plots
161 Scatter Plots
162 Scatter Plot Matrix
163 Scatter Plot Matrix for Curve Fitting

Three-Variable Charts

170 3D Scatter Plots
171 3D Surface Plots
172 Contour Plots
173 Grid Plots

Settings Windows

180 Color Selection Window
181 Symbol Settings Window
182 Text Settings Window
183 Line Settings Window
184 Axis-Line Settings Window
185 Grid / Tick Settings Window
186 Tick Label Settings Window
187 Heat Map Settings Window

References and Indices

References
Chapter Index
Index

User’s Guide II

Table of Contents

Descriptive Statistics

200 Descriptive Statistics
201 Descriptive Tables

Means

T-Tests

205 One-Sample or Paired
206 Two-Sample
207 Two-Sample (From Means and SD's)

Analysis of Variance

210 One-Way Analysis of Variance
211 Analysis of Variance for Balanced Data
212 General Linear Models (GLM)
213 Analysis of Two-Level Designs
214 Repeated Measures Analysis of Variance

Mixed Models

220 Mixed Models

Other

230 Circular Data Analysis
235 Cross-Over Analysis Using T-Tests
240 Nondetects Analysis

Quality Control

250 Xbar R (Variables) Charts
251 Attribute Charts
252 Levey-Jennings Charts
253 Pareto Charts
254 R & R Study

Design of Experiments

260 Two-Level Designs
261 Fractional Factorial Designs
262 Balanced Incomplete Block Designs
263 Latin Square Designs
264 Response Surface Designs
265 Screening Designs
266 Taguchi Designs
267 D-Optimal Designs
268 Design Generator

References and Indices

References
Chapter Index
Index

User's Guide III

Table of Contents

Regression

Linear and Multiple Regression

300 Linear Regression and Correlation
305 Multiple Regression
306 Multiple Regression with Serial Correlation

Variable Selection

310 Variable Selection for Multivariate Regression
311 Stepwise Regression
312 All Possible Regressions

Other Regression Routines

315 Nonlinear Regression
320 Logistic Regression
325 Poisson Regression
330 Response Surface Regression
335 Ridge Regression
340 Principal Components Regression
345 Nondetects Regression

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

Curve Fitting

Curve Fitting

350 Introduction to Curve Fitting
351 Curve Fitting – General
360 Growth and Other Models
365 Piecewise Polynomial Models

Ratio of Polynomials

370 Search – One Variable
371 Search – Many Variables
375 Fit – One Variable
376 Fit – Many Variables

Other

380 Sum of Functions Models
385 User-Written Models
390 Area Under Curve

References and Indices

References
Chapter Index
Index

User’s Guide V

Table of Contents

Tabulation

500 Frequency Tables
501 Cross Tabulation

Item Analysis

505 Item Analysis
506 Item Response Analysis

Proportions

510 One Proportion
515 Two Independent Proportions
520 Two Correlated Proportions (McNemar)
525 Mantel-Haenszel Test
530 Loglinear Models

Diagnostic Tests

Binary Diagnostic Tests

535 Single Sample
536 Paired Samples
537 Two Independent Samples
538 Clustered Samples

ROC Curves

545 ROC Curves

Survival / Reliability

550 Distribution (Weibull) Fitting
551 Beta Distribution Fitting
552 Gamma Distribution Fitting
555 Kaplan-Meier Curves (Logrank Tests)
560 Cumulative Incidence
565 Cox Regression
566 Parametric Survival (Weibull) Regression
570 Life-Table Analysis
575 Probit Analysis
580 Time Calculator
585 Tolerance Intervals

References and Indices

References
Chapter Index
Index

Chapter 400

Canonical Correlation

Introduction

Canonical correlation analysis is the study of the linear relations between two sets of variables. It is the multivariate extension of correlation analysis. Although we will present a brief introduction to the subject here, you will probably need a text that covers the subject in depth such as Tabachnick (1989).

Suppose you have given a group of students two tests of ten questions each and wish to determine the overall correlation between these two tests. Canonical correlation finds a weighted average of the questions from the first test and correlates this with a weighted average of the questions from the second test. The weights are constructed to maximize the correlation between these two averages. This correlation is called the first canonical correlation coefficient.

You can create another set of weighted averages unrelated to the first and calculate their correlation. This correlation is the second canonical correlation coefficient. This process continues until the number of canonical correlations equals the number of variables in the smallest group.

Discriminant analysis, MANOVA, and multiple regression are all special cases of canonical correlation. It provides the most general multivariate framework. Because of this generality, it is probably the least used of the multivariate procedures. Researchers would rather use the specific procedure designed for their data. However, there are instances when canonical correlation techniques are useful.

Variates and Variables

Canonical correlation terminology makes an important distinction between the words variables and variates. The term *variables* is reserved for referring to the original variables being analyzed. The term *variates* is used to refer to variables that are constructed as weighted averages of the original variables. Thus a set of Y variates is constructed from the original Y variables. Likewise, a set of X variates is constructed from the original X variables.

Basic Issues

Some of the issues that must be dealt with during a canonical correlation analysis are:

1. Determining the number of canonical variate pairs to use. The number of pairs possible is equal to the smaller of the number of variables in each set.

2. The canonical variates themselves often need to be interpreted. As in factor analysis, you are dealing with mathematically constructed variates that are usually difficult to interpret. However, in this case, you must relate two constructed variates to each other.
3. The importance of each variate must be evaluated from two points of view. You have to determine the strength of the relationship between the variate and the variables from which it was created. You also need to study the strength of the relationship between the corresponding X and Y variates.
4. Do you have a large enough sample size? In social science work you will often need a minimum of ten cases per variable. In fields with more reliable data, you can get by with a little less.

Canonical Correlation Checklist

Tabachnick (1989) provides the following checklist for conducting a canonical correlation analysis. We suggest that you consider these issues and guidelines carefully.

Missing Data

You should begin by screening your data for outliers. Pay particular attention to patterns of missing values. The program ignores rows with missing values. If it appears that most of the missing values occur in one or two variables, you might want to leave these out of the analysis in order to obtain more data on the remaining variables.

Multivariate Normality and Outliers

Canonical correlation analysis does not make strong normality assumptions. However, as with all least squares procedures, outliers can cause severe problems. You should screen your data carefully for outliers using the various univariate normality tests and plots.

Linearity

Canonical correlation analysis assumes linear relations among the variables. You should study scatter plots of each pair of variables, watching carefully for curvilinear patterns and for outliers. The occurrence of curvilinear relationship will reduce the effectiveness of the analysis.

Multicollinearity and Singularity

Multicollinearity occurs when one variable is almost a weighted average of the others. Singularity occurs when this relationship is exact. Since inverse matrices are needed during the analysis, you must check for this. Try running a principal components analysis on each set of variables, separately. If you have eigenvalues at or near zero, you have multicollinearity problems. You must omit the offending variables.

Technical Details

As the name suggests, canonical correlation analysis is based on the correlations between two sets of variables which we call **Y** and **X**.

The correlation matrix of all the variables is divided into four parts:

1. R_{xx}. The correlations among the **X** variables.
2. R_{yy}. The correlations among the **Y** variables.
3. R_{xy}. The correlations between the **X** and **Y** variables.
4. R_{yx}. The correlations between the **Y** and **X** variables.

Canonical correlation analysis may be defined using the singular value decomposition of a matrix ***C*** where:

$$C = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}$$

Define the singular value decomposition of ***C*** as:

$$C = U'\Lambda\hat{B}$$

The diagonal matrix Λ of the singular values of ***C*** is made up of the eigenvalues of ***C***. The i^{th} eigenvalue λ_i of the matrix ***C*** is equal to the square of the i^{th} canonical correlation which is called r_{ci}^2. Hence, the i^{th} canonical correlation is the square root of the i^{th} eigenvalue of ***C***.

Two sets of canonical coefficients (like regression coefficients) are used for each canonical correlation: one for the **X** variables and another for the **Y** variables. These coefficients are defined as follows:

$$B_y = R_{yy}^{-1/2}\hat{B}$$

$$B_x = \Lambda R_{xx}^{-1} R_{xy} B_y$$

The canonical scores for **X** and **Y** (denoted $\hat{X}$ and $\hat{Y}$) are calculated by multiplying the standardized data (subtract the mean and divide by the standard deviation) by these coefficient matrices. Thus we have:

$$\hat{X} = Z_x B_x$$

and

$$\hat{Y} = Z_y B_y$$

where Z_x and Z_y represent the standardized versions of **X** and **Y**.

To aid in the interpretation of the canonical variates, loading matrices are computed. These are the correlations between the original variables and the constructed variates. They are computed as follows:

$$A_x = R_{xx} B_x$$

$$A_y = R_{yy} B_y$$

The *average squared loadings* are given by

$$pv_{xc} = 100\sum_{i=1}^{k_x}\frac{a_{ixc}^2}{k_x}$$

$$pv_{yc} = 100\sum_{i=1}^{k_y}\frac{a_{iyc}^2}{k_y}$$

The *redundancy indices* are given by:

$$rd = (pv)(r_c^2)$$

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable.

Missing Values

Rows with missing values in any of the variables used in the analysis are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Data Variables

Y Variables

Specify the first set of one or more variables to be correlated with the second set of variables. Although we call these the Y variables, they are not dependent variables. Canonical correlation does not assume a dependent versus independent relationship between the two sets of variables. Rather, it analyzes their association. The results would be the same if the X and Y variables were reversed.

X Variables

Specify the second set of one or more variables to be correlated with the first set of variables.

Partial Variables

An optional set of variables whose influence on the X and Y variables is removed using partial correlation techniques.

The linear influence of these variables is removed from the X and Y variables using a statistical adjustment mechanism called partial correlation. This operation involves running a multiple regression using each of the X and Y variables as the dependent variable and the partial variables as the independent variables. The residuals from each of these multiple regressions are used to calculate a *partial* correlation matrix.

Partial correlation analysis has some serious limitations. First, partial correlation techniques only remove linear (straight-line) patterns. Curvilinear patterns are ignored. Second, like all algorithms based on least squares, the results may be severely distorted by the data outliers.

Labels

Y Variate Label

This is a label that will be associated with the Y variates (constructed as weighted averages from the Y variables).

X Variate Label

This is a label that will be associated with the X variates (constructed as weighted averages from the X variables).

Options

Zero Exponent

This is the exponent of 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.

You enter the exponent only. For example, if you wanted to use 1E-16, you enter 16 here.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Descriptive Statistics ... Scores Reports

Specify whether to display the indicated reports.

Report Options

Number of Correlations

This option specifies the number of canonical correlations that are reported on. One of the major attractions to canonical correlation analysis is the reduction in variable count, so this value is usually set to two or three. You would approach the selection of this number in much the same way as selecting the number of factors in factor analysis.

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.

Select Plots

Scores Plot

Specify whether to display the scores plot.

Plot Options

Plot Size

This option controls the size of the plots that are displayed. You can select *small*, *medium*, or *large*. *Medium* and *large* are displayed one per line, while *small* are displayed two per line.

Scores Plot Tab

These options control the attributes of the scores 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

The constructed variates may be stored on the current database for further analysis. This group of options lets you designate which variates (if any) should be stored and which variables should receive these variates. The data are automatically stored while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Y Variates

Store the values of the Y Variates in these variables.

X Variates

Store the values of the X Variates in these variables.

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 – Canonical Correlation Analysis

This section presents an example of how to run a canonical correlation analysis using data contained on the SAMPLE database. As an example, we will correlate variables Test1, Test2, and Test3 with variables Test4, Test5, and IQ.

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

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Canonical Correlation**. The Canonical 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 Canonical Correlation window, select the **Variables tab**.
- Double-click in the **Y Variables** text box. This will bring up the variable selection window.
- Select **Test4, Test5, IQ** from the list of variables and then click **Ok**. “Test4-IQ” will appear in the Y Variables box.
- Double-click in the **X Variables** text box. This will bring up the variable selection window.
- Select **Test1, Test2, Test3** from the list of variables and then click **Ok**. “Test1-Test3” will appear in the X Variables box.

4 Specify the reports.

- Select the **Reports tab**.
- Enter **3** in the **Number of Correlations** box.
- Check all reports and plots. Normally you would only view a few of these reports, but we are selecting them all 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          Non-Missing
Type        Variable      Mean          Deviation         Rows
Y           Test4         65.53333      13.95332          15
Y           Test5         69.93333      16.15314          15
Y           IQ            104.3333      11.0173           15
X           Test1         67.93333      17.39239          15
X           Test2         61.4          19.39735          15
X           Test3         72.33334      14.73415          15
```

This report displays the descriptive statistics for each variable. You should check that the mean is reasonable and that the number of nonmissing rows is accurate.

Correlation Section

```
Correlation Section

                 Test4        Test5           IQ        Test1        Test2        Test3
Test4         1.000000    -0.172864     0.371404     0.753937     0.719623    -0.140941
Test5        -0.172864     1.000000    -0.058064     0.013967    -0.281449     0.347335
IQ            0.371404    -0.058064     1.000000     0.225648     0.240651     0.074070
Test1         0.753937     0.013967     0.225648     1.000000     0.100018    -0.260801
Test2         0.719623    -0.281449     0.240651     0.100018     1.000000     0.057232
Test3        -0.140941     0.347335     0.074070    -0.260801     0.057232     1.000000
```

This report presents the simple correlations among all variables specified.

Canonical Correlations Section

```
Canonical Correlations Section

Variate     Canonical                                Num     Den        Prob     Wilks'
Number    Correlation     R-Squared     F-Value       DF      DF       Level     Lambda
1            0.995600      0.991219     16.58          9      22    0.000000     0.006819
2            0.467461      0.218519     0.67           4      20    0.617695     0.776503
3            0.079810      0.006370     0.07           1      11    0.795498     0.993630
F-value tests whether this canonical correlation and those following are zero.
```

This report presents the canonical correlations plus supporting material to aid in their interpretation.

Variate Number

This is the sequence number of the canonical correlation. Remember that the first correlation will be the largest, the second will be the next to largest, and so on.

Canonical Correlation

The value of the canonical correlation coefficient. This coefficient has the same properties as any other correlation: it ranges between minus one and one, a value near zero indicates low correlation, and an absolute value near one indicates near perfect correlation.

R-Squared

The square of the canonical correlation coefficient. This gives the R-squared value of fitting the Y canonical variate to the corresponding X canonical variate.

F-Value

The value of the F approximation for testing the significance of the Wilks’ lambda corresponding to this row and those below it. In this example, the first F-Value tests the significance of the first, second, and third canonical correlations while the second F-value tests the significance of only the second and third.

Num DF

The numerator degrees of freedom of the above F-ratio.

Den DF

The denominator degrees of freedom of the above F-ratio.

Prob Level

This is the probability value for the above F statistic. A value near zero indicates a significant canonical correlation. A cutoff value of 0.05 or 0.01 is often used to determine significance.

Wilks’ Lambda

The Wilks’ lambda value for the canonical correlation on this report row. Wilks’ lambda is the multivariate generalization of R-Squared. The Wilks’ lambda statistic is interpreted just the opposite of R-Squared: a value near zero indicates high correlation while a value near one indicates low correlation.

Variance Explained Section

```
Variation Explained Section

Canonical  Variation  Explained  Individual    Cumulative    Canonical
Variate    in these   by these      Percent       Percent  Correlation
Number     Variables  Variates    Explained     Explained      Squared
1          Y          Y                37.6          37.6       0.9912
2          Y          Y                32.1          69.7       0.2185
3          Y          Y                30.3         100.0       0.0064

1          Y          X                37.2          37.2       0.9912
2          Y          X                 7.0          44.3       0.2185
3          Y          X                 0.2          44.5       0.0064

1          X          Y                37.1          37.1       0.9912
2          X          Y                 5.4          42.5       0.2185
3          X          Y                 0.2          42.8       0.0064

1          X          X                37.4          37.4       0.9912
2          X          X                24.8          62.2       0.2185
3          X          X                37.8         100.0       0.0064
```

This report displays the percent of the variation in each set of variables explained by other sets of variables.

Canonical Variate Number

This is the sequence number of the canonical variable being reported on. Remember that the maximum number of variates is the minimum of the number of variables in each set.

Variation in these Variables

Each row of the report presents the results of how well a set of variables is explained by a particular canonical variate. This column designates which set of variables is being reported on.

Explained by these Variates

Each row of the report presents the results of how well a set of variables is explained by a particular canonical variate. This column designates which set of canonical variates is being reported on.

Individual Percent Explained

This column indicates the percentage of the variation in the designated set of variables that is explained by this canonical variate.

Cumulative Percent Explained

This column indicates the cumulative percentage of the variation in the designated set of variables that is explained by this canonical variate and those listed above it.

Canonical Correlation Squared

The square of the canonical correlation coefficient. This is repeated from an earlier report.

Standardized Canonical Coefficients Section

```
Standardized Y Canonical Coefficients Section
                   Y1          Y2          Y3
Test4        1.021375    0.104989    0.370860
Test5       -0.005995    0.990267    0.224017
IQ          -0.065358    0.229775   -1.050237

Standardized X Canonical Coefficients Section
                   X1          X2          X3
Test1        0.690657    0.592485    0.510311
Test2        0.655584   -0.428196   -0.636097
Test3       -0.008941    0.919574   -0.485199
```

These coefficients are used to estimate the standardized scores for the X and Y variates. They aid the interpretation of the variates by showing the weight given each variable in the construction of the variate. They are analogous to standardized beta coefficients in multiple regression.

Variable - Variate Correlations Section

```
Variable - Variate Correlations Section
                   Y1          Y2          Y3          X1          X2
Test4        0.998137    0.019146   -0.057927    0.993745    0.008950
Test5       -0.178759    0.958777    0.220890   -0.177972    0.448190
IQ           0.314333    0.211270   -0.925505    0.312950    0.098760
Test1        0.755221    0.144834    0.045750    0.758559    0.309832
Test2        0.720964   -0.147861   -0.048910    0.724151   -0.316308
Test3       -0.150877    0.346177   -0.052251   -0.151544    0.740547

                   X3
Test4       -0.004623
Test5        0.017629
IQ          -0.073865
Test1        0.573230
Test2       -0.612826
Test3       -0.654694
```

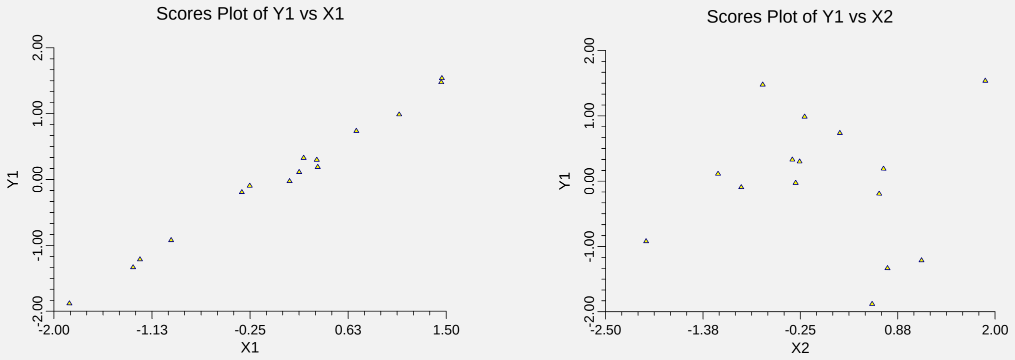
This report shows the correlations between the variables and the variates. By determining which variables are highly correlated with a particular variate, it is hoped that you can determine its interpretation. For example, you can see that variate Y1 is highly correlated with Test4. Hence, we assume that Y1 has the same interpretation as Test4.

Scores Section

```
Scores Section

Row          Y1          Y2          Y3          X1          X2          X3
1     -0.193124   -0.348044   -0.308495   -0.323303    0.660431    1.582089
2     -1.214743    0.350598    0.877022   -1.232224    1.150186    1.517131
3     -0.026336    0.135325    0.250782    0.103271   -0.304012   -1.369888
4      1.536744    1.992049   -0.657871    1.461462    1.887123   -0.138798
5      0.189923    0.709643    0.455333    0.354314    0.711949    0.757851
6      0.986597   -0.677646    0.115011    1.081350   -0.201044    0.489839
.             .           .           .           .           .           .
.             .           .           .           .           .           .
.             .           .           .           .           .           .
```

This report provides the canonical scores of each set of variates for each row of non-missing data. These are the values that are plotted in the score plots shown next.

Scores Plots



(seven more plots are displayed)

These reports show the relationship between each pair of canonical variates. The correlation coefficient of the data in the first plot (Y1 versus X1) is the first canonical correlation coefficient.

Chapter 401

Correlation Matrix

Introduction

This program calculates Pearsonian and Spearman-rank correlation matrices. It allows missing values to be deleted in a pair-wise or row-wise fashion.

When someone speaks of a correlation matrix, they usually mean a matrix of Pearson-type correlations. Unfortunately, these correlations are unduly influenced by outliers, unequal variances, nonnormality, and nonlinearities. One of the chief competitors of the Pearson correlation coefficient is the Spearman-rank correlation coefficient. This latter correlation is calculated by applying the Pearson correlation formulas to the ranks of the data rather than to the actual data values themselves. In so doing, many of the distortions that infect the Pearson correlation are reduced considerably.

To allow you to compare the two types of correlation matrices, a matrix of differences can be displayed. This allows you to determine which pairs of variables require further investigation.

This program lets you specify a set of partial variables. The linear influence of these variables is removed by sweeping them out of the matrix. This provides a statistical adjustment to the remaining variables using multiple regression. Note that in the case of Spearman correlations, this sweeping occurs after the complete correlation matrix has been formed.

Discussion

When there is more than one independent variable, the collection of all pair-wise correlations are succinctly represented in a correlation form. In regression analysis, the purpose of examining these correlations is two-fold: to find outliers and to identify collinearity. In the case of outliers, there should be major differences between the parametric measure, the Pearson correlation coefficient, and the nonparametric measure, the Spearman rank correlation coefficient. In the case of collinearity, high pair-wise correlations could be the first indicators of collinearity problems.

The Pearson correlation coefficient is unduly influenced by outliers, unequal variances, nonnormality, and nonlinearities. As a result of these problems, the Spearman correlation coefficient, which is based on the ranks of the data rather than the actual data, may be a better choice for examining the relationships between variables.

Finally, the patterns of missingness in multiple regression and correlation analysis can be very complex. As a result, missing values can be deleted in a pair-wise or a row-wise fashion. If there are only a few observations with missing values, it might be preferable to use the row-wise deletion, especially for large data sets. The row-wise deletion procedure omits the entire observation from the analysis. On the other hand, if the pattern of missingness is randomly dispersed throughout the data and the use of the row-wise deletion would omit at least 25% of the observations, the pair-wise deletion procedure for missing values would be a safer way to capture the essence of the relationships between variables. While this method appears to make full use of all your data, the resulting correlation matrix may have mathematical and interpretation difficulties. Mathematically, this correlation matrix may not have a positive determinant. Since each correlation may be based on a different set of rows, practical interpretations could be difficult, if not illogical.

The Spearman correlation coefficient measures the monotonic association between two variables in terms of ranks. It measures whether one variable increases or decreases with another even when the relationship between the two variables is not linear or bivariate normal. Computationally, each of the two variables is ranked separately, and the ordinary Pearson correlation coefficient is computed on the ranks. This nonparametric correlation coefficient is a good measure of the association between two variables when outliers, nonnormality, nonconstant variance, and nonlinearity may exist between the two variables being investigated.

Data Structure

The data are entered in two or more variables. An example of data appropriate for this procedure is shown in the table below. These 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 (subset)

YldA	YldB	YldC
452	546	785
874	547	458
554	774	886
447	465	536
356	459	
754	665	669
558	467	857
574	365	821
664	589	772
682	534	732
	456	689
547	651	654
	654	
435	665	297
	546	830
245	537	827

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Data Variables

Correlation Variables

Specify the variables whose correlations are to be formed. Only numeric data are analyzed.

Partial Variables

An optional set of variables that are to be "partialled out" of the correlation matrix. The influence of these variables is removed by sweeping them from the remaining variables. For the Pearson-type correlations, the resulting matrix is the same that would be formed if the regular variables were regressed on the partial variables, the residuals were stored, and the correlation matrix of these residuals was formed. The correlations that are formed are the partial correlations.

Options

Correlation Type

Specify the type of correlation to be computed

- **Pearson Product-Moment**

  Display the Pearson product-moment correlation matrix.

- **Spearman Rank**

  Display the Spearman-Rank correlation matrix.

- **Both**

  Display both the Pearson product-moment and the Spearman-Rank correlation matrices.

Missing Value Removal

This option indicates how you want the program to handle missing values.

- **Pair-wise**

  Pair-wise removal of missing values. Each correlation is based on all pairs of data values in which no missing values occur. Missing values occurring in other variables do not influence the calculations. Note that although this method appears to make full use of all your data, the resulting correlation matrix is difficult to analyze. Mathematically, it may not have a positive determinant. Practically, each correlation may be based on a different set of rows, making it difficult to interpret.

- **Row-wise**

  Row-wise removal of missing values. If a missing value occurs in any of the variables specified, the row of data is ignored in the calculation of all correlations.

Reports Tab

These options specify the reports.

Select Reports

Difference Report

Specify whether to display the matrix of differences between the Pearson and the Spearman correlations.

Report Options

Report Format

Specifies the length and format of the correlation matrix.

- **Short**
  Display only the correlation matrix.

- **Full**
  Display the full report with sample sizes and significance levels.

Variable Names

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

Storage Tab

Specify if and where the correlation matrices are to be stored.

Data Storage Variables

Pearson Correlations Storage Variables

A list of variables into which the Pearson correlation matrix is stored.

Spearman Correlations Storage Variables

A list of variables into which the Spearman correlation matrix is stored.

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 – Creating a Correlation Matrix

This section presents an example of how to run an 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 Correlation Matrix 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 Correlation Matrix window.**

- On the menus, select **Analysis**, then **Regression/Correlation**, then **Correlation Routines**, then **Correlation Matrix**. The Correlation Matrix 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 Correlation Matrix window, select the **Variables tab**.
- Double-click in the **Correlation Variables** text box. This will bring up the variable selection window.
- Select **YldA, YldB, YldC** from the list of variables and then click **Ok**. "YldA-YldB,YldC" will appear in the Correlation Variables box.
- Enter **Both** in the Correlation Type box.
- Enter **Pair Wise** in the Missing Value Removal box.

4 **Specify the reports.**

- On the Correlation Matrix window, select the **Reports tab**.
- Check the **Different Report** box.
- Enter **Full** in the Report Format box.

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).

Pearson Correlations Section

Full Report Format

```
Pearson Correlations Section   (Pair-Wise Deletion)

                  YldA            YldB              YldC
YldA              1.000000        .170692           -.361414
                  .000000         .577154           .248377
                  12.000000       13.000000         12.000000
YldB              .170692         1.000000          -.004071
                  .577154         .000000           .988980
                  13.000000       14.000000         14.000000
YldC              -.361414        -.004071          1.000000
                  .248377         .988980           .000000
                  12.000000       14.000000         14.000000
Cronbachs Alpha = -0.250337    Standardized Cronbachs Alpha = -0.223864
```

The above report displays the correlations, significance level, and sample size of each pair of variables. This format is obtained when Report Format is set to “Full.”

Reliability

Because of the central role of measurement in science, scientists of all disciplines are concerned with the accuracy of their measurements. Item analysis is a methodology for assessing the accuracy of measurements that are obtained in the social sciences where precise measurements are often hard to secure. The accuracy of a measurement may be broken down into two main categories: validity and reliability. The validity of an instrument refers to whether it accurately measures the attribute of interest. The reliability of an instrument concerns whether it produces identical results in repeated applications. An instrument may be reliable but not valid. However, it cannot be valid without being reliable.

The methods described here assess the reliability of an instrument. They do not assess its validity. This should be kept in mind when using the techniques of item analysis since they address reliability, not validity.

An instrument may be valid for one attribute but not for another. For example, a driver’s license exam may accurately measure an individual’s ability to drive. However, it does not accurately measure that individual’s ability to do well in college. Hence the exam is reliable and valid for measuring driving ability. It is reliable and invalid for measuring success in college.

Several methods have been proposed for assessing the reliability of an instrument. These include the retest method, alternative-form method, split-halves method, and the internal consistency method. We will focus on internal consistency here.

Cronbach’s Alpha

Cronbach’s alpha (or *coefficient alpha*) is the most popular of the internal consistency coefficients. It is calculated as follows:

$$\alpha = \frac{K}{K-1}\left[1-\frac{\sum_{i=1}^{K}\sigma_{ii}}{\sum_{i=1}^{K}\sum_{j=1}^{K}\sigma_{ij}}\right]$$

where K is the number of items (questions) and σ_{ij} is the estimated covariance between items i and j. Note the σ_{ii} is the variance (not standard deviation) of item i.

If the data are standardized by subtracting the item means and dividing by the item standard deviations before the above formula is used, we get the standardized version of Cronbach's alpha. A little algebra will show that this is equivalent to the following calculations based directly on the correlation matrix of the items:

$$\alpha = \frac{K\bar{\rho}}{1+\bar{\rho}(K-1)}$$

where K is the number of items (variables) and $\bar{\rho}$ is the average of all the correlations among the K items.

Cronbach's alpha has several interpretations. It is equal to the average value of alpha coefficients obtained for all possible combinations of dividing *2K* items into two groups of K items each and calculating the two-half tests. Also, alpha estimates the expected correlation of one instrument with an alternative form containing the same number of items. Furthermore, alpha estimates the expected correlation between an actual test and a hypothetical test which may never by written.

Since Cronbach's alpha is suppose to be a correlation, it should range between -1 and 1. However, it is possible for alpha to be less than -1 when several of the covariances are relatively large, negative numbers. In most cases, alpha is positive, although negative values arise occasionally. What value of alpha should be achieved? Carmines (1990) stipulates that as a rule, a value of at least 0.8 should be achieved for widely used instruments. An instrument's alpha value may be improved by either adding more items or by increasing the average correlation among the items.

Short Report Format

```
Pearson Correlations Section   (Pair-Wise Deletion)

              YldA            YldB              YldC
YldA          1.000000        .170692           -.361414
YldB          .170692         1.000000          -.004071
YldC          -.361414        -.004071          1.000000
Cronbachs Alpha = 0.219908    Standardized Cronbachs Alpha = 0.311396
```

The above report displays the correlation matrix only. This format is obtained when Report Format is set to "Short."

Spearman Correlations Section

Full Report Format

```
Spearman Correlations Section   (Pair-Wise Deletion)

               YldA         YldB          YldC
YldA           1.000000     .184319       -.153846
               .000000      .546634       .633091
               12.000000    13.000000     12.000000
YldB           .184319      1.000000      -.088106
               .546634      .000000       .764552
               13.000000    14.000000     14.000000
YldC           -.153846     -.088106      1.000000
               .633091      .764552       .000000
               12.000000    14.000000     14.000000
```

The above report displays the correlations, significance level, and sample size of each pair of variables. This format is obtained when Report Format is set to “Full.”

Short Report Format

```
Spearman Correlations Section   (Pair-Wise Deletion)

               YldA         YldB          YldC
YldA           1.000000     .184319       -.153846
YldB           .184319      1.000000      -.088106
YldC           -.153846     -.088106      1.000000
```

The above report displays the correlation matrix only. This format is obtained when Report Format is set to “Short.”

Difference Between Correlations Section

```
Difference Between Pearson and Spearman Correlations Section   (Pair-Wise Deletion)

               YldA         YldB          YldC
YldA           .000000      -.013627      -.207568
YldB           -.013627     .000000       .084035
YldC           -.207568     .084035       .000000
```

The above report displays the difference between the Pearson and the Spearman correlation coefficients. The report lets you find those pairs of variables for which these two correlation coefficients are very different. A large difference here indicates the presence of outliers, nonlinearity, nonnormality, and the like. You should investigate scatter plots of pairs of variables with large differences.

Storing the Correlations on the Database

When you specify variables in either the Pearson Correlations or the Spearman Correlations boxes, the correlation matrix will be stored in those variables during the execution of the program.

Chapter 402

Equality of Covariance

Introduction

Discriminant analysis, MANOVA, and other multivariate procedures assume that the individual group covariance matrices are equal (homogeneous across groups). This ***NCSS*** module lets you test this hypothesis using Box's M test, which was first presented by Box (1949). This module also performs Bartlett's univariate homogeneity of variance test for testing equality of variance among individual variables.

Box's M Test

The calculation of Box's M test proceeds as follows. Suppose you have *k* groups measured on each of *p* variables, with n_i observations per group. Represent the estimated within-group covariance as S_i (the divisor is $n_i - 1$). The calculations for Box's M and Bartlett's test are identical. Box's M is simply an extension of Bartlett's test to the multivariate case. To calculate Bartlett's test, set $p = 1$. The value of *M* is given by

$$M = (N - k) \log_e |S| - \sum_{i=1}^{k} (n_i - 1) \log_e |S_i|$$

where

$$N = \sum_{i=1}^{k} n_i$$

$$S = \frac{\sum_{i=1}^{k} (n_i - 1) S_i}{N - k}$$

We use the Chi-square and F-ratio to test the significance of the *M* value. These approximations are constructed as follows:

$$A_1 = \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left[\sum_{i=1}^{k} \left(\frac{1}{n_i - 1} \right) - \frac{1}{N - k} \right]$$

$$v_1 = \frac{p(p+1)(k-1)}{2}$$

$$A_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)^2 - \frac{1}{(N-k)^2}\right]$$

If $A_2 - A_1^2 > 0$ then

$$v_2 = \frac{v_1 + 2}{A_2 - A_1^2}$$

$$b = \frac{v_1}{1 - A_1 - (v_1 / v_2)}$$

$$F_{v_1,v_2} = \frac{M}{b}$$

If $A_2 - A_1^2 < 0$ then

$$v_2 = \frac{v_1 + 2}{A_1^2 - A_2}$$

$$b = \frac{v_2}{1 - A_1 + (2 / v_2)}$$

$$F_{v_1,v_2} = \frac{v_2 M}{v_1(b - M)}$$

$$\chi^2_{v_1} = M\,(1 - A_1)$$

Box’s M test is very sensitive to non-normality, so that a significant value indicates either unequal covariance matrices or non-normality or both. Hence, it is important to establish multivariate normality before using Box’s M test.

The Chi-square approximation should be used when all $n_i > 20$, $p < 6$, and $k < 6$. Otherwise, the F approximation is more accurate.

NCSS supplies both the multivariate Box’s M test and the individual Bartlett’s tests so that when Box’s M test is significant, you can determine which variables contribute to the variance inequality.

Data Structure

The data given in the table below are the first eight rows (out of the 150 in the database) of the famous “iris data” published by Fisher (1936). These data are measurements in millimeters of sepal length, sepal width, petal length, and petal width of fifty plants for each of three varieties of iris: (1) Iris setosa, (2) Iris versicolor, and (3) Iris virginica.

We will test to see if the covariance matrices are equal across the three varieties of iris. Here Iris is the group variable while SepalLength, SepalWidth, PetalLength, and PetalWidth are the regular variables.

FISHER dataset (subset)

SepalLength	SepalWidth	PetalLength	PetalWidth	Iris
50	33	14	2	1
64	28	56	22	3
65	28	46	15	2
67	31	56	24	3
63	28	51	15	3
46	34	14	3	1
69	31	51	23	3
62	22	45	15	2

Missing Values

If missing values are found in any of the variables being used, the row is omitted.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Group Variable

Y: Group Variable

This is the dependent, Y, grouping, or classification variable. It must be discrete--it can only have a few unique values. Each unique value represents a separate group of individuals. The values may be text or numeric.

Independent Variables

X's: Independent Variables

Specify the independent (X or Predictor) variables. These should be numeric variables whose values are either continuous or binary.

Reports Tab

The following options control the format of the reports.

Select Reports

Group Means - Box's M Test

These options let you specify which reports you want 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. 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.

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 – Equality of Covariance Analysis

This section presents an example of how to run an analysis. The data used are shown in the table above and found in the FISHER database. In this example, we will test whether the covariance matrices of the four measurements (SepalLength, SepalWidth, PetalLength, and PetalWidth) are equal across the three iris varieties.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Equality of Covariance window.

1 Open the Fisher dataset.

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

2 Open the Equality of Covariance window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Equality of Covariance**. The Equality of Covariance 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 Equality of Covariance window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Group Variable box.
- Double-click in the **X’s: Independent Variables** box. This will bring up the variable selection window.
- Select **Sepal Length** through **PetalWidth** from the list of variables and then click **Ok**. “SepalLength-PetalWidth” will appear in the Variables box.

4 Specify which reports.

- Select the **Reports** tab.
- **Check all reports**. Normally you would only view a few of these reports, but we are selecting them all so that we can document them.
- Enter **Labels** in the **Variable Names** box.
- Enter **Value Labels** in the **Value Labels** box.

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).

Group Means Report

Group Means

	Iris			
Variable	**Setosa**	**Versicolor**	**Virginica**	**Overall**
Sepal Length	50.06	59.36	65.88	58.43333
Sepal Width	34.28	27.7	29.74	30.57333
Petal Length	14.62	42.6	55.52	37.58
Petal Width	2.46	13.26	20.26	11.99333
Count	50	50	50	150

This report shows the means of each of the variables across each of the groups. The last row shows the count (number of observations) in the group. Note that the column headings come from the use of value labels for the group variable.

Group Standard Deviations Report

Group Standard Deviations

	Iris			
Variable	**Setosa**	**Versicolor**	**Virginica**	**Overall**
Sepal Length	3.524897	5.161712	6.358796	8.280662
Sepal Width	3.790644	3.137983	3.224966	4.358663
Petal Length	1.73664	4.69911	5.518947	17.65298
Petal Width	1.053856	1.977527	2.7465	7.622377
Count	50	50	50	150

This report shows the standard deviations of each of the variables across each of the groups. The last row shows the count or number of observations in the group.

Within Group Correlation\Covariance Matrices

Within-Group Correlation\Covariance For Type of Iris = Total

	Variable			
Variable	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Sepal Length	26.50082	9.272109	16.75143	3.840136
Sepal Width	.530236	11.53878	5.524354	3.27102
Petal Length	.756164	.377916	18.51878	4.266531
Petal Width	.364506	.470535	.484459	4.188163

Within-Group Correlation\Covariance For Type of Iris = Setosa

Variable	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	12.4249	9.921633	1.63551	1.033061
Sepal Width	.742547	14.36898	1.169796	.9297959
Petal Length	.267176	.177700	3.015918	.6069388
Petal Width	.278098	.232752	.331630	1.110612

Within-Group Correlation\Covariance For Type of Iris = Versicolor

Variable	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	26.64326	8.518368	18.2898	5.577959
Sepal Width	.525911	9.846939	8.265306	4.120408
Petal Length	.754049	.560522	22.08163	7.310204
Petal Width	.546461	.663999	.786668	3.910612

Within-Group Correlation\Covariance For Type of Iris = Virginica

Variable	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	40.43428	9.376327	30.32898	4.909388
Sepal Width	.457228	10.40041	7.137959	4.762857
Petal Length	.864225	.401045	30.45877	4.882449
Petal Width	.281108	.537728	.322108	7.543265

This report shows the within-group correlations in the lower-left portion of the matrix and the within-group covariances in the upper-right portion of the matrix. The within-group variances are displayed on the diagonal. The total within-group values are found by forming a weighted average of the group covariances, averaging across all groups.

The three individual-group reports show the correlations and covariances for each of the three iris varieties. These are the correlations and covariances that would be obtained if each group was analyzed separately. These are the group covariances that will be tested by Box's M test.

Bartlett-Box Homogeneity Tests

Bartlett-Box Homogeneity Tests

	Bartlett			F	F	Chi2	Chi2
Variable	Value	DF1	DF2	Approx	Prob	Approx	Prob
Sepal Length	16.1509	2	48620	8.01	.000334	16.00	.000335
Sepal Width	2.1100	2	48620	1.05	.351514	2.09	.351533
Petal Length	55.9252	2	48620	27.74	.000000	55.42	.000000
Petal Width	39.5688	2	48620	19.62	.000000	39.21	.000000
Box's M	146.6632	20	77567	7.05	.000000	140.94	.000000

This report gives Bartlett's test for each variable followed by the Box's M test for all variables together. These tests are used to determine whether the variances of each of the groups are close enough to each other so that they may be considered equal. For example, the first line of the report tests for equal group variances of sepal length (SepalLength). Since the probability levels are small (less than 0.01), we would assume that the variances are significantly different. As was mentioned earlier, this test is also sensitive to departures from normality, so a significant result should be interpreted to mean that the variances are different or the data is non-normal. You can run a normality test to check this assumption.

Notice that the probability levels of SepalWidth are 0.35151 and 0.35153. Hence, both tests indicate that the variances are essentially equal. This is the only variable that did not fail this test!

We should also make a point regarding sample size here. The size of the probability level is directly related to the size of the sample. This probability level is for statistical significance, which may or may not be related to practical significance. You will have to consider this by comparing the individual standard deviations from one of the prior reports.

Matrix Determinant Report

```
Matrix Determinant Section

                            Log of Covariance              Correlation
Type of Iris                      Determinant              Determinant
Setosa                               5.353320                  .353359
Versicolor                           7.546356                  .083594
Virginica                            9.493622                  .137390
Pooled (Overall)                     8.462142                  .199529
```

This report gives the logarithm (base e) of the determinant of each of the relevant covariance matrices and the determinant of each of the correlation matrices. This report is useful since Box's M test compares these values.

Eigenvalues of Covariance Matrices Report

```
Eigenvalues of Covariance Matrices

                  Iris
Number            Setosa          Versicolor      Virginica       Overall
1                 23.645569       48.787394       69.525484       44.356592
2                 3.691873        7.238410        10.655123       8.618331
3                 2.679640        5.477609        5.229543        5.535235
4                 .903326         .979036         3.426585        2.236372
```

This report gives the eigenvalues of each of the individual covariance matrices followed by the eigenvalues of the within-group covariance matrix. Each column gives a set of eigenvalues.

This report is useful because the eigenvalues summarize the covariance matrix into a few values. By comparing the largest eigenvalues across all groups, you can determine which groups are different. Also, eigenvalues near zero indicate singularities in your data.

Eigenvalues of Correlation Matrices Report

```
Eigenvalues of Covariance Matrices

                  Iris
Number            Setosa          Versicolor      Virginica       Overall
1                 2.05854         2.926341        2.454737        2.503762
2                 1.022178        .5462747        .9647126        .7251373
3                 .6678202        .3949976        .4522719        .5824012
4                 .2514613        .1323871        .1282783        .1886997
```

This report gives the eigenvalues of each of the individual correlation matrices followed by the eigenvalues of the within-group correlation matrix. Each column gives a set of eigenvalues.

This report is useful because the eigenvalues summarize the correlation matrix into a few values. By comparing the largest eigenvalues across all groups, you can determine which groups are different. Also, eigenvalues near zero indicate singularities in your data.

Chapter 405

Hotelling's One-Sample T2

Introduction

The one-sample Hotelling's *T*2 is the multivariate extension of the common one-sample or paired Student's *t*-test. In a one-sample *t*-test, the mean response is compared against a specific value. Hotelling's one-sample *T*2 is used when the number of response variables is two or more, although it can be used when there is only one response variable.

*T*2 makes the usual assumption that the data are approximately multivariate normal. *Randomization tests* are provided that do not rely on this assumption. These randomization tests should be used whenever you want exact results that do not rely on several assumptions.

One-Sample Case

The one-sample *T*2 is used to test hypotheses about a set of means simultaneously. Specifically, suppose a set of *p* response variables $Y_1, Y_2, \cdots, Y_p$ is measured. Assume that the population is distributed as $N_p(\mu, \Sigma)$, where $N_p(\mu, \Sigma)$ is the *p*-variable multivariate normal distribution with mean vector μ and covariance matrix Σ. The null hypothesis that $\mu = \mu_0$, where μ_0 is a vector of *p* constants (often 0's), can be tested using the test statistic

$$T2 = n(\bar{y} - \mu_0)' S^{-1} (\bar{y} - \mu_0)$$

where $\bar{y}$ is the sample mean vector, *n* is the sample size, and S^{-1} is the inverse of the sample covariance matrix.

If the null hypothesis that $\mu = \mu_0$ is true, then *T*2 follows Hotelling's *T*2 distribution. That is, $T2 \sim T^2_{p,n-1}$. Reject the null hypothesis if $T2 \geq T^2_{1-\alpha,p,n-1}$. Note that rejecting the null hypothesis concludes that at least one of the *p* means is not equal to its hypothesized value.

Equality of Means

A second null hypothesis may be of interest. This hypothesis is that all means are equal to each other. This hypothesis also tested using the one-sample *T*2 value calculated using the formula

$$T2' = n(C\bar{y})' (CSC')^{-1} (C\bar{y})$$

where C is a contrast matrix of the form

$$C = \begin{matrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{matrix}$$

Our choice of C tests the hypothesis that all p means are equal. In this case, $T2' \sim T^2_{p-1,n-1}$.

Paired-Sample Case

The one-sample *T*2 test may also be applied to the situation in which two samples are to be compared that had a natural pairing between two observation vectors. An example of this pairing occurs when responses are measured on each experimental subject before and after a treatment is administered. Thus, the one-sample *T*2 test may be applied in the one-factor repeated measures design.

When such pairing exists, the differences between the first and second measurements are formed—replacing the two observation vectors with one difference vector. This difference vector may then be used in the one-sample *T*2 test as described above.

Randomization Test

Because of the strict assumptions that must be made when using this procedure, ***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 sample data, calculating the test statistic for each permutation, and counting the number of permutations that result in a *T*2 value greater than or equal to the actual *T*2 value. Dividing this count by the number of permutations tried gives the significance level of the test. Each permutation is found by randomly multiplying each observation by a plus or a minus.

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.

Permutation results are provided for the equal covariance case, the unequal covariance case, and for all individual *t* tests.

Assumptions

The following assumptions are made when using *T*2.

1. The population follows the multivariate normal distribution.
2. The members of the sample are independent.

Data Structure

The data must be entered in a format that places the response variables side by side. An example of the data structure for a paired Hotelling's *T*2 design is shown below. In this example, measurements were taken at three points in time before and after a certain drug was administered. Each subject performed strenuous exercise between the first and second measurements of the before set and the after set. This database is stored in the file T2.

T2 dataset

Before1	Before2	Before3	After1	After2	After3
36	34	30	38	35	29
36	36	28	38	37	27
41	32	29	43	31	25
11	10	8	14	11	10
17	15	13	19	14	12
21	20	18	24	25	17
36	33	30	40	34	28
36	35	34	41	36	30
37	33	28	36	37	29
31	28	25	31	25	26

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options control which variables are used in the analysis.

Response Variables

Response Variables

One or more numeric response variables are specified here. These variables can contain the values to be analyzed or the first values (the X1's) when paired differences (X1-X2) are to be analyzed.

If the response variables have only two or three unique values, use the randomization test results.

Paired Variables

Paired Variables

Specify matching paired variables (the X2's) that are to be used with the Response Variables (the X1's) to create differences (X1-X2). The first variable specified here is subtracted from the first Response variable, the second variable specified here is subtracted from the second Response Variable, and so on. Hence, if these variables are used, their number must equal the number of Response Variables specified.

Leave this option blank if only the Response Variables are to be used.

Resampling

Run Randomization Tests

Check this option to run randomization tests. Note that these tests are computer-intensive and may require a great deal of time to run.

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 these tests.

Somewhere between 1,000 and 100,000 Monte Carlo samples are usually necessary. We suggest the use of 10,000.

Reports Tab

Select Reports

Means and Std. Deviations ... Correlation\Covariance

Specify whether to display the various reports.

Report Options

Confidence Coefficient

Specify the value of confidence coefficient for the confidence intervals. This is the value of one minus alpha. The value 0.95 is commonly used. However, you can specify any value between 0.50 and 0.99999.

Variable Names

Indicate whether to display the variable names or the variable labels.

Precision

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

Report Options – Decimal Places

T2 ... Correlation 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.

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 – Paired T2 Test

This section presents an example of how to run a paired T2 analysis of the T2 dataset shown earlier. In this analysis, the before and after variables will be compared.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Hotelling’s One-Sample T2 window.

1 Open the T2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **T2.s0**.
- Click **Open**.

2 Open the Hotelling’s One-Sample T2 window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Hotelling’s One-Sample T2**. The Hotelling’s One-Sample T2 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 Hotelling’s One-Sample T2 window, select the **Variables tab**.
- Set the **Response Variables** to **AFTER1-AFTER3**.
- Set the **Paired Variables** to **BEFORE1-BEFORE3**.
- Check **Run Randomization Tests** so that these tests will be included in the reports.

4 Specify the reports.

- On the Hotelling’s One-Sample T2 window, select the **Reports tab**.
- Check all of the reports. (Although all reports are not necessary, we will check them all so that they can all be documented.)
- Set **VC Decimals** to **4**.
- Set **Means Decimals** to **4**.

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

```
Variable             Mean Diff.       S.D. of Diff.
After1-Before1       2.2000           1.7512
After2-Before2       0.9000           2.3310
After3-Before3       -1.0000          2.0000
Count                10               10
```

This report provides the means, standard deviations, and counts of each difference (or variable). Look the values over to be certain that the right variables were selected.

Hotelling’s T2 Test Section

```
                                                  Parametric     Randomization
                                                        Test              Test
                                                        Prob              Prob
Hypothesis              T2      DF1         DF2        Level             Level
Means All Zero      17.034        3         9.0       0.0483            0.0370
Means All Equal      9.321        2         9.0       0.0582            0.0400
The randomization test results are based on 1000 Monte Carlo samples.
```

This report gives the results of the two *T*2 tests.

Hypothesis

This option indicates the null hypothesis that is tested on this output line. For the case of the paired observations, the first line (Means All Zero) is of most interest.

T2

The values of *T*2 are given here.

DF1

This is the number of response variables.

DF2

This is the degrees of freedom of the covariance matrix which is *n* - 1.

Parametric Test Prob Level

This is the significance level of the *T*2 test. If this value is less than 0.05, we say that the test was significant at the 0.05 level and at least one pair of means are significantly different. If the value is less than 0.01, we say that the test was significant at the 0.01 level. This result is accurate if the assumptions are met.

Randomization Test Prob Level

This is the significance level of the randomization test. The result is accurate even if the response variables were binary.

Individual Variables Section

```
Variable                 T2      Prob          T2      Prob          T2      Prob
Omitted              Others     Level      Change     Level       Alone     Level
After1-Before1        4.399    0.2036      12.635    0.0424      15.783    0.0032
After2-Before2       16.484    0.0156       0.550    0.7110       1.491    0.2531
After3-Before3       16.554    0.0154       0.480    0.7294       2.500    0.1483
```

This report provides information about the influence of each of the individual response variables on the overall *T*2 value. This is accomplished by calculating the change in *T*2 when a response variable is omitted.

Variable Omitted

This is the name of the variable shown on this line of the report.

T2 Others

This is the value of *T*2 calculated with all response variables except the variable listed its the left.

Prob Level

This is the significance level of the *T*2 value shown on the report to the left of this value.

T2 Change

This is the amount that *T*2 is reduced when the response variable shown on this line is omitted.

Prob Level

This is the significance level of the *T*2 change value shown on the report to the left of this value. It is computed using the fact that the change in *T*2 is related to an *F* distribution using the formula

$$F_{\alpha,1,n-p} = \frac{T_p^2 - T_{p-1}^2}{n-1+T_{p-1}^2}$$

Note that this quantity tests the drop in *T*2 when a variable is removed conditional on the other response variables that are included. Another way of looking at this quantity is that it tests whether the variable omitted significantly increases the distance between the two populations. See Rencher (1998) page 68 for further details.

T2 Alone

This is the value of *T*2 calculated when only this response variable is used. It is the square of the common one-sample *t* test. It is the two-sided test of the null hypothesis that the mean for this variable is equal to the hypothesized value (which is usually zero), ignoring all other variables.

Prob Level

This is the two-sided significance level of the *T*2 value to its left.

Student’s T-Test Section

```
                                       Parametric     Randomization
                                T2           Test              Test
                                or           Prob              Prob
Variable             |Student's T|          Level             Level
All (T2)                    17.034         0.0483            0.0370
After1-Before1               3.973         0.0032            0.0020
After2-Before2               1.221         0.2531            0.2490
After3-Before3               1.581         0.1483            0.1980

The randomization test results are based on 1000 Monte Carlo samples.
These individual t-test significance levels should only be used when the overall T2 value is significant.
```

This report provides the results of individually conducting a two-sided, paired *t*-test on each pair of response variables. You might think that since there are a series of *p* *t*-tests being employed, a Bonferroni adjustment should be applied to the significance levels. However, if these individual tests are only considered when the overall *T*2 is significant at the same level, such as 0.05, then their significance levels are “protected” by the *T*2 test and the unadjusted significance levels given here can be used.

Variable

The variable whose results are presented on this line. Note that the first line gives the overall results for *T*2.

T2 or |Student’s T|

The first line is the value of *T*2. The other lines are the two-sided Student’s *t*-test values.

Parametric Test Prob Level

These are the significance levels of the test statistics given to the left. Note that if the individual tests are only used when the overall test is significant, these significance levels are accurate even though several individual tests are made. The multivariate *T*2 test is said to “protect” the significance levels of the individual tests.

Randomization Test Prob Level

These are the results of randomization tests that are run on each of the variables. These tests are exact when the Monte Carlo sample size is large, say over 5000. These tests should be used when there is a even a hint that the regular assumptions of the *t*-tests are not valid. For example, this significance level is accurate even when the response variable takes on binary values (the *t*-test assumes a continuous, normal response variable).

Note that these values will change from run to run. As you increase the number of Monte Carlo samples, these values will become more and more stable. You may have to go as large as 100,000 before the results remain the same from run to run. This instability is due to the our use of a random sample of all the trillions of permutations that are possible. As you increase the Monte Carlo sample size, you reduce the sampling error (and greatly increase the time it takes to generate the results).

Confidence Intervals for the Mean Differences

Variable	Difference	Lower 95.0% Bonferroni Conf. Limit	Upper 95.0% Bonferroni Conf. Limit	Lower 95.0% Simultaneous Conf. Limit	Upper 95.0% Simultaneous Conf. Limit
After1-Before1	2.2000	0.5756	3.8244	-0.0675	4.4675
After2-Before2	0.9000	-1.2622	3.0622	-2.1182	3.9182
After3-Before3	-1.0000	-2.8552	0.8552	-3.5897	1.5897

This report provides confidence intervals for the mean differences (or the means) for each response variable. Two intervals are provided: Bonferroni and simultaneous.

Variable

The variable(s) whose results are presented on this line.

Difference

The actual difference for the corresponding response variable(s).

Bonferroni Confidence Interval

Bonferroni confidence intervals are based on the formula

$$\bar{d}_j \pm t_{\alpha/(2p),n-1}\sqrt{\left(\frac{s_{jj}}{n}\right)}$$

This formula is derived by applying a Bonferroni adjustment to the regular univariate confidence interval. This adjustment is made by dividing the alpha level by *p*, the number of such intervals to be created. These intervals are usually not as wide as the simultaneous intervals, yet still have an appropriate adjustment because of the multiple intervals that are being created.

Simultaneous Confidence Interval

Simultaneous confidence intervals are based on the formula

$$\bar{d}_j \pm \sqrt{T^2_{1-\alpha,p,n-1}\left(\frac{s_{jj}}{n}\right)}$$

This formula is derived from a formula for confidence intervals for *any* linear combination of the mean differences, including those that are generated after looking at the data. Because of this, these confidence intervals are extra wide and may not be of must use.

Correlation\Covariance Matrix

	Variable		
Variable	After1-Before1	After2-Before2	After3-Before3
After1-Before1	3.0667	0.3556	-2.0000
After2-Before2	0.0871	5.4333	0.4444
After3-Before3	-0.5710	0.0953	4.0000

This report displays correlations and covariances of the variables, or differences, analyzed. The correlations are shown in the lower-left half of the matrix and the covariances are shown on the diagonal and in the upper-right half of the matrix.

Example 2 – One-Sample T2 Test

This section presents an example of how to run a one-sample T2 analysis of the T2 dataset shown earlier. In this analysis, the analyst wants to test the null hypothesis that the three measurements conform to the response pattern: 30, 33, 30. These values are entered into each row of three new columns: H01, H02, and H03. The analysis will proceed as in the paired test of Example 1.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Hotelling's One-Sample T2 window.

1 Open the T2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **T2.s0**.
- Click **Open**.

2 Open the Hotelling's One-Sample T2 window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Hotelling's One-Sample T2**. The Hotelling's One-Sample T2 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 Hotelling's One-Sample T2 window, select the **Variables tab**.
- Set the **Response Variables** to **AFTER1-AFTER3**.
- Set the **Paired Variables** to **H01-H03**.
- Check **Run Randomization Tests** so that these tests will be included in the reports.

4 Specify the reports.

- On the Hotelling's One-Sample T2 window, select the **Reports tab**.
- Check all of the reports. (Although all reports are not necessary, we will check them all so that they can all be documented.)
- Set **VC Decimals** to **4**.
- Set **Means Decimals** to **4**.

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).

Hotelling’s T2 Test Output

```
Hotelling's T2 Test Section
                                                 Parametric   Randomization
                                                       Test            Test
                                                       Prob            Prob
Hypothesis               T2     DF1        DF2        Level           Level
Means All Zero       98.484       3        9.0       0.0004          0.0020
Means All Equal      51.939       2        9.0       0.0005          0.0030
The randomization test results are based on 1000 Monte Carlo samples.

Individual Variables Section
Variable                 T2     Prob        T2     Prob          T2     Prob
Omitted              Others    Level    Change    Level       Alone    Level
After1-H01           22.002   0.0071    76.482   0.0060       0.569   0.4698
After2-H02           98.226   0.0000     0.259   0.9009       2.221   0.1703
After3-H03           33.000   0.0021    65.485   0.0163       8.079   0.0193

Student's T-Test Section
                                   Parametric    Randomization
                            T2           Test             Test
                            or           Prob             Prob
Variable          |Student's T|         Level            Level
All (T2)                98.484         0.0004           0.0020
After1-H01               0.755         0.4698           0.4510
After2-H02               1.490         0.1703           0.1930
After3-H03               2.842         0.0193           0.0050
The randomization test results are based on 1000 Monte Carlo samples.
These individual t-test significance levels should only be used when the overall T2 value is significant.
```

The significance of the *T*2 value indicates that at least one mean does not equal the hypothesized value. A look at the individual *t*-tests indicates that the significance occurs with the third variable: After3. After1 and After2 are not significantly different from 30 and 33, respectively.

Chapter 410

Hotelling's Two-Sample T2

Introduction

The two-sample Hotelling's *T*2 is the multivariate extension of the common two-group Student's *t*-test. In a *t*-test, differences in the mean response between two populations are studied. *T*2 is used when the number of response variables are two or more, although it can be used when there is only one response variable. The null hypothesis is that the group means for all response variables are equal.

*T*2 makes the usual assumptions of equal variances and normally distributed residuals. Preliminary tests are provided that allow these assumptions to be evaluated. *Randomization tests* are provided that do not rely on these assumptions. These randomization tests should be used whenever you want exact results that do not rely on several assumptions.

Technical Details

Equal Covariance Case

The two-sample *T*2 is used to test the equality of the mean vectors of two populations. Specifically, suppose a set of *p* response variables $Y_1, Y_2, \cdots, Y_p$ is measured for each of two groups. Assume that population 1 is distributed as $N_p(\mu_1, \Sigma_1)$ and population 2 is distributed as $N_p(\mu_2, \Sigma_2)$, where $N_p(\mu, \Sigma)$ is the *p*-variable multivariate normal distribution with mean vector μ and covariance matrix Σ. The null hypothesis that $\mu_1 = \mu_2$ can be tested using the test statistic

$$T2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2)$$

where $\bar{y}_1$ and $\bar{y}_2$ are the two sample mean vectors, n_1 and n_2 are the two sample sizes, and S_{pl}^{-1} is the inverse of the pooled covariance matrix which is calculated using

$$S_{pl} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Here, S_1 and S_2 are the estimated covariance matrices calculated from the two samples.

If we make the additional assumption that $\Sigma_1 = \Sigma_2$, *T*2 follows Hotelling's *T*-squared distribution when the null hypothesis is true. That is, $T2 \sim T^2_{p,n_1+n_2-2}$. Reject the null hypothesis if $T2 \geq T^2_{1-\alpha,p,n_1+n_2-2}$. Note that rejecting the null hypothesis concludes that at least one pair of the *p* sets of group response means are unequal.

Unequal Covariance Case

When the experimental setting or a preliminary test such as Box's *M* test leads us to conclude that $\Sigma_1 \neq \Sigma_2$, an alternative to *T*2 must be used. Several such multivariate Behrens-Fisher tests have been suggested in the statistical literature. Following the suggestions of Rencher (1998) derived from a large simulation study, we use the procedure suggested by Nel and van der Merwe (1986) since it was shown to have near optimal power while maintaining reasonable type-I error rates. The test statistic is computed using the formula

$$T*2 = (\bar{y}_1 - \bar{y}_2)'\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1}(\bar{y}_1 - \bar{y}_2)$$

*T**2 is approximately distributed as $T^2_{p,v}$ where *v* is given by

$$v = \frac{tr\left[\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)\right] + \left[tr\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)\right]^2}{\frac{tr\left[\left(\frac{S_1}{n_1}\right)\left(\frac{S_1}{n_1}\right)\right] + \left[tr\left(\frac{S_1}{n_1}\right)\right]^2}{n_1 - 1} + \frac{tr\left[\left(\frac{S_2}{n_2}\right)\left(\frac{S_2}{n_2}\right)\right] + \left[tr\left(\frac{S_2}{n_2}\right)\right]^2}{n_2 - 1}}$$

Randomization Test

Because of the stringent assumptions that must be made when using this procedure, ***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 sample data, calculating the test statistic for each permutation, and counting the number of permutations that result in a T2 value greater than or equal to the actual T2 value. 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.

Permutation results are provided for the equal covariance case, the unequal covariance case, and for all individual t tests.

Assumptions

The following assumptions are made when using *T*2.

1. Each population follows the multivariate normal distribution.
2. The two samples are independent.
3. The two covariance matrices are equal.

These are a set of restrictive assumptions that must be evaluated for each set of data. Box’s *M* test may be used to test whether two covariances matrices are equal. Unfortunately, the accuracy of Box’s *M* test is very sensitive to departures form multivariate normality (assumption 1).

Data Structure

The data must be entered in a format that places the response variables and values for the group side by side. An example of the data structure for a Hotelling’s *T*2 design is shown below. In this example, *WRATR* and *WRATA* are the two response variables. *Treatment* is the group variable. Note that this database has a fourth variable, *Disability*, that is ignored in this analysis. This database is stored in the file MANOVA1.

MANOVA1 dataset

WRATR	WRATA	Treatment	Disability
115	108	1	1
98	105	1	1
107	98	1	1
90	92	2	1
85	95	2	1
80	81	2	1
100	105	1	2
105	95	1	2
95	98	1	2
70	80	2	2
85	68	2	2
78	82	2	2
89	78	1	3
100	85	1	3
90	95	1	3
65	62	2	3
80	70	2	3

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options control which variables are used in the analysis.

Response Variables

Response Variables

One or more numeric response variables are specified here. If two or more variables are specified, the analysis is based on Hotelling’s T2. If only one variable is specified, the analysis is based on Student's t distribution. If the response variables have only two or three unique values, only use the randomization test results.

Group Variable

Group Variable

This variable contains values that identify the group to which each row belongs. A separate analysis is conducted for each unique pair of group values.

The values in this variable may be text or numeric.

Resampling

Run Randomization Tests

Check this option to run randomization tests. Note that these tests are computer-intensive and may require a great deal of time to run.

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 these tests.

Somewhere between 1,000 and 100,000 Monte Carlo samples are usually necessary. We suggest the use of 10,000.

Reports Tab

Select Reports

Means and Std. Deviations ... VC Determinants

Specify whether to display the various reports.

Report Options

Confidence Coefficient

Specify the value of confidence coefficient for the confidence intervals. This is the value of one minus alpha. The value 0.95 is commonly used. However, you can specify any value between 0.50 and 0.99999.

Variable Names

Indicate whether to display the variable names or the variable labels.

Value Labels

Indicate whether to display the data values or their labels.

Precision

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

Report Options – Decimal Places

T2 ... Correlation 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.

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 – Hotelling’s Two-Sample T2 Test

This section presents an example of how to run an analysis of the MANOVA1 data shown earlier. In this analysis, two groups are to be compared on two variables: WRATR and WRATA.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Hotelling’s Two-Sample T2 window.

1 Open the MANOVA1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MANOVA1.s0**.
- Click **Open**.

2 Open the Hotelling’s Two-Sample T2 window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Hotelling’s Two-Sample T2**. The Hotelling’s Two-Sample T2 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 Hotelling’s Two-Sample T2 window, select the **Variables tab**.
- Set the **Response Variables** to **WRATR-WRATA**.
- Set the **Group Variable** to **Treatment**.
- Check **Run Randomization Tests** so that these tests will be included in the reports.

4 Specify the reports.

- On the Hotelling’s Two-Sample T2 window, select the **Reports tab**.
- Check all of the reports. (Although all reports are not necessary, we will check them all so that they can all be documented.)

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

```
               Means                          Standard Deviations
Variable       1             2                1             2
WRATR          99.88889      78.33334         8.283182      8.046739
WRATA          96.33334      78.11111         9.746795      10.94811
Count          9             9                9             9
```

This report provides the means, standard deviations, and counts of each group. Look the values over to be certain that the right variables were selected.

Hotelling’s T2 Test Section

```
                                                 Parametric    Randomization
                                                       Test             Test
Covariance                                             Prob             Prob
Assumption       T2        DF1      DF2               Level            Level
Equal         31.640         2      16.0             0.0003           0.0020
Unequal       31.640         2      15.9             0.0003           0.0020
The randomization test results are based on 1000 Monte Carlo samples.
```

This report gives the results of the two *T*2 tests: first for equal, and then for unequal, covariance matrices.

Covariance Assumption

Indicates the type of *T*2 test displayed on this line. Either *Equal* or *Unequal* are possible. Even though you actually use only one of these two tests, both are reported here.

T2

The values of the two test statistics are shown in this column. The top value is *T*2. The bottom value is *T**2. Note that these value will be equal if the two sample sizes are equal.

DF1

This is the number of response variables.

DF2

For the top line, this is $n_1 + n_2 - 2$. For the bottom line, this is the value of *v*.

Parametric Test Prob Level

This is the significance level of the test. If this value is less than 0.05, we say that the test was significant at the 0.05 level. If the value is less than 0.01, we say that the test was significant at the 0.01 level. This result is accurate if all the assumptions of the corresponding test are met.

Randomization Test Prob Level

This is the significance level of the randomization test. If this value is less than 0.05, we say that the test was significant at the 0.05 level. If the value is less than 0.01, we say that the test was significant at the 0.01 level.

This result is an exact result even if the data were not obtained by random sampling. The result is accurate even if the response variables were binary.

Individual Variables Section

```
Variable         T2      Prob            T2      Prob            T2     Prob
Omitted      Others     Level        Change     Level         Alone     Level
WRATR        13.909    0.0018        17.731    0.0093        31.357     0.0000
WRATA        31.357    0.0000         0.283    0.7688        13.909     0.0018
```

This report provides information about the influence of each of the individual response variables on the overall *T*2 value. This is accomplished by calculating the change in *T*2 when a response variable is omitted.

Variable Omitted

This is the name of the variable shown on this line of the report.

T2 Others

This is the value of *T*2 calculated with all response variables except the variable listed to the left.

Prob Level

This is the significance level of the *T*2 value shown on the report to the left of this value.

T2 Change

This is the amount that *T*2 is reduced when the response variable shown on this line is omitted.

Prob Level

This is the significance level of the *T*2 change value shown on the report to the left of this value. It is computed using the fact that the change in *T*2 is related to an *F* distribution using the formula

$$F_{\alpha,1,n_1+n_2-2-(p-1)} = \frac{T_p^2 - T_{p-1}^2}{n_1 + n_2 - 2 + T_{p-1}^2}$$

Note that this quantity tests the drop in *T*2 when a variable is removed conditional on the other response variables that are included. Another way of looking at this quantity is that it tests whether the variable omitted significantly increases the distance between the two populations. See Rencher (1998) page 109 for further details.

T2 Alone

This is the value of *T*2 calculated when only this response variable is used. It is the square of the common *t* test. It is the two-sided test of the null hypothesis that the means for this variable are equal, ignoring all other variables.

Prob Level

This is the two-sided significance level of the above *T*2 value.

Student’s T-Test Section

```
                                   Parametric      Randomization
                           T2            Test               Test
                           or            Prob               Prob
Variable        |Student's T|           Level              Level
All (T2)               31.640          0.0003             0.0030
WRATR                   5.600          0.0000             0.0010
WRATA                   3.729          0.0018             0.0020
The randomization test results are based on 1000 Monte Carlo samples.
```

This report provides the results of individually conducting a two-sided *t*-test on each of the response variables. You might think that since there are a series of *p* t-tests being employed, a Bonferroni adjustment should be applied to the significance levels. However, if these individual tests are only considered when the overall *T*2 is significant at the same level, such as 0.05, then their significance levels are “protected” by the *T*2 test and the unadjusted significance levels given here can be used.

Variable

The variable whose results are presented on this line. Note that the first line gives the overall results for *T*2.

T2 or |Student's T|

The first line is the value of *T*2. The other lines are the two-sided Student's *t*-test values.

Parametric Test Prob Level

These are the significance levels of the test statistics given to the left. Note that if the individual tests are only used when the overall test is significant, these significance levels are accurate even though several individual tests are made. The multivariate *T*2 test is said to "protect" the significance levels of the individual tests.

Randomization Test Prob Level

These are the results of randomization tests that are run on each of the variables. These tests are exact when the Monte Carlo sample size is large, say over 5000. These tests should be used when there is a even a hint that the regular assumptions of the *t*-tests are not valid. For example, this significance level is accurate even when the response variable takes on binary values (the *t*-test assumes a continuous, normal response variable).

Note that these values will change from run to run. As you increase the number of Monte Carlo samples, these values will become more and more stable. You may have to go as large as 100,000 before the results remain the same from run to run. This instability is due to the our use of a random sample of all the trillions of permutations that are possible. As you increase the Monte Carlo sample size, you reduce the sampling error (and greatly increase the time it takes to generate the results).

Confidence Intervals for the Mean Differences

```
                          Lower 95.0%   Upper 95.0%   Lower 95.0%   Upper 95.0%
                           Bonferroni    Bonferroni  Simultaneous  Simultaneous
Variable     Difference    Conf. Limit   Conf. Limit   Conf. Limit   Conf. Limit
WRATR          21.55556       12.03645      31.07466       10.7665      32.34611
WRATA          18.22222       6.139622      30.30482      4.527671      31.91677
```

This report provides confidence intervals for the differences between the group means for each response variable. Two intervals are provided: Bonferroni and simultaneous.

Variable

The variable whose results are presented on this line.

Difference

The actual difference between the means for the corresponding response variable.

Bonferroni Confidence Interval

Bonferroni confidence intervals are based on the formula

$$\bar{y}_{1j} - \bar{y}_{2j} \pm t_{\alpha/2p,n_1+n_2-2}\sqrt{\left(\frac{n_1+n_2}{n_1 n_2} s_{pl,jj}\right)}$$

This formula is derived by applying a Bonferroni adjustment to the regular univariate confidence interval. This adjustment is made by dividing the alpha level by *p*, the number of such intervals to be created. These intervals are usually not as wide as the simultaneous intervals, yet still have an appropriate adjustment because of the multiple intervals that are being created.

Simultaneous Confidence Interval

Simultaneous confidence intervals are based on the formula

$$\bar{y}_{1j} - \bar{y}_{2j} \pm \sqrt{T^2_{1-\alpha,p,n_1+n_2-2}\left(\frac{n_1+n_2}{n_1 n_2} s_{pl,jj}\right)}$$

This formula is derived from a formula for confidence intervals for *any* linear combination of the mean differences, including those that are generated after looking at the data. Because of this, these confidence intervals are extra wide and may not be of must use.

Bartlett-Box Homogeneity of Variance Tests

```
Variable(s)                Test                      F         F      Chi2      Chi2
Tested                    Value   DF1    DF2    Approx      Prob    Approx      Prob
Box's M Test
ALL                       0.153     3  46080     0.044    0.9877     0.133    0.9877
Bartlett Individual Variable Tests
WRATR                     0.007     1    768     0.006    0.9367     0.006    0.9368
WRATA                     0.108     1    768     0.101    0.7503     0.101    0.7505
```

This report provides a preliminary test of the assumption of equality of covariance matrices. If the data fails this test, you should use the unequal variance version of the *T*2 test or the randomization test. The calculation of these tests is documented in the Technical Details section of the Equality of Covariance Matrices chapter and will not be repeated here.

Box's *M* test is very sensitive to non-normality. A significant value indicates either unequal covariance matrices, non-normality, or both. Hence, it is important to establish multivariate normality before concluding unequal covariance matrices using Box's *M* test.

The Chi-square approximation should be used when all group sample sizes are greater than 20 and *p* is less than 6. Otherwise, the F approximation is more accurate. ***NCSS*** supplies both the multivariate Box's M test and the individual Bartlett's tests so that when Box's *M* test is significant, you can determine which variables contribute to the variance.

Covariance and Correlation Matrix Determinants

```
                Log of Covariance        Correlation
Treatment             Determinant        Determinant
1                          8.3863             0.6730
2                          8.4949             0.6301
All                        8.4502             0.6528
```

This report gives the logarithm (base *e*) of the determinant of each of the covariance matrices and the determinant of each of the correlation matrices. The assumption of equality of covariance matrices forces us to also assume that these values are equal. Box's *M* test compares these values.

Eigenvalues of Covariance Matrices

	Treatment		
Number	1	2	All
1	129.8207	152.5590	140.9319
2	33.7904	32.0521	33.1793

This report gives the eigenvalues of each of the individual covariance matrices as well as the pooled covariance matrix. Each column gives a set of eigenvalues. These eigenvalues summarize the covariance matrix into a few values. By comparing the largest eigenvalues across both groups, you can determine if the groups are different. Also, eigenvalues near zero indicate singularities in your data.

Eigenvalues of Correlation Matrices

	Treatment		
Number	1	2	All
1	1.5718	1.6082	1.5893
2	0.4282	0.3918	0.4107

This report gives the eigenvalues of each of the individual correlation matrices followed by the eigenvalues of the within-group correlation matrix. Each column gives a set of eigenvalues.

This report is useful because the eigenvalues summarize the correlation matrix. By comparing the largest eigenvalues across all groups, you can determine if the groups are different. Also, eigenvalues near zero indicate singularities in your data.

Within Group Correlations\Covariances

Within Group Correlation\Covariance For Treatment = All

	Variable	
Variable	WRATR	WRATA
WRATR	66.68056	49.875
WRATA	0.5893	107.4306

Within Group Correlation\Covariance For Treatment = 1

	Variable	
Variable	WRATR	WRATA
WRATR	68.61111	46.16667
WRATA	0.5718	95

Within Group Correlation\Covariance For Treatment = 2

	Variable	
Variable	WRATR	WRATA
WRATR	64.75	53.58333
WRATA	0.6082	119.8611

This report displays correlations and covariances. The covariance matrices were labeled S_{sp}, S_1, and S_2 in the formulas given earlier in this chapter. The correlations are shown in the lower-left half of the matrix and the covariances are shown on the diagonal and in the upper-right half of the matrix.

Chapter 415

Multivariate Analysis of Variance (MANOVA)

Introduction

Multivariate analysis of variance (MANOVA) is an extension of common analysis of variance (ANOVA). In ANOVA, differences among various group means on a single-response variable are studied. In MANOVA, the number of response variables is increased to two or more. The hypothesis concerns a comparison of vectors of group means. When only two groups are being compared, the results are identical to Hotelling's T^2 procedure.

The multivariate extension of the F-test is not completely direct. Instead, several test statistics are available, such as Wilks' Lambda and Lawley's trace. The actual distributions of these statistics are difficult to calculate, so we rely on approximations based on the F-distribution.

Technical Details

A MANOVA has one or more factors (each with two or more levels) and two or more dependent variables. The calculations are extensions of the general linear model approach used for ANOVA.

Unlike the univariate situation in which there is only one statistical test available (the F-ratio), the multivariate situation provides several alternative statistical tests. We will describe these tests in terms of two matrices, *H* and *E*. *H* is called the *hypothesis matrix* and *E* is the *error matrix*. These matrices may be computed using a number of methods. In ***NCSS***, we use the standard general linear models (GLM) approach in which a sum of squares and cross-products matrix is computed. This matrix is based on the dependent variables and independent variables generated for each degree of freedom in the model. It may be partitioned according to the terms in the model.

MANOVA Test Statistics

For a particular p-variable multivariate test, assume that the matrices *H* and *E* have *h* and *e* degrees of freedom, respectively. Four tests may be defined as follows. See Seber (1984) for details. Let θ_i, ϕ_i, and λ_i be the eigenvalues of $H(E+H)^{-1}$, HE^{-1}, and $E(E+H)^{-1}$ respectively. Note that these eigenvalues are related as follows:

$$\theta_i = 1 - \lambda_i = \frac{\phi_i}{1+\phi_i}$$

$$\phi_i = \frac{\theta_i}{1-\theta_i} = \frac{1-\lambda_i}{\lambda_i}$$

$$\lambda_i = 1 - \theta_i = \frac{1}{1+\phi_i}$$

Wilks' Lambda

Define Wilks' Lambda as follows:

$$\Lambda_{p,h,e} = \frac{|E|}{|E+H|} = \prod_{j=1}^{p}(1-\theta_j)$$

with $e \geq p$.

The following approximation based on the F-distribution is used to determine significance levels:

$$F_{ph,ft-g} = \frac{(ft-g)(1-\Lambda^{1/t})}{ph\,\Lambda^{1/t}}$$

where

$$f = e - \frac{1}{2}(p-h+1)$$

$$g = \frac{ph-2}{2}$$

$$t = \begin{cases} \sqrt{\dfrac{p^2h^2-4}{p^2+h^2-5}} & \text{if } p^2+h^2-5>0 \\ 1 & \text{otherwise} \end{cases}$$

This approximation is exact if *p* or $h \geq 2$.

Lawley - Hotelling Trace

The trace statistic, T_g^2, is defined as follows:

$$T_g^2 = e\sum_{j=1}^{s}\phi_j$$

where

$$s = \min(p,h)$$

The following approximation based on the F-distribution is used to determine significance levels:

$$F_{a,b} = \frac{T_g^2}{ce}$$

where

$$a = ph$$

$$b = 4 + (a + 2)/(B - 1)$$

$$c = \frac{a(b - 2)}{b(e - p - 1)}$$

$$B = \frac{(e + h - p - 1)(e - 1)}{(e - p - 3)(e - p)}$$

Pillai’s Trace

Pillai’s trace statistic, $V^{(s)}$, is defined as follows:

$$V^{(s)} = \sum_{j=1}^{s}\theta_j = tr(H(E + H\ )^{-1})$$

where

$$s = \min(p,h)$$

The following approximation based on the F-distribution is used to determine significance levels:

$$F_{s(2m+s+1),s(2n+s+1)} = \frac{(2n + s + 1)V^{(s)}}{(2m + s + 1)(s - V^{(s)})}$$

where

$$s = \min(p,h)$$

$$m = (|\,p - h\,| -1)/2$$

$$n = (e - p - 1)/2$$

Roy's Largest Root

Roy's largest root, $\phi_{\max}$, is defined as the largest of the ϕ_i 's. The following approximation based on the F-distribution is used to determine significance levels:

$$F_{(2\nu_1+2),(2\nu_2+2)} = \frac{2\nu_2+2}{2\nu_1+2}\phi_{\max}$$

where

$$s = \min(p,h)$$

$$\nu_1 = (|\,p - h\,| - 1)/2$$

$$\nu_2 = (e - p - 1)/2$$

Which Test to Use

When the hypothesis degrees of freedom, *h*, is one, all four test statistics will lead to identical results. When *h*>1, the four statistics will usually lead to the same result. When they do not, the following guidelines from Tabachnick (1989) may be of some help.

Wilks' Lambda, Lawley's trace, and Roy's largest root are often more powerful than Pillai's trace if *h*>1 and one dimension accounts for most of the separation among groups. Pillai's trace is more robust to departures from assumptions than the other three.

Tabachnick (1989) provides the following checklist for conducting a MANOVA. We suggest that you consider these issues and guidelines carefully.

Assumptions and Limitations

The following assumptions are made when using a MANOVA.

1. The response variables are continuous.
2. The residuals follow the multivariate-normal probability distribution with means equal to zero.
3. The variance-covariance matrices of each group of residuals are equal.
4. The individuals are independent.

Multivariate Normality and Outliers

MANOVA is robust to modest amount of skewness in the data. A sample size that produces 20 degrees of freedom in the univariate F-test is adequate to ensure robustness. Non-normality caused by the presence of outliers can cause severe problems that even the robustness of the test will not overcome. You should screen your data for outliers and run it through various univariate and multivariate normality tests and plots to determine if the normality assumption is reasonable.

Homogeneity of Covariance Matrices

MANOVA makes the assumption that the within-cell (group) covariance matrices are equal. If the design is balanced so that there is an equal number of observations in each cell, the robustness of the MANOVA tests is guaranteed. If the design is unbalanced, you should test the equality of covariance matrices using Box's M test. If this test is significant at less than .001, there may be severe distortion in the alpha levels of the tests. You should only use Pillai's trace criterion in this situation.

Linearity

MANOVA assumes linear relationships among the dependent variables within a particular cell. You should study scatter plots of each pair of dependent variables using a different color for each level of a factor. Look carefully for curvilinear patterns and for outliers. The occurrence of curvilinear relationships will reduce the power of the MANOVA tests.

Multicollinearity and Singularity

Multicollinearity occurs when one dependent variable is almost a weighted average of the others. This collinearity may only show up when the data are considered one cell at a time. The R^2-Other Y's in the Within-Cell Correlations Analysis report lets you determine if multicollinearity is a problem. If this R^2 value is greater than .99 for any variable, you should take corrective action (remove one of the variables). To correct for multicollinearity, begin removing the variables one at a time until all of the R^2's are less than .99. Do not remove them all at once! *Singularity* is the extreme form of multicollinearity in which the R^2 value is one.

Forms of multicollinearity may show up when you have very small cell sample sizes (when the number of observations is less than the number of variables). In this case, you must reduce the number of dependent variables.

Data Structure

The data must be entered in a format that places the dependent variables and values of each factor side by side. An example of the data for a MANOVA design is shown in the table below. In this example, *WRATR* and *WRATA* are the two dependent variables. *Treatment* and *Disability* are two factor variables. This database is stored in the file MANOVA1.

MANOVA1 dataset (subset)

WRATR	WRATA	Treatment	Disability
115	108	1	1
98	105	1	1
107	98	1	1
90	92	2	1
85	95	2	1
80	81	2	1
100	105	1	2
105	95	1	2
95	98	1	2
70	80	2	2

Unequal Sample Size and Missing Data

You should begin by screening your data. Pay particular attention to patterns of missing values. When using MANOVA, you should have more observations per factor category than you have dependent variables so that you can test the equality of covariance matrices using Box’s M test.

NCSS ignores rows with missing values. If it appears that most of the missing values occur in one or two variables, you might want to leave these out of the analysis in order to obtain more data and hence more power.

NCSS uses the GLM procedure for calculating the hypothesis and error matrices. Each matrix is calculated as if it were fit last in the model. This is the recommended way of obtaining these matrices. This method is valid even when the sample sizes for the various groups are unequal.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These options control which variables are used in the analysis.

Response Variables

Response Variables

Specifies the response (dependent) variables to be analyzed.

Factor Specification

Factor Variable (1-10)

At least one factor variable must be specified. This variable’s values indicates how the values of the response variable should be categorized. Examples of factor variables are gender, age groups, ‘yes’ or ‘no’ responses, etc. Note that the values in the variable may be either numeric or text. The treatment of text variables is specified for each variable by the Data Type option on the data base.

Type

This option specifies whether the factor is fixed or random.

- **Fixed**

  The factor includes all possible levels, like male and female for gender, includes representative values across the possible range of values, like low, medium, and high temperatures, or includes a set of values to which inferences will be limited, like New York, California, and Maryland.

- **Random**

  The factor is one in which the chosen levels represent a random sample from the population of values. For example, you might select four classes from the hundreds in your state or you might select ten batches from an industrial process. The key is that a random sample is chosen. In ***NCSS***, a random factor is "crossed" with other random and fixed factors. Two factors are crossed when each level of one includes all levels of the other.

Model Specification

This section specifies the experimental design model.

Which Model Terms

A design in which main effect and interaction terms are included is called a saturated model. Often, it is useful to omit various interaction terms from the model. This option lets you specify which interactions to keep very easily. If the selection provided here is not flexible enough for your needs, you can specify custom here and enter the model directly.

The options included are as follows.

- **Full Model**

  The complete, saturated model is analyzed. This option requires that you have no missing cells, although you can have an unbalanced design. Hence, you cannot use this option with Latin square or fractional factorial designs.

- **Up to 1-Way**

  A main-effects only model is run. All interactions are omitted.

- **Up to 2-Way**

  All main-effects and two-way interactions are included in the model.

- **Up to 3-Way**

  All main-effects, two-way, and three-way interactions are included in the model.

- **Up to 4-Way**

  All main-effects, two-way, three-way, and four-way interactions are included in the model.

- **Custom**

  This option indicates that you want the Custom Model (given in the next box) to be used.

Write Model in 'Custom Model' Field

When this option is checked, no 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 delete selected terms from the model without having to enter all the terms you want to keep.

Custom Model

When a Custom Model is selected (see Which Model Terms), the model itself is entered here. If all main effects and interactions are desired, you can enter the word "ALL" here. For complicated designs, it is usually easier to check the next option, Write Model in 'Custom Model' Field, and run the procedure. The appropriate model will be generated and placed in this box. You can then edit it as you desire.

The model is entered using letters separated by the plus sign. For example, a three-factor factorial in which only two-way interactions are needed would be entered as follows:

A+B+AB+C+AC+BC

Note that repeated-measures designs are not allowed.

Reports Tab

Select Reports

EMS Report ... Univariate F's

Specify whether to display the indicated reports.

Select Plots

Means Plot(s) and Subject Plot

Specify whether to display the indicated plots.

Report Options

Test Alpha

The value of alpha for the statistical tests and power analysis. 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 where available. You should determine a value appropriate for your particular study.

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

Indicate whether to display the variable names or the variable labels.

Value Labels

Indicate whether to display the data values or their labels.

Means Plot Tab

The following few options specify the plot(s) of group means.

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.

Y Scaling

Specify the method for calculating the minimum and maximum along the vertical axis. *Separately* means that each plot is scaled independently. *Uniform* means that all plots use the overall minimum and maximum of the data. This option is ignored if a minimum or maximum is specified.

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.

Connect Lines

Click this box to connect the points for a particular factor. This makes it easier to spot patterns in the means.

Plot Settings – Legend

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 appropriate factor name.

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.

Symbols Tab

Plotting Symbols

Group (1-15)

The symbol used to designate a point on the scatter plot. Each option represents the corresponding factor.

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 – Multivariate Analysis of Variance

This section presents an example of how to run an analysis of the data contained in the MANOVA1 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Multivariate Analysis of Variance (MANOVA) window.

1 Open the MANOVA1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MANOVA1.s0**.
- Click **Open**.

2 Open the MANOVA window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **MANOVA**. The Multivariate Analysis of Variance (MANOVA) 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 Multivariate Analysis of Variance (MANOVA) window, select the **Variables tab**.
- Double-click in the **Response Variable** box. This will bring up the variable selection window.
- Select **WRATR** and **WRATA** from the list of variables and then click **Ok**. "WRATR-WRATA" will appear in the Response Variable box.
- Double-click in the **first Factor Variable** box. This will bring up the variable selection window.
- Select **Treatment** from the list of variables and then click **Ok**. "Treatment" will appear in the first Factor Variable box.
- Double-click in the **second Factor Variable** box. This will bring up the variable selection window.
- Select **Disability** from the list of variables and then click **Ok**. "Disability" will appear in the second Factor Variable 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).

Expected Mean Squares Section

```
Expected Mean Squares Section

Source                     Term       Denominator    Expected
Term              DF       Fixed?     Term           Mean Square
A (Treatment)     1        Yes        S(AB)          S+bsA
B (Disability)    2        Yes        S(AB)          S+asB
AB                2        Yes        S(AB)          S+sAB
S(AB)             12       No                        S
Note: Expected Mean Squares are for the balanced cell-frequency case.
```

The Expected Mean Square expressions are provided to show the appropriate error term for each factor. The correct error term for a factor is that term that is identical except for the factor being tested.

Source Term

The source of variation or term in the model.

DF

The degrees of freedom. The number of observations "used" by this term.

Term Fixed?

Indicates whether the term is "fixed" or "random."

Denominator Term

Indicates the term used as the denominator in the F-ratio.

Expected Mean Square

This expression represents the expected value of the corresponding mean square if the design was completely balanced. "S" represents the expected value of the mean square error (sigma). The uppercase letters represent either the adjusted sum of squared treatment means if the factor is fixed, or the variance component if the factor is random. The lowercase letter represents the number of levels for that factor, and “s” represents the number of replications of the experimental layout.

These EMS expressions are provided to determine the appropriate error term for each factor. The correct error term for a factor is that term whose EMS is identical except for the factor being tested.

MANOVA Tests Section

```
Manova Tests Section

Term(DF)                       Test                              Prob    Decision
Test Statistic                Value   DF1  DF2  F-Ratio         Level    (0.05)
A(1):Treatment
Wilks' Lambda              0.137721     2   11    34.44      0.000018    Reject
Hotelling-Lawley Trace     6.261036     2   11    34.44      0.000018    Reject
Pillai's Trace             0.862279     2   11    34.44      0.000018    Reject
Roy's Largest Root         6.261036     2   11    34.44      0.000018    Reject
WRATR                    2090.88889     1   12    46.12      0.000019    Reject
WRATA                    1494.22222     1   12    33.25      0.000089    Reject
B(2):Disability
Wilks' Lambda              0.255263     4   22     5.39      0.003528    Reject
Hotelling-Lawley Trace     2.895034     4   20     7.24      0.000896    Reject
Pillai's Trace             0.750481     4   24     3.60      0.019460    Reject
Roy's Largest Root         2.887241     2   12    17.32      0.000290    Reject
WRATR                    260.388889     2   12     5.74      0.017784    Reject
WRATA                    563.388889     2   12    12.54      0.001151    Reject
AB(2)
Wilks' Lambda              0.908068     4   22     0.27      0.893037    Accept
Hotelling-Lawley Trace     0.100954     4   20     0.25      0.904790    Accept
Pillai's Trace             0.092192     4   24     0.29      0.881598    Accept
Roy's Largest Root         0.098039     2   12     0.59      0.570550    Accept
WRATR                      1.055556     2   12     0.02      0.977029    Accept
WRATA                     26.388889     2   12     0.59      0.571116    Accept
```

This report gives the results of the various significance tests. Usually, the four multivariate tests will lead to the same conclusions. When they do not, refer to the discussion of these tests found earlier in this chapter. Once a multivariate test has found a term significant, use the univariate ANOVA to determine which of the variables and factors are “causing” the significance.

Term(DF)

The term in the design model with the degrees of freedom of the term in parentheses.

Test Statistic

The name of the statistical test shown on this row of the report. The four multivariate tests are followed by the univariate F-tests of each variable.

Test Value

The value of the test statistic.

DF1

The numerator degrees of freedom of the F-ratio corresponding to this test.

DF2

The denominator degrees of freedom of the F-ratio corresponding to this test.

F-Ratio

The value of the F-test corresponding to this test. In some cases, this is an exact test. In other cases, this is an approximation to the exact test. See the discussion of each test to determine if it is exact or approximate.

Prob Level

The significance level of the above F-ratio. The probability of an F-ratio larger than that obtained by this analysis. For example, to test at an alpha of 0.05, this probability would have to be less than 0.05 to make the F-ratio significant.

Decision(0.05)

The decision to accept or reject the null hypothesis at the given level of significance. Note that you specify the level of significance when you select Alpha.

Within Correlations\Covariances Section

```
Within Correlations\Covariances Section

          WRATR       WRATA
WRATR     45.33333    2.583333
WRATA     0.0572313   44.94444
```

This report displays the correlations and covariances formed by averaging across all the individual group covariance matrices. The correlations are shown in the lower-left half of the matrix. The within-group covariances are shown on the diagonal and in the upper-right half of the matrix.

Within-Cell Correlations Analysis Section

```
Within-Cell Correlations Analysis

            R-Squared   Canonical                Percent     Cumulative
Variable    Other Y's   Variate     Eigenvalue   of Total    Total
Wratr       0.003275    1           1.057231     52.86       52.86
Wrata       0.003275    2           0.942769     47.14       100.00
```

This report analyzes the within-cell correlation matrix. It lets you diagnose multicollinearity problems as well as determine the number of dimensions that are being used. This is useful in determining if Pillai’s trace should be used.

R-Squared Other Y’s

This is the R-Squared index of this variable with the other variables. When this value is larger than 0.99, severe multicollinearity problems exist. If this happens, you should remove the variable with the largest R-Squared and re-run your analysis.

Canonical Variate

The identification numbers of the canonical variates that are generated during the analysis. The total number of variates is the smaller of the number of variables and the number of degrees of freedom in the model.

Eigenvalue

The eigenvalues of the within correlation matrix. Note that this value is not associated with the variable at the beginning of the row, but rather with the canonical variate number directly to the left.

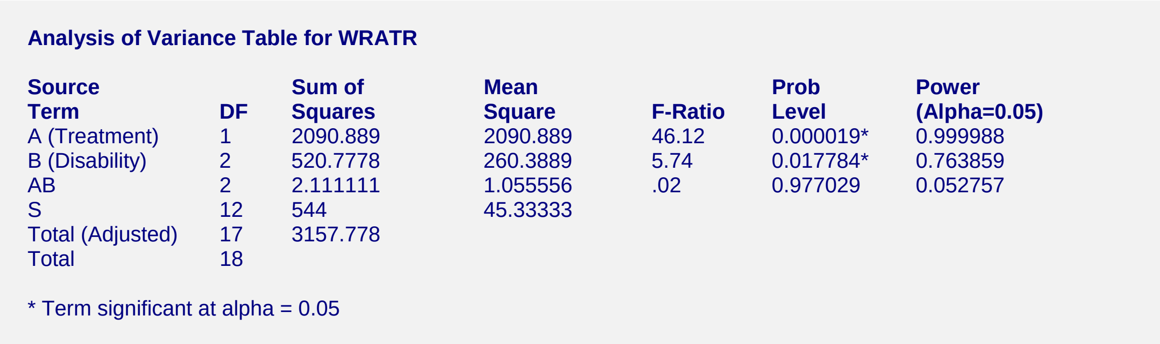
Percent of Total

The percent that the eigenvalue is of the total. Note that the sum of the eigenvalues will equal the number of variates. If the percentage accounted for by the first eigenvalue is relatively large (70 or 80 percent), Pillai's trace will be less powerful than the other three multivariate tests.

Cumulative Total

The cumulative total of the Percent of Total column.

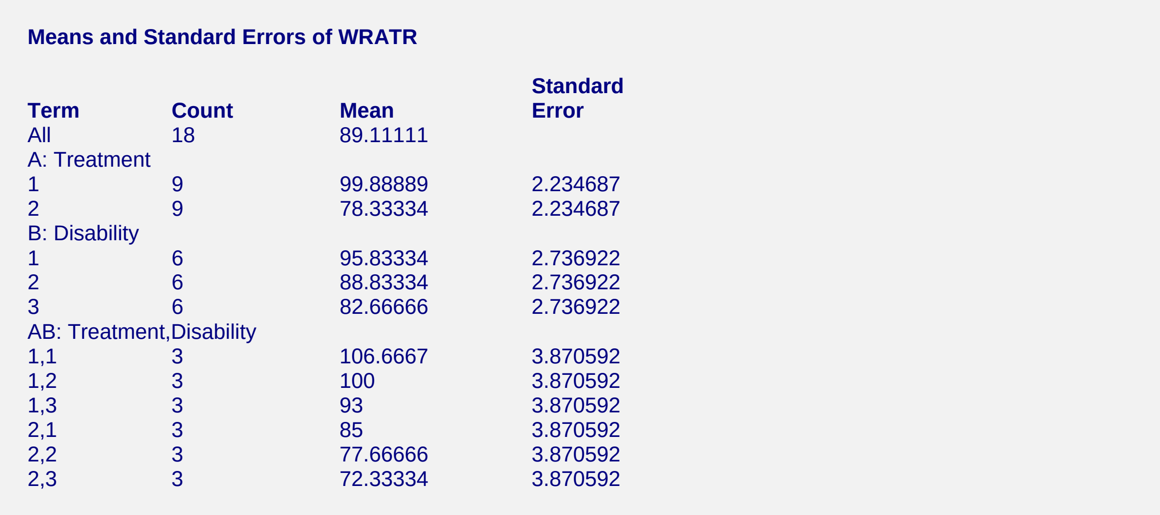
Univariate Analysis of Variance Section



```
Analysis of Variance Table for WRATR

Source                       Sum of          Mean                      Prob         Power
Term               DF        Squares         Square         F-Ratio    Level        (Alpha=0.05)
A (Treatment)      1         2090.889        2090.889       46.12      0.000019*    0.999988
B (Disability)     2         520.7778        260.3889       5.74       0.017784*    0.763859
AB                 2         2.111111        1.055556       .02        0.977029     0.052757
S                  12        544             45.33333
Total (Adjusted)   17        3157.778
Total              18

* Term significant at alpha = 0.05
```

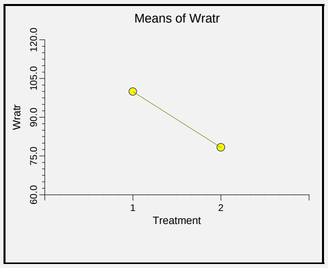
This is the standard ANOVA report as documented in the General Linear Models chapter. A separate report is displayed for each of the dependent variables.

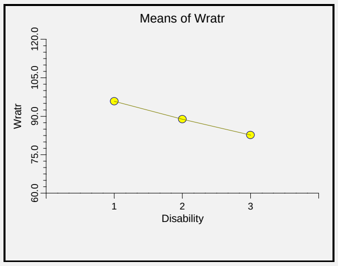
Means and Plots Section

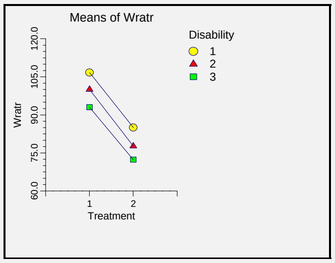


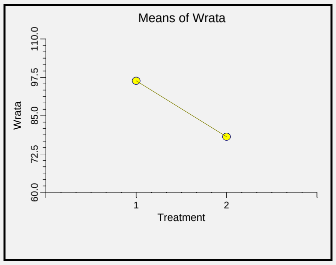
```
Means and Standard Errors of WRATR

                                                  Standard
Term            Count           Mean              Error
All             18              89.11111
A: Treatment
1               9               99.88889          2.234687
2               9               78.33334          2.234687
B: Disability
1               6               95.83334          2.736922
2               6               88.83334          2.736922
3               6               82.66666          2.736922
AB: Treatment,Disability
1,1             3               106.6667          3.870592
1,2             3               100               3.870592
1,3             3               93                3.870592
2,1             3               85                3.870592
2,2             3               77.66666          3.870592
2,3             3               72.33334          3.870592
```

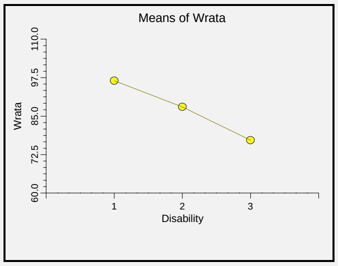
Plots Section

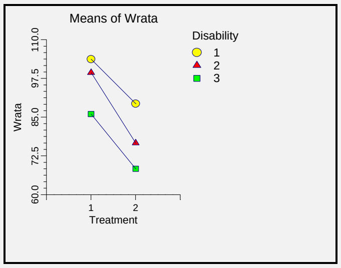












This report provides the least-squares means and standard errors for each variable. Note that the standard errors are calculated from the mean square error of the ANOVA table. They are not the standard errors that would be calculated from the individual cells.

Chapter 420

Factor Analysis

Introduction

Factor Analysis (FA) is an exploratory technique applied to a set of observed variables that seeks to find underlying factors (subsets of variables) from which the observed variables were generated. For example, an individual's response to the questions on a college entrance test is influenced by underlying variables such as intelligence, years in school, age, emotional state on the day of the test, amount of practice taking tests, and so on. The answers to the questions are the observed variables. The underlying, influential variables are the factors.

Factor analysis is carried out on the correlation matrix of the observed variables. A factor is a weighted average of the original variables. The factor analyst hopes to find a few factors from which the original correlation matrix may be generated.

Usually the goal of factor analysis is to aid data interpretation. The factor analyst hopes to identify each factor as representing a specific theoretical factor. Therefore, many of the reports from factor analysis are designed to aid in the interpretation of the factors.

Another goal of factor analysis is to reduce the number of variables. The analyst hopes to reduce the interpretation of a 200-question test to the study of 4 or 5 factors. One of the most subtle tasks in factor analysis is determining the appropriate number of factors.

Factor analysis has an infinite number of solutions. If a solution contains two factors, these may be rotated to form a new solution that does just as good a job at reproducing the correlation matrix. Hence, one of the biggest complaints of factor analysis is that the solution is not unique. Two researchers can find two different sets of factors that are interpreted quite differently yet fit the original data equally well.

NCSS provides the *principal axis method* of factor analysis. The results may be rotated using varimax or quartimax rotation. The factor scores may be stored for further analysis.

Many books are devoted to factor analysis. We suggest you obtain a book on the subject from an author in your own field. An excellent introduction to the subject is provided by Tabachnick (1989).

Technical Details

Mathematical Development

This section will document the basic formulas used by **NCSS** in performing a factor analysis. The following table lists many of the matrices that are used in the discussion to follow.

Label	Matrix Name	Size	Description
R	Correlation	pxp	Matrix of correlations between each pair of variables.
X	Data	Nxp	Observed data matrix with N rows (observations) and p columns (variables).
Z	Standardized data	Nxp	Matrix of standardized data. The standardization of each variable is made by subtracting its mean and dividing by its standard deviation.
A	Factor loading	pxm	Matrix of correlations between the original variables and the factors. Also represents the contribution of each factor in estimating the original variables.
L	Eigenvalue	mxm	Diagonal matrix of eigenvalues. Only the first m eigenvalues are considered.
V	Eigenvector	pxm	Matrix of eigenvectors. Only the first m columns of this matrix are used.
B	Factor-score coefficients	pxm	Matrix of regression weights used to construct the factor scores from the original variables.
U	Uniqueness	pxp	Matrix of uniqueness values.
F	Factor score	Nxm	Matrix of factor scores. For each observation in the original data, the values of each of the retained factors are estimated. These are the factor scores.

The principal-axis method is used by **NCSS** to solve the factor analysis problem. Factor analysis assumes the following partition of the correlation matrix, *R*:

$$R = AA' + U$$

The principal-axis method proceeds according to the following steps:

1. Estimate *U* from the communalities as discussed below.
2. Find *L* and *V*, the eigenvalues and eigenvectors of *R-U* using standard eigenvalue analysis.
3. Calculate the loading matrix as follows:

$$A = VL^{\frac{1}{2}}$$

4. Calculate the score matrix as follows:

$$B = VL^{-\frac{1}{2}}$$

5. Calculate the factor scores as follows:

$$F = ZB$$

Steps 1-3 may be iterated since a new *U* matrix may be estimated from the current loading matrix.

Initial Communality Estimation

We close this section with a discussion of obtaining an initial value of U. **NCSS** uses the initial estimation of Cureton (1983), which will be outlined here. The initial communality estimates, c_{ii}, are calculated from the correlation and inverse correlation matrices as follows:

$$c_{ii} = \left(1 - \frac{1}{R^{ii}}\right) \frac{\sum_{k=1}^{p} \max_{over\ j \neq k} \left(\left|r_{jk}\right|\right)}{\sum_{k=1}^{p} \left(1 - \frac{1}{R^{kk}}\right)}$$

where R_{ii} is the i^{th} diagonal element of R-1 and r_{jk} is an element of R. The value of U is then estimated by 1-c_{ii}.

Missing Values and Robust Estimation

Missing values and robust estimation are done the same way as in principal components analysis. Refer to that chapter for details.

How Many Factors

Several methods have been proposed for determining the number of factors that should be kept for further analysis. Several of these methods will now be discussed. However, remember that important information about possible outliers and linear dependencies may be determined from the factors associated with the relatively small eigenvalues, so these should be investigated as well.

Kaiser (1960) proposed dropping factors whose eigenvalues are less than one since these provide less information than is provided by a single variable. Jolliffe (1972) feels that Kaiser's criterion is too large. He suggests using a cutoff on the eigenvalues of 0.7 when correlation matrices are analyzed. Other authors note that if the largest eigenvalue is close to one, then holding to a cutoff of one may cause useful factors to be dropped. However, if the largest factors are several times larger than one, then those near one may be reasonably dropped.

Cattell (1966) documented the *scree graph*, which will be described later in this chapter. Studying this chart is probably the most popular method for determining the number of factors, but it is subjective, causing different people to analyze the same data with different results.

Another criterion is to preset a certain percentage of the variation that must be accounted for and then keep enough factors so that this variation is achieved. Usually, however, this cutoff percentage is used as a lower limit. That is, if the designated number of factors do not account for at least 50% of the variance, then the whole analysis is aborted.

Varimax and Quartimax Rotation

Factor analysis finds a set of dimensions (or coordinates) in a subspace of the space defined by the set of variables. These coordinates are represented as axes. They are orthogonal (perpendicular) to one another. For example, suppose you analyze three variables that are represented in three-dimensional space. Each variable becomes one axis. Now suppose that the data lie near a two-dimensional plane within the three dimensions. A factor analysis of this data should uncover two factors that would account for the two dimensions. You may rotate the axes of this two-dimensional plane while keeping the 90-degree angle between them, just as the blades of a helicopter propeller rotate yet maintain the same angles among themselves. The hope is that rotating the axes will improve your ability to interpret the “meaning” of each factor.

Many different types of rotation have been suggested. Most of them were developed for use in factor analysis. **NCSS** provides two orthogonal rotation options: varimax and quartimax.

Varimax Rotation

Varimax rotation is the most popular orthogonal rotation technique. In this technique, the axes are rotated to maximize the sum of the variances of the squared loadings within each column of the loadings matrix. Maximizing according to this criterion forces the loadings to be either large or small. The hope is that by rotating the factors, you will obtain new factors that are each highly correlated with only a few of the original variables. This simplifies the interpretation of the factor to a consideration of these two or three variables. Another way of stating the goal of varimax rotation is that it clusters the variables into groups; each “group” is actually a new factor.

Since varimax seeks to maximize a specific criterion, it produces a unique solution (except for differences in sign). This has added to its popularity. Let the matrix $G = \{g_{ij}\}$ represent the rotated factors. The goal of varimax rotation is to maximize the quantity:

$$Q_1 = \sum_{j=1}^{k} \left(\frac{p \sum_{i=1}^{p} g_{ij}^4 - \sum_{i=1}^{p} g_{ij}^2}{p} \right)$$

This equation gives the raw varimax rotation. This rotation has the disadvantage of not spreading the variance very evenly among the new factors. Instead, it tends to form one large factor followed by many small ones. To correct this, **NCSS** uses the normalized-varimax rotation. The quantity maximized in this case is:

$$Q_N = \sum_{j=1}^{k} \left[\frac{p \sum_{i=1}^{p} \left(\frac{g_{ij}}{c_i} \right)^4 - \sum_{i=1}^{p} \left(\frac{g_{ij}}{c_i} \right)^2}{p^2} \right]$$

where c_i is the square root of the communality of variable i.

Quartimax Rotation

Quartimax rotation is similar to varimax rotation except that the rows of G are maximized rather than the columns of G. This rotation is more likely to produce a “general” factor than will varimax. Often, the results are quite similar. The quantity maximized for the quartimax is:

$$Q_N = \sum_{j=1}^{k} \left[\frac{\sum_{i=1}^{p} \left(\frac{g_{ij}}{c_i} \right)^4}{p} \right]$$

Miscellaneous Topics

Using Correlation Matrices Directly

Occasionally, you will be provided with only the correlation matrix from a previous analysis. This happens frequently when you want to analyze data that is presented in a book or a report. You can perform a factor analysis on a correlation matrix using **NCSS**.

NCSS can store the correlation matrix on the current database. If it takes a great deal of computer time to build the correlation matrix, you might want to save it so you can use it while you determine the number of factors. You could then return to the original data to analyze the factor scores.

Principal Component Analysis versus Factor Analysis

Both principal component analysis (PCA) and factor analysis (FA) seek to reduce the dimensionality of a data set. The most obvious difference is that while PCA is concerned with the total variation as expressed in the correlation matrix, R, FA is concerned with a correlation in a partition of the total variation called the common portion. That is, FA separates R into two matrices R_c (common factor portion) and R_u (unique factor portion). FA models the R_c portion of the correlation matrix. Hence, FA requires the discovery of R_c as well as a model for it. The goals of FA are more concerned with finding and interpreting the underlying, common factors. The goals of PCA are concerned with a direct reduction in the dimensionality.

Put another way, PCA is directed towards reducing the diagonal elements of R. Factor analysis is directed more towards reducing the off-diagonal elements of R. Since reducing the diagonal elements reduces the off-diagonal elements and vice versa, both methods achieve much the same thing.

Data Structure

The data for a factor analysis consists of two or more variables. We have created an artificial data set in which each of the six variables (X1 - X6) were created using weighted averages of two original variables (V1 and V2) plus a small random error. For example, X1 = .33 V1 + .65 V2 + error. Each variable had a different set of weights (.33 and .65 are the weights) in the weighted average.

Rows two and three of the data set were modified to be outliers so that their influence on the analysis could be observed. Note that even though these two rows are outliers, their values on each of the individual variables are not outliers. This shows one of the challenges of multivariate analysis: multivariate outliers are not necessarily univariate outliers. In other words, a point may be an outlier in a multivariate space and yet you cannot detect it by scanning the data one variable at a time.

This data set is contained in the database PCA2. The data given below are the first few rows of this data set.

PCA2 dataset (subset)

X1	X2	X3	X4	X5	X6
50	102	103	70	75	102
4	2	5	11	11	5
81	98	94	5	85	97
31	81	86	46	50	74
65	50	51	60	57	53
22	30	39	17	15	17
36	33	39	29	27	25
31	91	96	50	56	85

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Input Variables

Variables

Designates the variables to be analyzed. If matrix input is selected, indicate the variables containing the matrix. Note that for matrix input, the number of rows used is equal to the number of variables specified. Other rows will be ignored.

Data Input Format

Indicates whether raw data is to be analyzed or if a previously summarized correlation or covariance matrix is to be used.

- **Regular Data**
  The data is to be input in its raw format.

- **Lower-Triangular**
  The data is in a correlation or covariance matrix in lower-triangular format. This matrix could have been created by a previous run of an **NCSS** program or from direct keyboard input.

- **Upper-Triangular**
  The data is in a correlation or covariance matrix in upper triangular format. The number of rows used is equal to the number of variables specified. This matrix could have been created by a previous run of an **NCSS** program or from direct keyboard input.

Covariance Estimation Options

Robust Covariance Matrix Estimation

This option indicates whether robust estimation is to be used. A full discussion of robust estimation is provided in the PCA chapter. If checked, robust estimates of the means, variances, and covariances are formed.

Robust Weight

This option specifies the value of ν1 for robust estimation. This parameter controls the weighting function in robust estimation of the covariance matrix. Jackson (1991) recommends the value 4.

Missing Value Estimation

This option indicates the type of missing value imputation method that you want to use. (Note that if the number of iterations is zero, this option is ignored.)

- **None**

  No missing value imputation. Rows with missing values in any of the selected variables are ignored.

- **Average**

  The average-value imputation method is used. Each missing value is estimated by the average value of that variable. The process is iterated as many times as is indicated in the second box.

- **Multivariate Normal**

  The multivariate-normal method. Each missing value is estimated using a multiple regression of the missing variable(s) on the variables that contain data in that row. This process is iterated as many times as indicated. See the discussion of missing value imputation methods elsewhere in this chapter.

Maximum Iterations

This option specifies the number of iterations used by either Missing Value Imputation or Robust Covariance Estimation. Robust estimation usually requires only four or five iterations to converge. Missing value imputation may require as many as twenty iterations if there are a lot of missing values.

When using this option, it is better to specify too many iterations than too few. After considering the Percent Change values in the Iteration Report, you can decide upon an appropriate number of iterations and re-run the problem.

Factor Options

Factor Rotation

Specifies the type of rotation, if any, that should be used on the solution. If rotation is desired, either varimax or quartimax rotation is available.

Number of Factors

This option specifies the number of factors to be used. On the first run, you would set this rather large (say eight or so). After viewing the eigenvalues you would reset this appropriately and make a second run.

Communality Options

Communality Iterations

This option specifies how many iterations to use in estimating the communalities. Some authors suggest a value of one here. Others suggest as many as four or five.

Reports Tab

The following options control the format of the reports.

Select Reports

Descriptive Statistics - Scores Report

These options let you specify which reports are displayed.

Select Plots

Scores Plot - Loadings Plot

These options let you specify which reports and plots are displayed.

Row Numbers

Indicates whether to display row numbers on the individual points in the corresponding plot.

Report Options

Minimum Loading

Specifies the minimum absolute value that a loading can have and still remain in the Variable List report.

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.

Plot Options

Number Factors Plotted

You can limit the number of plots generated using this parameter. Usually, you will only have interest in the first three or four factors.

Scores Plot Tab and Loadings Plot Tab

These sections specify the pair-wise plots of the scores and loadings.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis 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.

Storage Tab

The factor scores and/or the correlation matrix may be stored on the current database for further analysis. This group of options let 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. Note that existing data are replaced.

Data Storage Variables

Factor Scores

You can automatically store the factor scores for each row into the variables specified here. These scores are generated for each row of data in which all independent variable values are nonmissing.

Correlation Matrix

You can automatically store the correlation matrix to 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 – Factor Analysis

This section presents an example of how to run a factor analysis. The data used are shown in the table above and found in the PCA2 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Factor Analysis window.

1 Open the PCA2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **PCA2.s0**.
- Click **Open**.

2 Open the Factor Analysis window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Factor Analysis**. The Factor Analysis 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 Factor Analysis window, select the **Variables tab**.
- Double-click in the **Variables** box. This will bring up the variable selection window.
- Select **X1** through **X6** from the list of variables and then click **Ok**. “X1-X6” will appear in the Variables box.
- Select **Varimax** in the **Factor Rotation** list box.
- Enter **2** in the **Number of Factors** box.
- Enter **6** in the **Communality Iterations** box.
- Select **Robust** in the **Covariance Estimation** list box.
- Enter **6** in the **Maximum Iterations** box.

4 Specify which reports.

- Select the **Reports tab**.
- Check all reports and plots. Normally you would only view a few of these reports, but we are selecting them all 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).

Robust and Missing-Value Iteration Section

This report is only produced when robust or missing value estimation is used.

```
Robust and Missing-Value Estimation Iteration Section
                        Trace of             Percent
No.     Count           Covar Matrix         Change
0       30              4907.795               0.00
1       30              4907.795               0.00
2       30              4423.718              -9.86
3       30              4423.718               0.00
4       30              4353.748              -1.58
5       30              4353.748               0.00
6       30              4335.77               -0.41
```

This report presents the progress of the robust iterations. The trace of the covariance matrix gives a measure of what is happening at each iteration. When this value stabilizes, the program has converged. The percent change is reported to let you determine how much the trace has changed. In this particular example, we see very little change between iterations five and six. We would feel comfortable stopping at this point. A look at the Descriptive Statistics section will let you see how much the means and standard deviations have changed.

A look at the Residual Section will let you see the robust weights that are assigned to each row. Those weights that are near zero indicate observations whose influence have been removed by the robust procedure.

Descriptive Statistics Section

```
Descriptive Statistics Section
                                              Standard
Variables       Count       Mean              Deviation          Communality
X1              30          42.83667          23.18579           0.997983
X2              30          53.25062          27.93123           0.999791
X3              30          57.13034          26.3737            0.999585
X4              30          43.5617           24.56474           0.992023
X5              30          43.20835          25.75021           1.00007
X6              30          48.61827          32.49559           0.999944
```

Count, Mean, and Standard Deviation

These are the familiar summary statistics of each variable. They are displayed to allow you to make sure that you have specified the correct variables. Note that using missing value imputation or robust estimation will change these values.

Communality

The communality shows how well this variable is predicted by the retained factors. It is similar to the R-Squared that would be obtained if this variable were regressed on the factors that were kept. However, remember that this is not based directly on the correlation matrix. Instead, calculations are based on an adjusted correlation matrix.

Correlation Section

```
Correlation Section
                  Variables
Variables            X1          X2          X3          X4          X5
X1             1.000000    0.271780    0.127016    0.881604    0.814686
X2             0.271780    1.000000    0.988909    0.683206    0.778649
X3             0.127016    0.988909    1.000000    0.568933    0.677480
X4             0.881604    0.683206    0.568933    1.000000    0.986945
X5             0.814686    0.778649    0.677480    0.986945    1.000000
X6             0.484907    0.973093    0.928454    0.831949    0.901975
Phi=0.769781  Log(Det|R|)=-29.547320  Bartlett Test=773.15  DF=15  Prob=0.000000

                  Variables
Variables            X6
X1             0.484907
X2             0.973093
X3             0.928454
X4             0.831949
X5             0.901975
X6             1.000000
Phi=0.769781  Log(Det|R|)=-29.547320  Bartlett Test=773.15  DF=15  Prob=0.000000

Bar Chart of Absolute Correlation Section
                       Variables
Variables      X1                    X2                    X3                    X4                    X5
X1                                   ||||||                |||                   ||||||||||||||||||    |||||||||||||||||
X2             ||||||                                      ||||||||||||||||||||  ||||||||||||||        ||||||||||||||||
X3             |||                   ||||||||||||||||||||                        ||||||||||||          ||||||||||||||
X4             ||||||||||||||||||    ||||||||||||||        ||||||||||||                                ||||||||||||||||||||
X5             |||||||||||||||||     ||||||||||||||||      ||||||||||||||        ||||||||||||||||||||
X6             ||||||||||            ||||||||||||||||||||  |||||||||||||||||||   |||||||||||||||||     |||||||||||||||||||
Phi=0.769781  Log(Det|R|)=-29.547320  Bartlett Test=773.15  DF=15  Prob=0.000000

                       Variables
Variables      X6
X1             ||||||||||
X2             ||||||||||||||||||||
X3             |||||||||||||||||||
X4             |||||||||||||||||
X5             |||||||||||||||||||
X6
Phi=0.769781  Log(Det|R|)=-29.547320  Bartlett Test=773.15  DF=15  Prob=0.000000
```

This report gives the correlations alone for a test of the overall correlation structure in the data. In this example, we notice several high correlation values. The Gleason-Staelin redundancy measure, phi, is 0.736, which is quite large. There is apparently some correlation structure in this data set that can be modeled. If all the correlations are small (say less then .3), there would be no need for a factor analysis.

Correlations

The simple correlations between each pair of variables. Note that using the missing value imputation or robust estimation options will affect the correlations in this report. When the above options are not used, the correlations are constructed from those observations having no missing values in any of the specified variables.

Phi

This is the Gleason-Staelin redundancy measure of how interrelated the variables are. A zero value of φ2 means that there is no correlation among the variables, while a value of one indicates perfect correlation among the variables. This coefficient may have a value less than 0.5 even when there is obvious structure in the data, so care should to be taken when using it. This statistic is especially useful for comparing two or more sets of data. The formula for computing φ3 is:

$$\varphi = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=1}^{p} r_{ij}^{2} - p}{p(p-1)}}$$

Log(Det|R|)

This is the log (base e) of the determinant of the correlation matrix. If you used the covariance matrix, this is the log (base e) of the determinant of the covariance matrix.

Bartlett Test, df, Prob

This is Bartlett’s sphericity test (Bartlett, 1950) for testing the null hypothesis that the correlation matrix is an identity matrix (all correlations are zero). If you get a probability (Prob) value greater than 0.05, you should not perform a factor analysis on the data. The test is valid for large samples (N>150). It uses a Chi-square distribution with p(p-1)/2 degrees of freedom. Note that this test is only available when you analyze a correlation matrix. The formula for computing this test is:

$$\chi^{2} = \frac{(11 + 2p - 6N)}{6} \mathrm{Log}_{e} |R|$$

Bar Chart of Absolute Correlation Section

This chart graphically displays the absolute values of the correlations. It lets you quickly find high and low correlations.

Eigenvalues Section

```
Eigenvalues after Varimax Rotation
                              Individual     Cumulative
No.        Eigenvalue         Percent        Percent        Scree Plot
1            3.288191           54.89          54.89        |||||||||||
2            2.701207           45.09          99.99        ||||||||||
3            0.001207            0.02         100.01        |
4           -0.000099            0.00         100.01        |
5           -0.000121            0.00         100.00        |
6           -0.000295            0.00         100.00        |
```

Eigenvalues

The eigenvalues of the *R-U* matrix. Often, these are used to determine how many factors to retain. (In this example, we would retain the first two eigenvalues.)

One rule-of-thumb is to retain those factors whose eigenvalues are greater than one. The sum of the eigenvalues is equal to the number of variables. Hence, in this example, the first factor retains the information contained in 3.3 of the original variables.

Note that, unlike in PCA where all eigenvalues are positive, the eigenvalues may be negative in factor analysis. Usually, these factors would be discarded and the analysis would be re-run.

Individual and Cumulative Percents

The first column gives the percentage of the total variation in the variables accounted for by this factor. The second column is the cumulative total of the percentage. Some authors suggest that the user pick a cumulative percentage, such as 80% or 90%, and keep enough factors to attain this percentage.

Scree Plot

This is a rough bar plot of the eigenvalues. It enables you to quickly note the relative size of each eigenvalue. Many authors recommend it as a method of determining how many factors to retain.

The word *scree*, first used by Cattell (1966), is usually defined as the rubble at the bottom of a cliff. When using the scree plot, you must determine which eigenvalues form the “cliff” and which form the “rubble.” You keep the factors that make up the cliff. Cattell and Jaspers (1967) suggest keeping those that make up the cliff plus the first factor of the rubble.

Interpretation of the Example

The first question that we would ask is how many factors should be kept. The scree plot shows that the first two factors are indeed the largest. The cumulative percentages show that the first two factors account for over 99.99% of the variation.

Eigenvectors Section

```
Eigenvectors after Varimax Rotation
                        Factors
Variables               Factor1        Factor2
X1                    -0.303444      -0.662220
X2                    -0.416551       0.378018
X3                    -0.382768       0.491154
X4                    -0.428167      -0.317929
X5                    -0.448606      -0.204840
X6                    -0.450912       0.185189

Bar Chart of Absolute Eigenvectors after Varimax Rotation
                      Factors
Variables             Factor1        Factor2
X1                    |||||||        ||||||||||||||
X2                    |||||||||      ||||||||
X3                    ||||||||       ||||||||||
X4                    |||||||||      |||||||
X5                    |||||||||      |||||
X6                    ||||||||||     ||||
```

Eigenvector

The eigenvectors of the R-U matrix.

Bar Chart of Absolute Eigenvectors

This chart graphically displays the absolute values of the eigenvectors. It lets you quickly interpret the eigenvector structure. By looking at which variables correlate highly with a factor, you can determine what underlying structure it might represent.

Factor Loadings Section

```
Factor Loadings after Varimax Rotation
                        Factors
Variables               Factor1          Factor2
X1                     -0.019936        -0.998792
X2                     -0.967470        -0.252572
X3                     -0.994037        -0.107126
X4                     -0.478418        -0.873578
X5                     -0.594943        -0.803812
X6                     -0.883654        -0.468080

Bar Chart of Absolute Factor Loadings after Varimax Rotation
                       Factors
Variables              Factor1              Factor2
X1                     |                    ||||||||||||||||||||
X2                     ||||||||||||||||||||  ||||||
X3                     ||||||||||||||||||||  |||
X4                     ||||||||||           ||||||||||||||||||
X5                     ||||||||||||         |||||||||||||||||
X6                     ||||||||||||||||||   ||||||||||
```

Factor Loadings

These are the correlations between the variables and factors.

Bar Chart of Absolute Factor Loadings

This chart graphically displays the absolute values of the factor loadings. It lets you quickly interpret the correlation structure. By looking at which variables correlate highly with a factor, you can determine what underlying structure it might represent.

Communality Section

```
Communality after Varimax Rotation
                       Factors
Variables              Factor1          Factor2          Communality
X1                     0.000397         0.997586         0.997983
X2                     0.935999         0.063793         0.999791
X3                     0.988109         0.011476         0.999585
X4                     0.228883         0.763139         0.992023
X5                     0.353958         0.646114         1.000072
X6                     0.780845         0.219099         0.999944

Bar Chart of Communalities after Varimax Rotation
                       Factors
Variables              Factor1              Factor2              Communality
X1                     |                    ||||||||||||||||||||  ||||||||||||||||||||
X2                     ||||||||||||||||||   ||                   ||||||||||||||||||||
X3                     ||||||||||||||||||||  |                    ||||||||||||||||||||
X4                     |||||                ||||||||||||||||     ||||||||||||||||||||
X5                     ||||||||             |||||||||||||        ||||||||||||||||||||
X6                     ||||||||||||||||     |||||                ||||||||||||||||||||
```

Communality

The communality is the proportion of the variation of a variable that is accounted for by the factors that are retained. It is similar to the R-Squared value that would be achieved if this variable were regressed on the retained factors. This table value gives the amount added to the communality by each factor.

Bar Chart of Communalities

This chart graphically displays the values of the communalities.

Factor Structure Summary Section

```
Factor Structure Summary after Varimax Rotation
                     Factors
Factor1              Factor2
X3                   X1
X2                   X4
X6                   X5
X5                   X6
X4
```

Interpretation

This report is provided to summarize the factor structure. Variables with an absolute loading greater than the amount set in the Minimum Loading option are listed under each factor. Using this report, you can quickly see which variables are related to each factor. Note that it is possible for a variable to have high loadings on several factors, although varimax rotation makes this very unlikely.

Score Coefficients Section

```
Score Coefficients after Varimax Rotation
                     Factors
Variables            Factor1              Factor2
X1                   0.268188             -0.5366089
X2                   -0.3613275           0.1312937
X3                   -0.4135219           0.2176106
X4                   2.901571E-02         -0.3414549
X5                   -0.0422971           -0.2712604
X6                   -0.2641693           -8.934696E-03
```

Score Coefficients

These are the coefficients that are used to form the factor scores. The factor scores are the values of the factors for a particular row of data. These score coefficients are similar to the eigenvectors. They have been scaled so that the scores produced have a variance of one rather than a variance equal to the eigenvalue. This causes each of the factors to have the same variance.

You would use these scores if you wanted to calculate the factor scores for new rows not included in your original analysis.

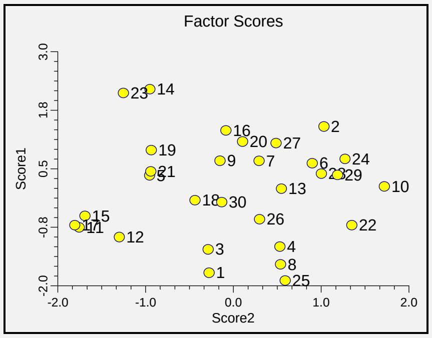
Factor Scores Section

```
Factor Scores after Varimax Rotation
                      Factors
Row                   Factor1       Factor2
1                     -1.7219       -0.2752
2                      1.4002        1.0317
3                     -1.2231       -0.2862
4                     -1.1632        0.5302
(report continues through all thirty rows)
```

Factor1 - Factor2

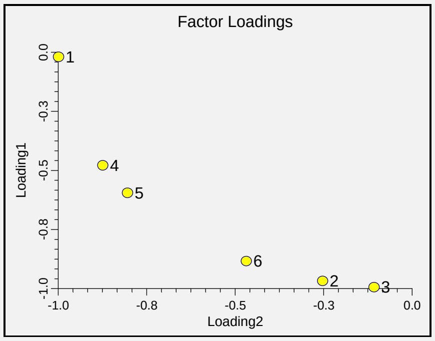
The factor scores are the values of the factors for a particular row of data. They have been scaled so they have a variance of one.

Factor Score Plots



This set of plots shows each factor plotted against every other factor.

Factor Loading Plots



This set of plots shows each of the factor loading columns plotted against each other.

Chapter 425

Principal Components Analysis

Introduction

Principal Components Analysis, or *PCA*, is a data analysis tool that is usually used to reduce the dimensionality (number of variables) of a large number of interrelated variables, while retaining as much of the information (variation) as possible. PCA calculates an uncorrelated set of variables (*factors* or *pc's*). These factors are ordered so that the first few retain most of the variation present in all of the original variables. Unlike its cousin Factor Analysis, PCA always yields the same solution from the same data (apart from arbitrary differences in the sign).

The computations of PCA reduce to an eigenvalue-eigenvector problem. **NCSS** uses a double-precision version of the modern QL algorithm as described by Press (1986) to solve the eigenvalue-eigenvector problem.

Note that PCA is a data analytical, rather than statistical, procedure. Hence, you will not find many t-tests or F-tests in PCA. Instead, you will make subjective judgments requiring you to spend a little extra time getting acquainted with the technique.

This **NCSS** program performs a PCA on either a correlation or a covariance matrix. Missing values may be dealt with using one of three methods. The analysis may be carried out using robust estimation techniques.

Chapters on PCA are contained in books dealing with multivariate statistical analysis. Books that are devoted solely to PCA include Dunteman (1989), Jolliffe (1986), Flury (1988), and Jackson (1991).

Technical Details

Mathematical Development

This section will document the basic formulas used by **NCSS** in performing a principal components analysis. We begin with an adjusted data matrix, *X*, which consists of *n* observations (rows) on *p* variables (columns). The adjustment is made by subtracting the variable's mean from each value. That is, the mean of each variable is subtracted from all of that variable's values. This adjustment is made since PCA deals with the covariances among the original variables, so the means are irrelevant.

New variables are constructed as weighted averages of the original variables. These new variables are called the factors, latent variables, or principal components. Their specific values on a specific row are referred to as the factor scores, the component scores, or simply the scores. The matrix of scores will be referred to as the matrix *Y*. The basic equation of PCA is, in matrix notation, given by:

$$Y = W'X$$

where *W* is a matrix of coefficients that is determined by PCA. This matrix is provided in **NCSS** in the *Score Coefficients* report. For those not familiar with matrix notation, this equation may be thought of as a set of *p* linear equations that form the factors out of the original variables. These equations are also written as:

$$y_{ij} = w_{1i} x_{1j} + w_{2i} x_{2j} + \ldots + w_{pi} x_{pj}$$

As you can see, the factors are a weighted average of the original variables. The weights, *W*, are constructed so that the variance of y_1, $Var(y_1)$, is maximized. Also, so that $Var(y_2)$ is maximized and that the correlation between y_1 and y_2 is zero. The remaining y_i's are calculated so that their variances are maximized, subject to the constraint that the covariance between y_i and y_j, for all *i* and *j* (*i* not equal to *j*), is zero.

The matrix of weights, *W*, is calculated from the variance-covariance matrix, *S*. This matrix is calculated using the formula:

$$s_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ik} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_j\right)}{n-1}$$

Later, we will discuss how this equation may be modified both to be robust to outliers and to deal with missing values.

The singular value decomposition of *S* provides the solution to the PCA problem. This may be defined as:

$$U'SU = L$$

where *L* is a diagonal matrix of the eigenvalues of *S*, and *U* is the matrix of eigenvectors of *S*. *W* is calculated from *L* and *U*, using the relationship:

$$W = UL^{-\frac{1}{2}}$$

It is interesting to note that *W* is simply the eigenvector matrix *U*, scaled so that the variance of each factor, y_i, is one.

The correlation between an i^{th} factor and the j^{th} original variable may be computed using the formula:

$$r_{ij} = \frac{u_{ji}\sqrt{l_i}}{s_{jj}}$$

Here u_{ij} is an element of *U*, l_i is a diagonal element of *L*, and s_{jj} is a diagonal element of *S*. The correlations are called the factor loadings and are provided in the *Factor Loadings* report.

When the correlation matrix, *R*, is used instead of the covariance matrix, *S*, the equation for *Y* must be modified. The new equation is:

$$Y = W' D^{-\frac{1}{2}} X$$

where D is a diagonal matrix made up of the diagonal elements of S. In this case, the correlation formula may be simplified since the s_{jj} are equal to one.

Missing Values

Missing values may be dealt with by ignoring rows with missing values, estimating the missing value with the variable's average, or estimating the missing value by regressing it on variables whose values are not missing. These will now be described in detail. Most of this information comes from Jackson (1991) and Little (1987).

When estimating statistics from data sets with missing values, you should first consider the mechanism that created the missing values. This mechanism determines whether your method of dealing with the missing values is appropriate. The worst case arises when the probability of obtaining a missing value is dependent on one or more variables in your study. For example, suppose one of your variables was a person's income level. You might suspect that the higher a person's income, the less likely he is to reveal it to you. When the probability of obtaining a missing value is dependent on one or more variables, serious biases can occur in your results. A complete discussion of missing value mechanisms is given in Little (1987).

NCSS provides three methods of dealing with missing values. In all three cases, the overall strategy is to deal with the missing values while estimating the covariance matrix, S. Hence, the rest of the section will consider estimating S.

Complete-Case Missing-Value Analysis

One method of dealing with missing values is to remove all cases (observations or rows) that contain missing values from the analysis. The analysis is then performed only on those cases that are "complete."

The advantages of this approach are *speed* (since no iteration is required), *comparability* (since univariate statistics, such as the mean, calculated on individual variables, will be equal to the results of the multivariate calculations), and *simplicity* (since the method is easy to explain).

Disadvantages of this approach are *inefficiency* and *bias*. This method is inefficient since as the number of missing values increases, the number of discarded cases also increases. In the extreme case, suppose a data set has 100 variables and 200 cases. Suppose one value is missing at random in 80 cases, so these cases are deleted from the study. Hence, of the 20,000 values in the study, 80 values or 0.4% were missing. Yet this method has us omit 8000 values or 40%, even though 7920 of those values were actually available. This is similar to the saying that one rotten apple ruins the whole barrel.

A certain amount of bias may occur if the pattern of missing values is related to at least one of the variables in the study. This could lead to gross distortions if this variable were correlated with several other variables.

One method of determining if the complete-case methodology is causing bias is to compare the means of each variable calculated from only complete cases, with the corresponding means of each variable calculated from cases that were dropped but had this variable present. This comparison could be run using a statistic like the t-test, although we would also be interested in comparing the variances, which the t-test does not do. Significant differences would indicate the presence of a strong bias introduced by the pattern of missing values.

A modification of the complete-case method is the pairwise available-case method in which covariances are calculated one at a time from all cases that are complete for those two variables. This method is not available in this program for three reasons: the univariate statistics change from pair to pair causing serious numeric problems (such as correlations greater than one), the resulting covariance matrix may not be positive semi-definite, and the method is dominated by other methods that are available in this program.

Filling in Missing Values with Averages

A growing number of programs offer the ability to fill in (or impute) the missing values. The naive choice is to fill in with the variable average. **NCSS** offers this option, implemented iteratively. During the first iteration, no imputation occurs. On the second, third, and additional iterations, each missing value is estimated using the mean of that variable from the previous iteration. Hence, at the end of each iteration, a new set of means is available for imputation during the next iteration. The process continues until it converges.

The advantages of this method are greater efficiency (since it takes advantage of the cases in which missing values occur) and speed (since it is much faster than the EM algorithm to be presented next).

The disadvantages of this method are biases (since it consistently underestimates the variances and covariances), unreliability (since simulation studies have shown it unreliable in some cases), and domination (since it is dominated by the EM algorithm, which does much better although that method requires more computations).

Multivariate-Normal Missing-Value Imputation

Little (1987) has documented the use of the EM algorithm for estimating the covariance matrix, *S*, when the data follow the multivariate normal distribution. This might also be referred to as a regression approach or modified conditional means approach. The assumption of a multivariate normal distribution may seem limiting, but the procedure produces estimates that are consistent under weaker assumptions. We will now define the algorithm for you.

1. Estimate the covariance matrix, *S*, with the complete-case method.
2. The E step consists of calculating the sums and sums of squares using the following formulas:

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} x_{ij}^{(t)}}{n}$$

$$s_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n}\left[\left(x_{ij}^{(t)} - \hat{\mu}_j^{(t+1)}\right)\left(x_{ik}^{(t)} - \hat{\mu}_k^{(t+1)}\right) + c_{jki}^{(t)}\right]}{n-1}$$

$$x_{ij}^{(t)} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed} \\ E\left(x_{ij} \mid x_{obs,i}, \hat{\mu}, S^{(t)}\right), & \text{if } x_{ij} \text{ is missing} \end{cases}$$

$$c_{jki}^{(t)} = \begin{cases} 0 & \text{if } x_{ij} \text{ or } x_{ik} \text{ are observed} \\ Cov(x_{ij}, x_{ik} \mid x_{obs,i}, S^{(t)}) & \text{if } x_{ij} \text{ and } x_{ik} \text{ are missing} \end{cases}$$

where $x_{obs,i}$ refers to that part of observation *i* that is not missing and $E\left(x_{ij}|x_{obs,i},\hat{\mu},S^{(t)}\right)$ refers to the regression of the variables that are missing on the variables that are observed. This regression is calculated by sweeping *S* by the variables that are observed and using the observed values as the values of the independent variables in the resulting regression equation. Essentially, we are fitting a multiple regression of each missing value on the values that are observed, using the *S(t)* matrix as our matrix of sums of squares and cross products. When both x_{ij} and x_{ik} are missing, the value of c_{jki} is the ij^{th} element of the swept *S* matrix.

Verbally, the algorithm may be stated as follows. Each missing data value is estimated by regressing it on the values that are observed. The regression coefficients are calculated from the current covariance matrix. Since this regression tends to underestimate the true covariance values, these are inflated by an appropriate amount. Once each missing value is estimated, a new covariance matrix is calculated and the process is repeated. The procedure is terminated when it converges. This convergence is measured by the trace of the covariance matrix.

NCSS first sorts the data according to the various patterns of missing values, so that the regression calculations (the sweeping of *S*) are performed a minimum number of times: once for each particular missing-value pattern.

This method has the disadvantage that it is computationally intensive and it may take twenty or more iterations to converge. However, it provides the maximum-likelihood estimate of the covariance matrix, it provides a positive semi-definite covariance matrix, and it seems to do well even when the occurrences of missing values are correlated with the values of the variables being studied. That is, it corrects for biases caused by the pattern of missing values.

Robust Estimation

Robust estimation refers to estimation techniques that decrease or completely remove the influence of observations that are outliers. These outliers can seriously distort the estimated means and covariances. The EM algorithm is employed as the robust technique used in **NCSS**. This algorithm uses weights that are inversely proportional to how “outlying” the observation is. The usual estimates of the means and covariances are modified to use these weights. The process is iterated until it converges. Note that since *S* is estimated robustly, the estimated correlation matrix is robust also.

One advantage of the EM algorithm is that it can be modified to deal with missing values and robust estimation at the same time. Hence, **NCSS** provides robust estimates that use the information in rows with missing values as well. The robust estimation formulas are:

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^{(t)} x_{ij}^{(t)}}{\sum_{i=1}^{n} w_i^{(t)}}$$

$$s_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n}\left[w_i^{(t)}\left(x_{ij}^{(t)} - \hat{\mu}_j^{(t+1)}\right)\left(x_{ik}^{(t)} - \hat{\mu}_k^{(t+1)}\right) + c_{jki}^{(t)}\right]}{n-1}$$

The weights, w_i, are calculated using the formula:

$$w_i = \frac{(\nu + p_i)}{(\nu + d_i^2)}$$

where ν is a parameter you supply, p_i is the number of nonmissing values in the i^{th} row, and

$$d_i^2 = \sum_{j=1}^{p} \sum_{k=1}^{p} \delta_{ijk} \left(x_{ij} - \hat{\mu}_j \right) \left(x_{ik} - \hat{\mu}_k \right) b^{jk}$$

where δ_{ijk} is equal to one if both variables x_j and x_k are observed in row i and is zero otherwise. The b^{jk} are the indicated elements of the inverse of S ($B = S^{-1}$). Note that B is found by sweeping S on all variables.

When using robust estimation, it is wise to run the analysis with the robust option turned on and then study the robust weights. When the weight is less than .4 or .3, the observation is being "removed." You should study rows that have such a weight to determine if there was an error in data entry or measurement, or if the values are valid. If the values are all valid, you have to decide whether this row should be kept or discarded. Next, make a second run without the discarded rows and without using the robust option. In this way, your results do not depend quite so much on the particular formula that was used to create the weights. Note that the weights are listed in the *Residual Report* after the values of Q_k and T^2.

How Many Factors

Several methods have been proposed for determining the number of factors that should be kept for further analysis. Several of these methods will now be discussed. However, remember that important information about possible outliers and linear dependencies may be determined from the factors associated with the relatively small eigenvalues, so these should be investigated as well.

Kaiser (1960) proposed dropping factors whose eigenvalues are less than one, since these provide less information than is provided by a single variable. Jolliffe (1972) feels that Kaiser's criterion is too large. He suggests using a cutoff on the eigenvalues of 0.7 when correlation matrices are analyzed. Other authors note that if the largest eigenvalue is close to one, then holding to a cutoff of one may cause useful factors to be dropped. However, if the largest factors are several times larger than one, then those near one may be reasonably dropped.

Cattell (1966) documented the *scree graph*, which will be described later in this chapter. Studying this chart is probably the most popular method for determining the number of factors, but it is subjective, causing different people to analyze the same data with different results.

Another criterion is to preset a certain percentage of the variation that must be accounted for and then keep enough factors so that this variation is achieved. Usually, however, this cutoff percentage is used as a lower limit. That is, if the designated number of factors do not account for at least 50% of the variance, then the whole analysis is aborted.

We cannot give a definitive answer as to which criterion is best, since most of these techniques were developed for use in factor analysis, not PCA. Perhaps the best advise we can give is to use the number of factors that agrees with the goals of your analysis. If you want to look for outliers in multivariate data, then you will want to keep most, if not all, factors during the early stages of the analysis. If you want to reduce the dimensionality of your database, then you should keep enough factors so that you account for a reasonably large percentage of the variation.

Varimax and Quartimax Rotation

PCA finds a set of dimensions (or coordinates) in a subspace of the space defined by the set of variables. These coordinates are represented as axes. They are orthogonal (perpendicular) to one another. For example, suppose you analyze three variables that are represented in three-dimensional space. Each variable becomes one axis. Now suppose that the data lie near a two-dimensional plane within the three dimensions. A PCA of this data should uncover two factors that would account for the two dimensions. You may rotate the axes of this two-dimensional plane while keeping the 90-degree angle between them, just as the blades of a helicopter propeller rotate yet maintain the same angles among themselves. The hope is that rotating the axes will improve your ability to interpret the meaning of each component.

Many different types of rotation have been suggested. Most of them were developed for use in factor analysis. **NCSS** provides two orthogonal rotation options: varimax and quartimax.

Varimax Rotation

Varimax rotation is the most popular orthogonal rotation technique. In this technique, the axes are rotated to maximize the sum of the variances of the squared loadings within each column of the loadings matrix. Maximizing according to this criterion forces the loadings to be either large or small. The hope is that by rotating the factors, you will obtain new factors that are each highly correlated with only a few of the original variables. This simplifies the interpretation of the factor to a consideration of these two or three variables. Another way of stating the goal of varimax rotation is that it clusters the variables into groups, where each group is actually a new factor.

Since varimax seeks to maximize a specific criterion, it produces a unique solution (except for differences in sign). This has added to its popularity. Let the matrix $B = \{b_{ij}\}$ represent the rotated factors. The goal of varimax rotation is to maximize the quantity:

$$Q_1 = \sum_{j=1}^{k} \left(\frac{p \sum_{i=1}^{p} b_{ij}^4 - \sum_{i=1}^{p} b_{ij}^2}{p} \right)$$

This equation gives the raw varimax rotation. This rotation has the disadvantage of not spreading the variance very evenly among the new factors. Instead, it tends to form one large factor followed by many small ones. To correct this, **NCSS** uses the normalized-varimax rotation. The quantity maximized in this case is:

$$Q_N = \sum_{j=1}^{k} \left[\frac{p \sum_{i=1}^{p} \left(\frac{b_{ij}}{h_i} \right)^4 - \sum_{i=1}^{p} \left(\frac{b_{ij}}{h_i} \right)^2}{p^2} \right]$$

where h_i is the square root of the communality of variable i.

Quartimax Rotation

Quartimax rotation is similar to varimax rotation, except that the rows of *B* are maximized rather than the columns of *B*. This rotation is more likely to produce a general factor than will varimax. Often, the results are quite similar. The quantity maximized for the quartimax is:

$$Q_N = \sum_{j=1}^{k} \left[\frac{\sum_{i=1}^{p} \left(\frac{b_{ij}}{h_i} \right)^4}{p} \right]$$

Miscellaneous Topics

Using Correlation Matrices Directly

Occasionally, you will be provided with only the correlation (or covariance) matrix from a previous analysis. This happens frequently when you want to analyze data that is presented in a book or a report. You can perform a partial PCA on a correlation matrix using **NCSS**. We say partial because you cannot analyze the individual scores, the row-by-row values of the factors. These are often very useful to investigate, but they require the raw data.

NCSS can store the correlation (or covariance) matrix on the current database. If it takes a great deal of computer time to build the correlation matrix, you might want to save it so you can use it while you determine the number of factors**.** You could then return to the original data to analyze the factor scores.

Using PCA to Select a Subset of the Original Variables

There are at least two reasons why a researcher might want to select a subset of the original variables for further use. These will now be discussed.

1. In some data sets the number of original variables is too large, making interpretation and analysis difficult. Also, the cost of obtaining and managing so many variables is prohibitive.
2. When using PCA, it is often difficult to find a reasonable interpretation for all the factors that are kept. Instead of trying to interpret each factor, McCabe (1984) has suggested finding the principal variables. Suppose you start with *p* variables, run a PCA, and decide to retain *k* factors. McCabe suggests that it is often possible to find *k+2* or *k+3* of the original variables that will account for the same amount of variability as the *k* factors. The interpretation of the variables is much easier than the interpretation of the factors.

Jolliffe (1986) discusses several methods to reduce the number of variables in a data set while retaining most of the variability. Using **NCSS**, one of the most effective methods for selecting a subset of the original variables can easily be implemented. This method is outlined next.

1. Perform a PCA. Save the *k* most important factor scores onto your database for further analysis.

2. Use the Multivariate Variable Selection procedure to reduce the number of variables. This is done by using the saved factor scores as the dependent variables and the original variables as the independent variables. The variable selection process finds the best subset of the original variables that predicts the group of factor scores. Since the factor scores represent the original variables, you are actually finding the best subset of the original variables.

   You will usually have to select two or three more variables than you did factors, but you will end up with most of the information in your data set being represented by a fraction of the variables.

Principal Component versus Factor Analysis

Both PCA and factor analysis (FA) seek to reduce the dimensionality of a data set. The most obvious difference is that while PCA is concerned with the total variation as expressed in the correlation matrix, *R*, FA is concerned with a correlation in a partition of the total variation called the common portion. That is, FA separates *R* into two matrices R_c (common factor portion) and R_u (unique factor portion). FA models the R_c portion of the correlation matrix. Hence, FA requires the discovery of R_c as well as a model for it. The goals of FA are more concerned with finding and interpreting the underlying, common factors. The goals of PCA are concerned with a direct reduction in the dimensionality.

Put another way, PCA is directed towards reducing the diagonal elements of *R*. Factor analysis is directed more towards reducing the off-diagonal elements of *R*. Since reducing the diagonal elements reduces the off-diagonal elements and vice versa, both methods achieve much the same thing.

Further Reading

There are several excellent books that provide detailed discussions of PCA. We suggest you first read the inexpensive monograph by Dunteman (1989). More complete (and mathematical) accounts are given by Jackson (1991) and Jolliffe (1986). Several books on multivariate methods provide excellent introductory chapters on PCA.

Data Structure

The data for a PCA consist of two or more variables. We have created an artificial data set in which each of the six variables (X1 - X6) were created using weighted averages of two original variables (V1 and V2) plus a small random error. For example, X1 = 0.33 V1 + 0.65 V2 + error. Each variable had a different set of weights (0.33 and 0.65 are the weights) in the weighted average.

Rows two and three of the data set were modified to be outliers so that their influence on the analysis could be observed. Note that even though these two rows are outliers, their values on each of the individual variables are not outliers. This shows one of the challenges of multivariate analysis: multivariate outliers are not necessarily univariate outliers. In other words, a point may be an outlier in a multivariate space, and yet you cannot detect it by scanning the data one variable at a time.

This data set is contained in the database PCA2. The data given in the table below are the first few rows of this data set.

PCA2 dataset (subset)

X1	X2	X3	X4	X5	X6
50	102	103	70	75	102
4	2	5	11	11	5
81	98	94	5	85	97
31	81	86	46	50	74
65	50	51	60	57	53
22	30	39	17	15	17
36	33	39	29	27	25
31	91	96	50	56	85

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Input Variables

Variables

Designates the variables to be analyzed. If matrix input is selected, indicate the variables containing the matrix. Note that for matrix input, the number of rows used is equal to the number of variables specified. Other rows will be ignored.

Data Input Format

Indicates whether raw data is to be analyzed or if a previously summarized correlation or covariance matrix is to be used.

- **Regular Data**
  The data is to be input in its raw format.

- **Lower-Triangular**
  The data is in a correlation or covariance matrix in lower-triangular format. This matrix could have been created by a previous run, or from direct keyboard input.

- **Upper-Triangular**
  The data is in a correlation or covariance matrix in upper-triangular format. The number of rows used is equal to the number of variables specified. This matrix could have been created by a previous run, or from direct keyboard input.

Covariance Estimation Options

Robust Covariance Matrix Estimation

This option indicates whether robust estimation is to be used. A full discussion of robust estimation is provided at the beginning of this chapter. If checked, robust estimates of the means, variances, and covariances are formed.

Robust Weight

This option specifies the value of ν1. This parameter controls the weighting function in robust estimation of the covariance matrix. Jackson (1991) recommends the value 4.

Missing Value Estimation

This option indicates the type of missing value imputation method that you want to use. (Note that if the number of iterations is zero, this option is ignored.)

- **None**

  No missing value imputation. Rows with missing values in any of the selected variables are ignored.

- **Average**

  The average-value imputation method is used. Each missing value is estimated by the average value of that variable. The process is iterated as many times as is indicated in the second box.

- **Multivariate Normal**

  The multivariate-normal method. Each missing value is estimated using a multiple regression of the missing variable(s) on the variables that contain data in that row. This process is iterated as many times as indicated. See the discussion of missing value imputation methods elsewhere in this chapter.

Maximum Iterations

This option specifies the number of iterations used by either Missing Value Imputation or Robust Covariance Estimation. Robust estimation usually requires only four or five iterations to converge. Missing value imputation may require as many as twenty iterations if there are a lot of missing values.

When using this option, it is better to specify too many iterations than too few. After considering the Percent Change values in the Iteration Report, you can decide upon an appropriate number of iterations and re-run the problem.

Type of Matrix Used in Analysis

Matrix Type

This option indicates whether the analysis is to be run on a correlation or covariance matrix. Normally, the analysis is run on the scale-invariant correlation matrix since the scale of the variables changes the analysis when the covariance matrix is used. (For example, a variable that was measured in yards results in a different analysis than if it were measured in feet when a covariance matrix was used.)

Factor (Component) Options

Factor Rotation

Specifies the type of rotation, if any, that should be used on the solution. If rotation is desired, either varimax or quartimax rotation is available.

Factor Selection - Method

This option specifies which of the following three methods is used to set the number of factors retained in the analysis.

- **Percent of Eigenvalues**

  Specify the total percent of variation that must be accounted for. Enough factors will be included to account for this percentage (or slightly greater) of the variation in the data.

- **Number of Factors**

  Specify the number of factors.

- **Eigenvalue Cutoff**

  Specify the minimum eigenvalue amount. All factors whose eigenvalues are greater than or equal to this value will be retained. Older statistical texts suggest that you should only keep factors whose eigenvalues are greater than one.

Factor Selection - Value

This option sets a quantity corresponding to the method selected by the last option. For example, if you specified a Percent of Eigenvalues in the last option, you would enter the percentage (perhaps 80) here.

Reports Tab

The following options control the format of the reports.

Select Reports

Descriptive Statistics - Residuals (Q and T2)

These options let you specify which reports are displayed.

Select Plots

Scores Plot - Loadings Plot

These options let you specify which reports and plots are displayed.

Row Numbers

Indicates whether to display row numbers on the individual points in the corresponding plot.

Report Options

Alpha

The alpha value that is used in the residual reports to test if the observation is an outlier.

Minimum Loading

Specifies the minimum absolute value that a loading can have and still remain in the Variable List report.

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 variable names, variable labels, or both.

Plot Options

Number Factors Plotted

You can limit the number of plots generated using this parameter. Usually, you will only have interest in the first three or four factors.

Scores Plot Tab and Loadings Plot Tab

These sections specify the pair-wise plots of the scores and loadings.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis 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.

Storage Tab

The factor scores and/or the correlation matrix may be stored on the current database for further analysis. This group of options let 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. Note that existing data are replaced.

Data Storage Variables

Factor Scores

You can automatically store the factor scores for each row into the variables specified here. These scores are generated for each row of data in which all independent variable values are nonmissing.

Correlation or Covariance Matrix

Specifies variables to receive the correlation or covariance matrix. The type of matrix saved (i.e., correlation matrix or covariance matrix) depends on the Matrix Type option specified on the Variables tab.

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 Analysis

This section presents an example of how to run a principal components analysis. The data used are found in the PCA2 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Principal Components Analysis window.

1 Open the PCA2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **PCA2.s0**.
- Click **Open**.

2 Open the Principal Components Analysis window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Principal Components Analysis**. The Principal Components Analysis 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 Analysis window, select the **Variables tab**.
- Double-click in the **Variables** box. This will bring up the variable selection window.
- Select **X1** through **X6** from the list of variables and then click **Ok**. “X1-X6” will appear in the Variables box.

4 Specify which reports.

- Select the **Reports tab**.
- Check all reports and plots. Normally you would only view a few of these reports, but we are selecting them all 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
Variables       Count     Mean            Deviation        Communality
X1              30        44.2            24.66241         1.000000
X2              30        51.53333        30.57803         1.000000
X3              30        54.93333        29.05753         1.000000
X4              30        41.7            25.3175          1.000000
X5              30        43.66667        26.65143         1.000000
X6              30        47.63334        34.18962         1.000000
```

This report lets us compare the relative sizes of the standard deviations. In this data set, they are all about the same size, so we could analyze either the correlation or the covariance matrix. We will analyze the correlation matrix.

Count, Mean, and Standard Deviation

These are the familiar summary statistics of each variable. They are displayed to allow you to make sure that you have specified the correct variables. Note that using missing value imputation or robust estimation will change these values.

Communality

The communality shows how well this variable is predicted by the retained factors. It is the R^2 that would be obtained if this variable were regressed on the factors that were kept. In this example, all factors were kept, so the R^2 is one.

Correlation Section

```
Correlation Section
                    Variables
Variables                X1           X2           X3           X4           X5
X1                 1.000000     0.347229     0.224730     0.734112     0.819983
X2                 0.347229     1.000000     0.990372     0.557526     0.799049
X3                 0.224730     0.990372     1.000000     0.475404     0.710086
X4                 0.734112     0.557526     0.475404     1.000000     0.830195
X5                 0.819983     0.799049     0.710086     0.830195     1.000000
X6                 0.514102     0.974167     0.935223     0.693869     0.907416
Phi=0.735970  Log(Det|R|)=-22.779188  Bartlett Test=596.06  DF=15  Prob=0.000000

                    Variables
Variables                X6
X1                 0.514102
X2                 0.974167
X3                 0.935223
X4                 0.693869
X5                 0.907416
X6                 1.000000
Phi=0.735970  Log(Det|R|)=-22.779188  Bartlett Test=596.06  DF=15  Prob=0.000000

Bar Chart of Absolute Correlation Section
                      Variables
Variables             X1                    X2                    X3                    X4                    X5
X1                                          |||||||               |||||                 |||||||||||||||       |||||||||||||||||
X2                    |||||||                                     ||||||||||||||||||||  ||||||||||||          ||||||||||||||||
X3                    |||||                 ||||||||||||||||||||                        ||||||||||            |||||||||||||||
X4                    |||||||||||||||       ||||||||||||          ||||||||||                                  |||||||||||||||||
X5                    |||||||||||||||||     ||||||||||||||||      |||||||||||||||       |||||||||||||||||
X6                    |||||||||||           ||||||||||||||||||||  |||||||||||||||||||   ||||||||||||||        |||||||||||||||||||
Phi=0.735970  Log(Det|R|)=-22.779188  Bartlett Test=596.06  DF=15  Prob=0.000000

Bar Chart of Absolute Correlation Section
                      Variables
Variables             X6
X1                    |||||||||||
X2                    ||||||||||||||||||||
X3                    |||||||||||||||||||
X4                    ||||||||||||||
X5                    |||||||||||||||||||
X6
Phi=0.735970  Log(Det|R|)=-22.779188  Bartlett Test=596.06  DF=15  Prob=0.000000
```

The report gives the correlations for a test of the overall correlation structure in the data. In this example, we notice several high correlation values. The Gleason-Staelin redundancy measure, phi, is 0.736, which is quite large. There is apparently some correlation structure in this data set that can be modeled. If all the correlations were small, there would be no need for a PCA.

Correlations

The simple correlations between each pair of variables. Note that using the missing value imputation or robust estimation options will affect the correlations in this report. When the above options are not used, the correlations are constructed from those observations having no missing values in any of the specified variables.

Phi

This is the Gleason-Staelin redundancy measure of how interrelated the variables are. A zero value of φ means that there is no correlation among the variables, while a value of one indicates perfect correlation among the variables. This coefficient may have a value less than 0.5 even when there is obvious structure in the data, so care should to be taken when using it. This statistic is especially useful for comparing two or more sets of data. The formula for computing φ is:

$$\varphi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{2} - p}{p(p-1)}}$$

Log(Det|R|)

This is the log (base e) of the determinant of the correlation matrix. If you used the covariance matrix, this is the log (base e) of the determinant of the covariance matrix.

Bartlett Test, DF, Prob

This is Bartlett's sphericity test (Bartlett, 1950) for testing the null hypothesis that the correlation matrix is an identity matrix (all correlations are zero). If you get a probability (Prob) value greater than 0.05, you should not perform a PCA on the data. The test is valid for large samples (*N>150*). It uses a Chi-square distribution with *p(p-1)/2* degrees of freedom. Note that this test is only available when you analyze a correlation matrix. The formula for computing this test is:

$$\chi^2 = \frac{(11 + 2p - 6N)}{6} \mathrm{Log}_e |R|$$

Bar Chart of Absolute Correlation Section

This chart graphically displays the absolute values of the correlations. It lets you quickly find high and low correlations.

Eigenvalues Section

```
Eigenvalues
                               Individual    Cumulative
No.          Eigenvalue          Percent       Percent    Scree Plot
1              4.562633            76.04         76.04    ||||||||||||||||
2              1.171509            19.53         95.57    ||||
3              0.242834             4.05         99.62    |
4              0.022878             0.38        100.00    |
5              0.000105             0.00        100.00    |
6              0.000041             0.00        100.00    |
```

Eigenvalue

The eigenvalues. Often, these are used to determine how many factors to retain. (In this example, we would retain the first two eigenvalues.)

When the PCA is run on the correlations, one rule-of-thumb is to retain those factors whose eigenvalues are greater than one. The sum of the eigenvalues is equal to the number of variables. Hence, in this example, the first factor retains the information contained in 4.563 of the original variables.

When the PCA is run on the covariances, the sum of the eigenvalues is equal to the sum of the variances of the variables.

Individual and Cumulative Percents

The first column gives the percentage of the total variation in the variables accounted for by this factor. The second column is the cumulative total of the percentage. Some authors suggest that the user pick a cumulative percentage, such as 80% or 90%, and keep enough factors to attain this percentage.

Scree Plot

This is a rough bar plot of the eigenvalues. It enables you to quickly note the relative size of each eigenvalue. Many authors recommend it as a method of determining how many factors to retain.

The word *scree*, first used by Cattell (1966), is usually defined as the rubble at the bottom of a cliff. When using the scree plot, you must determine which eigenvalues form the "cliff" and which form the "rubble." You keep the factors that make up the cliff. Cattell and Jaspers (1967) suggest keeping those that make up the cliff plus the first factor of the rubble.

Interpretation of the Example

This table presents the eigenvalues of the correlation (covariance) matrix. The first question that we would ask is how many factors should be kept. The scree plot shows that the first two factors are indeed the largest. The cumulative percentages show that the first two factors account for over 95% of the variation. Only the first two eigenvalues are greater than one. We begin to get the impression that the correct answer is that two factors will adequately approximate these data.

We note in passing that the third and fourth eigenvalues are several orders of magnitude larger than the fifth and sixth. We will keep our eyes on these factors as well. Although they are not significant, they certainly represent some artifact in the data.

Eigenvectors Section

```
Eigenvectors
                  Factors
Variables         Factor1       Factor2       Factor3       Factor4       Factor5
X1              -0.315300     -0.639819      0.507144      0.437189      0.012097
X2              -0.427719      0.372949      0.077702      0.188932      0.786262
X3              -0.399630      0.476348      0.016354      0.485911     -0.590471
X4              -0.379732     -0.385432     -0.830977      0.126310      0.027636
X5              -0.452963     -0.197773      0.209428     -0.567541     -0.140774
X6              -0.456690      0.192251      0.046011     -0.446098     -0.111393

                  Factors
Variables         Factor6
X1               0.206739
X2              -0.134256
X3              -0.168389
X4              -0.002258
X5              -0.608219
X6               0.735489
```

```
Bar Chart of Absolute Eigenvectors
                    Factors
Variables           Factor1          Factor2          Factor3          Factor4          Factor5
X1                  |||||||          |||||||||||||    |||||||||||      |||||||||        |
X2                  |||||||||        ||||||||         ||               ||||             ||||||||||||||||
X3                  ||||||||         ||||||||||       |                ||||||||||       ||||||||||||
X4                  ||||||||         ||||||||         |||||||||||||||||  |||              |
X5                  ||||||||||       ||||             |||||            ||||||||||||     |||
X6                  ||||||||||       ||||             |                |||||||||        |||

Bar Chart of Absolute Eigenvectors
                    Factors
Variables           Factor6
X1                  |||||
X2                  |||
X3                  ||||
X4                  |
X5                  |||||||||||||
X6                  |||||||||||||||
```

Eigenvector

The eigenvectors are the weights that relate the scaled original variables, $x_i = (X_i - Mean_i)/Sigma_i$, to the factors. For example, the first factor, $Factor_1$, is the weighted average of the scaled variables, the weight of each variable given by the corresponding element of the first eigenvector. Mathematically, the relationship is given by:

$$Factor_1 = v_{11}x_{11} + v_{12}x_{12} + \ldots + v_{1p}x_{1p}$$

These coefficients may be used to determine the relative importance of each variable in forming the factor. Often, the eigenvectors are scaled so that the variances of the factor scores are equal to one. These scaled eigenvectors are given in the *Score Coefficients* section described later.

Bar Chart of Absolute Eigenvectors

This chart graphically displays the absolute values of the eigenvectors. It lets you quickly interpret the eigenvector structure. By looking at which variables correlate highly with a factor, you can determine what underlying structure it might represent.

Factor Loadings Section

```
Factor Loadings
                       Factors
Variables              Factor1       Factor2       Factor3       Factor4       Factor5
X1                   -0.673491     -0.692517      0.249911      0.066127      0.000124
X2                   -0.913622      0.403667      0.038290      0.028577      0.008051
X3                   -0.853623      0.515581      0.008059      0.073497     -0.006046
X4                   -0.811120     -0.417177     -0.409490      0.019105      0.000283
X5                   -0.967543     -0.214062      0.103202     -0.085844     -0.001441
X6                   -0.975505      0.208086      0.022673     -0.067475     -0.001141

                       Factors
Variables              Factor6
X1                    0.001326
X2                   -0.000861
X3                   -0.001080
X4                   -0.000014
X5                   -0.003900
X6                    0.004716
```

```
Bar Chart of Absolute Factor Loadings
                    Factors
Variables           Factor1               Factor2           Factor3           Factor4           Factor5
X1                  ||||||||||||||        ||||||||||||||    |||||             ||                |
X2                  |||||||||||||||||||   |||||||||         |                 |                 |
X3                  ||||||||||||||||||    |||||||||||       |                 ||                |
X4                  |||||||||||||||||     |||||||||         |||||||||         |                 |
X5                  ||||||||||||||||||||  |||||             |||               ||                |
X6                  ||||||||||||||||||||  |||||             |                 ||                |

Bar Chart of Absolute Factor Loadings
                    Factors
Variables           Factor6
X1                  |
X2                  |
X3                  |
X4                  |
X5                  |
X6                  |
```

Factor Loadings

These are the correlations between the variables and factors.

Bar Chart of Absolute Factor Loadings

This chart graphically displays the absolute values of the factor loadings. It lets you quickly interpret the correlation structure. By looking at which variables correlate highly with a factor, you can determine what underlying structure it might represent.

Interpretation of the Example

We now go through the interpretation of each factor. Factor one appears to be an average of all six variables. Although the weights of all variables are large, the weights on X5 and X6 are the largest. Factor two appears to be a contrast (difference) of X2+X3 and X1+X4. Factor three is most highly correlated to X4. Factor four appears to be associated with several variables, but most highly with X5. Factor five is a contrast of X2 and X3. Factor six is a contrast of X5 and X6. If these data were real, we could try to attach meaning to these patterns.

Communality Section

```
Communalities
                    Factors
Variables           Factor1     Factor2     Factor3     Factor4     Factor5
X1                  0.453590    0.479580    0.062456    0.004373    0.000000
X2                  0.834705    0.162947    0.001466    0.000817    0.000065
X3                  0.728671    0.265824    0.000065    0.005402    0.000037
X4                  0.657916    0.174037    0.167682    0.000365    0.000000
X5                  0.936140    0.045823    0.010651    0.007369    0.000002
X6                  0.951610    0.043300    0.000514    0.004553    0.000001

                    Factors
Variables           Factor6     Communality
X1                  0.000002    1.000000
X2                  0.000001    1.000000
X3                  0.000001    1.000000
X4                  0.000000    1.000000
X5                  0.000015    1.000000
X6                  0.000022    1.000000
```

```
Bar Chart of Communalities
                  Factors
Variables         Factor1               Factor2      Factor3      Factor4      Factor5
X1                ||||||||||            ||||||||||   ||           |            |
X2                |||||||||||||||||     ||||         |            |            |
X3                |||||||||||||||       ||||||       |            |            |
X4                ||||||||||||||        ||||         ||||         |            |
X5                |||||||||||||||||||   |            |            |            |
X6                ||||||||||||||||||||  |            |            |            |

                  Factors
Variables         Factor6      Communality
X1                |            ||||||||||||||||||||
X2                |            ||||||||||||||||||||
X3                |            ||||||||||||||||||||
X4                |            ||||||||||||||||||||
X5                |            ||||||||||||||||||||
X6                |            ||||||||||||||||||||
```

Communality

The communality is the proportion of the variation of a variable that is accounted for by the factors that are retained. It is the R² value that would be achieved if this variable were regressed on the retained factors. This table value gives the amount added to the communality by each factor.

Bar Chart of Communalities

This chart graphically displays the values of the communalities.

Factor Structure Summary Section

```
Factor Structure Summary Section


Factor1      Factor2      Factor3      Factor4      Factor5      Factor6
X6           X1           X4
X5           X3
X2           X4
X3           X2
X4
X1
```

Interpretation

This report is provided to summarize the factor structure. Variables with an absolute loading greater than the amount set in the *Minimum Loading* option are listed under each factor. Using this report, you can quickly see which variables are related to each factor. Notice that it is possible for a variable to have high loadings on several factors.

Score Coefficients Section

```
Score Coefficients
              Factors
Variables     Factor1       Factor2       Factor3        Factor4       Factor5
X1            -0.1476102    -0.5911322    1.029146       2.890409      1.181451
X2            -0.20024      0.3445697     0.1576812      1.249092      76.79078
X3            -0.1870899    0.4401002     3.318755E-02   3.212524      -57.6688
X4            -0.1777746    -0.356102     -1.686299      0.8350784     2.699127
X5            -0.2120581    -0.1827235    0.4249913      -3.752211     -13.7488
X6            -0.2138031    0.1776219     9.337004E-02   -2.949311     -10.87929

Score Coefficients
              Factors
Variables     Factor6
X1            32.24512
X2            -20.93996
X3            -26.26369
X4            -0.3522398
X5            -94.86387
```

Score Coefficients

These are the coefficients that are used to form the factor scores. The factor scores are the values of the factors for a particular row of data. These score coefficients are similar to the eigenvectors. They have been scaled so that the scores produced have a variance of one rather than a variance equal to the eigenvalue. This causes each of the factors to have the same variance.

You would use these scores if you wanted to calculate the factor scores for new rows not included in your original analysis.

Residual Section

```
Residual Section
Row        T2   T2 Prob      Q0      Q1      Q2      Q3      Q5
1        4.68   0.6932    10.68    0.91    0.11    0.00    0.00
2      27.94*   0.0078    12.76    0.66    0.65   0.57*    0.00
3      28.02*   0.0077    12.93   6.84*   6.23*    0.01    0.00
4        2.25   0.9250     3.04    1.61    0.06    0.00    0.00
5        8.20   0.3742     1.53    0.95    0.01    0.00   0.00*
6        4.20   0.7427     4.52    0.23    0.01    0.01    0.00
7        1.33   0.9785     1.86    0.01    0.00    0.00    0.00
8        3.06   0.8573     5.47    2.29    0.07    0.00    0.00
(report continues through all thirty rows)
```

This report is useful for detecting outliers--observations that are very different from the bulk of the data. To do this, two quantities are displayed: T^2 and Q_k. We will now define these two quantities.

T^2 measures the combined variability of all the variables in a single observation. Mathematically, T^2 is defined as:

$$T^2 = [\underline{x} - \underline{\bar{x}}]' S^{-1} [\underline{x} - \underline{\bar{x}}]$$

where $\underline{x}$ represents a p-variable observation, $\underline{\bar{x}}$ represents the p-variable mean vector and S^{-1} represents the inverse of the covariance matrix.

T is not affected by a change in scale. It is the same whether the analysis is performed on the covariance or the correlation matrix. T^2 gives a scaled distance measure of an individual observation from the overall mean. The closer an observation is to its mean, the smaller will be the value of T^2.

If the variables follow a multivariate normal distribution, then the probability distribution of T^2 may be related to the common F distribution using the formula:

$$T^{2}_{p,n,\alpha} = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha}$$

Using this relationship, we can perform a statistical test at a given level of significance to determine if the observation is significantly different from the vector of means. You set the α value using the *Alpha* option. Since this test is being performed *N* times, you would anticipate about $N(1\text{-}\alpha)$ observations to be significant by chance variation. In our current example, rows two and three are starred (which means they were significant at the .05 significance level). You would probably want to check for data entry or transcription errors. (Of course, in this data set, these rows were made to be outliers.)

T^2 is really not part of a normal PCA since it may be calculated independently. It is presented to help detect observations that may have an undue influence on the analysis. You can read more about its use and interpretation in Jackson (1991).

The other quantity shown on this report is Q_k. Q_k represents the sum of squared residuals when an observation is predicted using the first *k* factors. Mathematically, the formula for Q_k is:

$$\begin{aligned} Q_k &= (\underline{x} - \underline{\hat{x}})'(\underline{x} - \underline{\hat{x}}) \\ &= \sum_{i=1}^{p} \left(x_i - {}_k\hat{x}_i \right)^2 \\ &= \sum_{i=k+1}^{p} \lambda_i \left(pc_i \right)^2 \end{aligned}$$

Here ${}_k x_i$ refers to the value of variable *i* predicted from the first *k* factors, λ_i refers to the i^{th} eigenvalue, and pc_i is the score of the i^{th} factor for this particular observation. Further details are given in Jackson (1991) on pages 36 and 37.

An upper limit for Q_k is given by the formula:

$$Q_\alpha = a\left[\frac{z_\alpha \sqrt{2b\,h^2}}{a} + \frac{bh(h-1)}{a^2} + 1 \right]^{1/h}$$

where

$$a = \sum_{i=k+1}^{p} \lambda_i$$

$$b = \sum_{i=k+1}^{p} \lambda_i^2$$

$$c = \sum_{i=k+1}^{p} \lambda_i^3$$

$$h = 1 - \frac{2ac}{3\,b^2}$$

and

z_α is the upper normal deviate of area α if *h* is positive or the lower normal deviate of area α if *h* is negative.

This limit is valid for any value of *k*, whether too many or too few factors are kept. Note that these formulas are for the case when the correlation matrix is being used. When the analysis is being run on the covariance matrix, the pc_i*'s* must be adjusted. Further details are given in Jackson (1991).

Notice that significant (starred) values of Q_k indicate observations that are not duplicated well by the first *k* factors. These should be checked to see if they are valid. Q_k and T^2 provide an initial data screening tool.

Interpretation of the Example

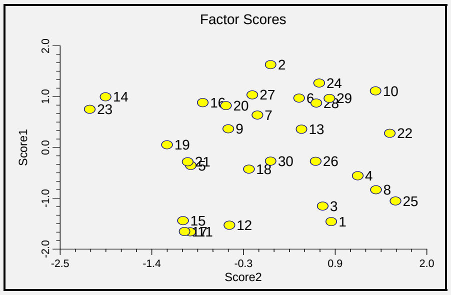
We are interested in two columns in this report: Q2 and T2. Notice that rows two and three are significantly large (shown by the asterisk) for both measurements. If these were real data, we would investigate these two rows very carefully. We would first check for data entry errors and next for errors that might have occurred when the measurements were actually taken. In our case, we know that these two rows are outliers (since they were artificially made to be outliers).

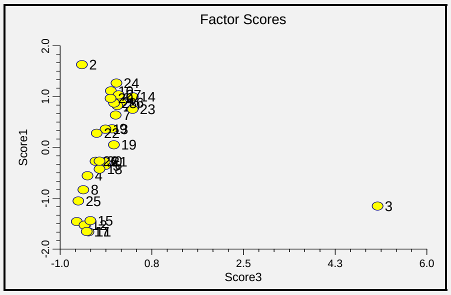
Factor Scores Section

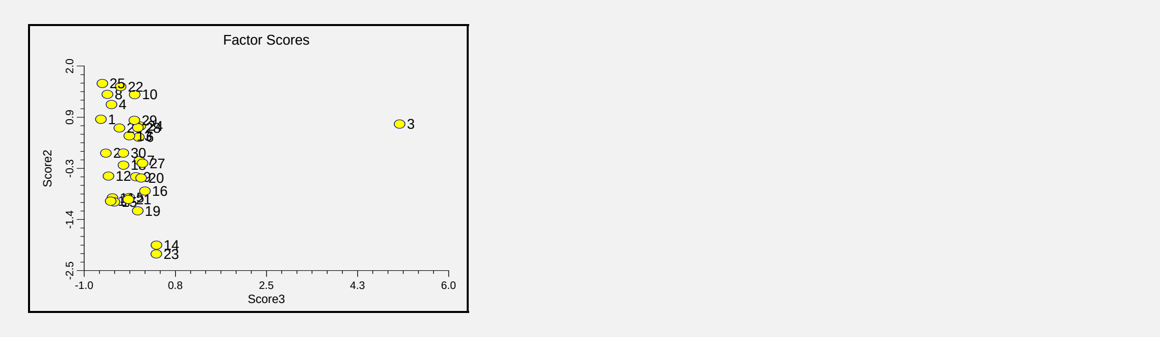
```
Factor Score
         Factors
Row      Factor1     Factor2     Factor3     Factor4     Factor5     Factor6
1        -1.4627      0.8272     -0.6797     -0.1124      1.1732      0.0689
2         1.6286      0.0834     -0.5825     -4.9911     -0.0743      0.1499
3        -1.1560      0.7225      5.0582     -0.7581     -0.0226      0.0079
4        -0.5595      1.1520     -0.4768      0.0670      0.5125      0.3465
5        -0.3573     -0.8963     -0.1360      0.2037     -1.6830      2.0931
6         0.9696      0.4329      0.0488      0.6208     -1.6155     -0.2796
7         0.6364     -0.0783      0.0624      0.4003     -0.8677     -0.0678
8        -0.8339      1.3759     -0.5545     -0.0806     -0.3899     -0.0447
(report continues through all thirty rows)
```

This report presents the individual factor scores scaled so each column has a mean of zero and a standard deviation of one. These are the values that are plotted in the plots to follow. Remember, there is one row of score values for each observation and one column for each factor that was kept.

Factor Score Plots





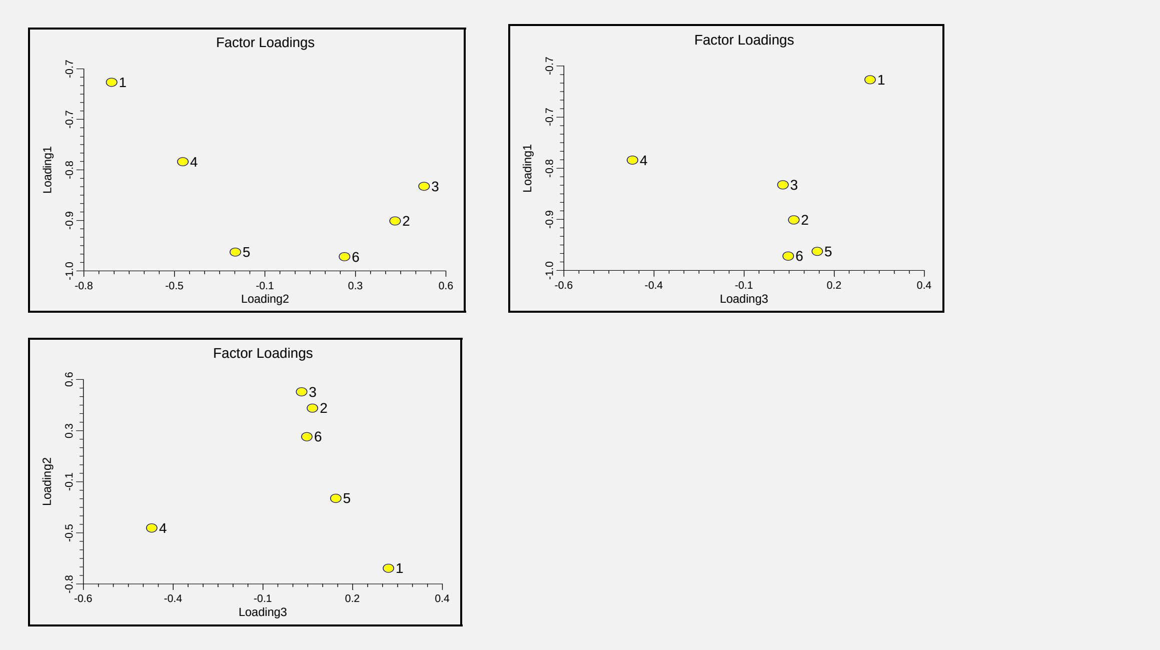


This set of plots shows each factor plotted against every other factor. The first k factors (where k is the number of large eigenvalues) usually show the major structure that will be found in the data. The rest of the factors show outliers and linear dependencies. Note that in our present example, outliers are displayed in both plots that include factor three. We would now investigate these rows much more closely.

Interpretation of the Example

We first notice the presence of an outlier in plots containing factor three. If we were not convinced before that this row was an outlier, we would be now. Factor three fits row three. If we had called for the plot of factor four, we would see that it fits row four. Hence, these plots of nonsignificant factors show outliers.

Factor Loading Plots



Discussion of Factor Loading Plots

This set of plots shows each of the factor loading columns plotted against each other. The data points represent variables. The plot allows you to find variables that are highly correlated with both factors. It is anticipated that this will aid in the interpretation of the factors.

Robust and Missing-Value Iteration Section

The following report is not part of the preceding tutorial. We have re-run the problem calling for robust estimation so that we could show you this iteration report. We have set the number of robust iterations at six.

```
Robust and Missing-Value Estimation Iteration Section
                            Trace of          Percent
No.       Count             Covar Matrix      Change
0         30                4907.795            0.00
1         30                4907.795            0.00
2         30                4423.718           -9.86
3         30                4423.718            0.00
4         30                4353.748           -1.58
5         30                4353.748            0.00
6         30                4335.77            -0.41
```

This report presents the progress of the iterations. The trace of the covariance matrix gives a measure of what is happening at each iteration. When this value stabilizes, the program has converged. The percent change is reported to let you determine how much the trace has changed. In this particular example, we see very little change between iterations five and six. We would feel comfortable in stopping at this point. A look at the Descriptive Statistics section will let you see how much the means and standard deviations have changed.

A look at the Residual Section will let you see the robust weights that are assigned to each row. Those weights that are near zero indicate observations whose influence has been removed by the robust procedure.

Chapter 430

Correspondence Analysis

Introduction

Correspondence analysis (CA) is a technique for graphically displaying a two-way table by calculating *coordinates* representing its rows and columns. These coordinates are analogous to factors in a principal components analysis (used for continuous data), except that they partition the Chi-square value used in testing independence instead of the total variance.

For those of you new to CA, we suggest that you obtain Greenacre (1993). This is an excellent introduction to the subject, is very readable, and is suitable for self study. If you want to understand the technique in detail, you should obtain this (paperback) book.

Discussion

We will explain CA using the following example. Suppose an aptitude survey consisting of eight yes or no questions is given to a group of tenth graders. The instructions on the survey allow the students to answer only those questions that they want to. The results of the survey are tabulated as follows.

Aptitude Survey Results – Counts

Question	Yes	No	Total
Q1	155	938	1093
Q2	19	63	82
Q3	395	542	937
Q4	61	64	125
Q5	1336	876	2212
Q6	22	14	36
Q7	864	354	1218
Q8	920	185	1105
Total	3772	3036	6808

Take a few moments to study this table and see what you can discover. The most obvious pattern is that many of the students did not answer all the questions. This makes response patterns between rows difficult to analyze.

To solve this problem of differential response rates, we create a table of row percents (or *row profiles* as they are called in CA).

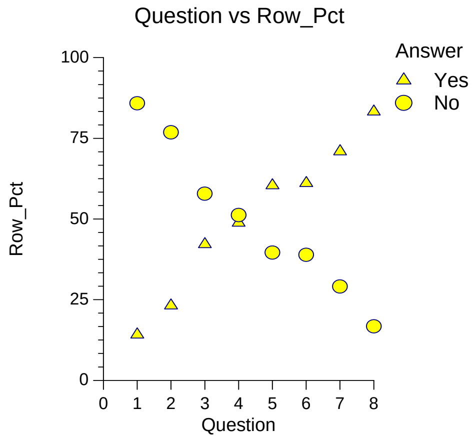
Aptitude Survey Results – Row Profiles

Question	Yes	No	Total
Q1	14.18	85.82	100.00
Q2	23.17	76.83	100.00
Q3	42.16	57.84	100.00
Q4	48.80	51.20	100.00
Q5	60.40	39.60	100.00
Q6	61.11	38.89	100.00
Q7	70.94	29.06	100.00
Q8	83.26	16.74	100.00
Total	55.41	44.59	100.00

This table allows us to see the underlying patterns within the table. We note that only 14% answered yes to question one while 83% answered yes to question eight.

Although we can inspect this table directly, a picture of the data will allow us to find patterns much more quickly. This is done in the following scatter plot. The plot shows the row profiles with the questions on the horizontal axis and the row percents on the vertical axis. Notice that there are two possible responses (yes or no) and two corresponding plotting symbols.

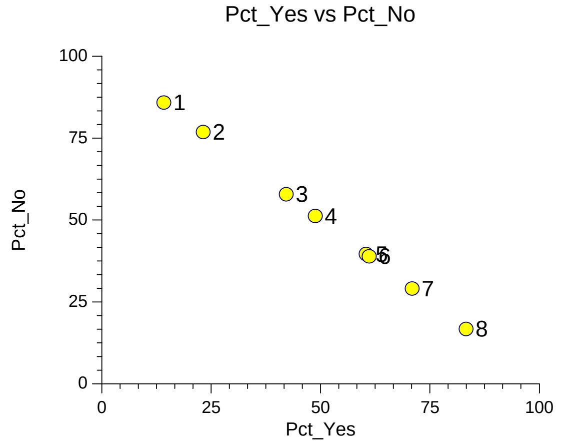
Aptitude Survey Results – Row Percents versus Questions



We notice a steady gain in the percent of students answering yes as we move from question one to question eight. Also, we can see the obvious relationship between the percent answering yes and the percent answering no. In fact, if you think about it for a moment you will realize that we really only need to plot the yes's or the no's, but not both since they both relate the same information.

Another way of plotting this data is to plot the percentage of each possible answer on a different axis. Since, in this example, we have two possible answers, we plot the yes percentage on one axis and the no percentage on the other axis.

Aptitude Survey Results – Scatter Plots



Spend some time analyzing this plot. Can you see the connection between the last plot and this plot? In the previous plot, each possible answer was a horizontal set of points. In this plot, each answer is an axis. Hence, if our survey was made up of multiple-choice questions each with three possible answers, this plot would need to have been three dimensional.

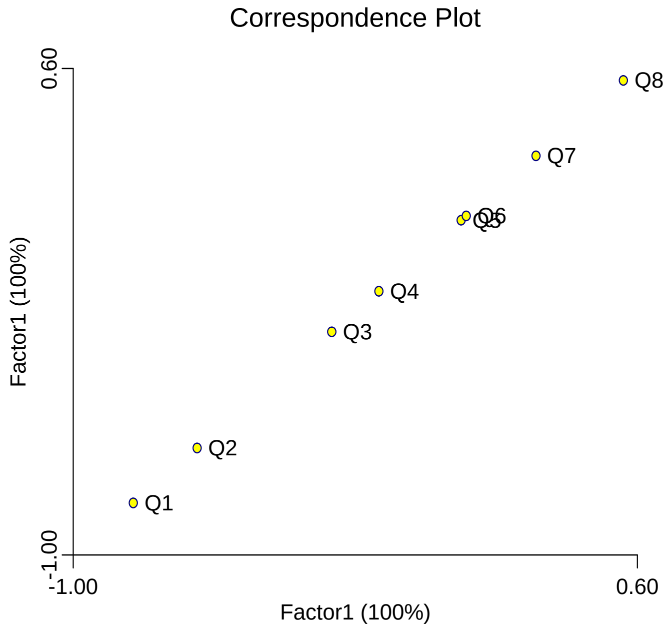
This plot is an example of a *correspondence map*, the primary output of CA. It is important to understand the features of this plot. Each axis of the plot represents a column and each point represents a row of the original table. If you were to draw a bar chart for this data, you would create bars representing the distances from each point to each axis. Since there are eight points and two axes, the bar chart would have sixteen bars.

Notice also how you interpret the plot. Each point represents a specific yes or no combination. For example, question eight had about 83% answering yes and 17% answering no. This high proportion of yes respondents positions the point very close to the Pct_Yes (horizontal) axis. Conversely, question one had a large percentage of no answers and is very near the Pct_No axis. We see that points relatively close to the end of a particular axis have a high percentage value on that axis.

Also notice that points that are close to each other have very similar patterns of yes or no answers. Look at the row profiles (percents) for questions five and six (two points that are almost in the same position). Notice that they differ by only one percentage point.

The following figure shows the CA plot of this data generated by the program. Although the orientation of the plot is different, the distances between the points are the same.

Aptitude Survey Results – CA Plot



This plot and the last plot appear the same because, in this example, the CA plot reproduces the plot of the row percents. This occurs because there were only two possible answers to each question. If the questions in our survey had been multiple choice so that there were four or five answers, we would have needed to create a four or five dimensional plot to reproduce the Row Profile Plot.

To summarize, then, a CA plot is a plot of the row profiles (percentages) constructed so that each column category becomes a different dimension. Since it is only possible to view two dimensions at a time, we must project this high dimensional space onto a two-dimensional subspace. This projection is constructed so as to maintain as much of the original information (or variation) as possible.

Technical Details

We will now present an outline of the computational methods used to perform the analysis. We will use standard matrix terminology to present the steps.

1. Read in the *n* (rows) by *m* (columns) data matrix, ***K***. Note that the elements of ***K*** must be non-negative and that none of the row or column totals is zero.
2. Compute the proportion matrix, **P**, by dividing the elements of **K** by the total of all numbers in **K**. Mathematically, we write

$$\mathbf{P} = \left\{ p_{ij} \right\} = \left\{ k_{ij} / k_{..} \right\}$$

3. Compute the totals of the rows of **P** and the columns of **P**, putting the results in the vectors **r** and **c**. Using standard matrix notation, we write

$$\mathbf{r} = \mathbf{P1}$$

$$\mathbf{c} = \mathbf{P'1}$$

where **1** is an appropriately dimensioned vector of ones.

4. Change the square roots of the vectors **r** and **c** into diagonal matrices and take the inverse of the resulting square matrices.

$$\mathbf{D}_r = \left[diag(\mathbf{r})\right]^{-1/2}$$

$$\mathbf{D}_c = \left[diag(\mathbf{c})\right]^{-1/2}$$

5. Compute the scaled matrix, **A**.

$$\mathbf{A} = \mathbf{D}_r \mathbf{P} \mathbf{D}_c$$

6. Compute the Singular Value Decomposition (SVD) of **A**.

$$\langle \mathbf{B}, \mathbf{W}, \mathbf{C} \rangle = SVD(\mathbf{A})$$

7. Compute the coordinate matrices, **F** and **G**, as follows:

$$\mathbf{F} = \mathbf{D}_r \mathbf{B} \mathbf{W}$$

$$\mathbf{G} = \mathbf{D}_c \mathbf{C} \mathbf{W}'$$

8. Compute the eigenvalues, **V**.

$$\mathbf{V} = \mathbf{W}\mathbf{W}'$$

9. Compute the row distances, d_i, and the column distances, d_j.

$$d_i = \sum_j \left(\frac{1}{p_{.j}}\right)\left(\frac{p_{ij}}{p_{i.}} - p_{.j}\right)^2$$

$$d_j = \sum_i \left(\frac{1}{p_{i.}}\right)\left(\frac{p_{ij}}{p_{.j}} - p_{i.}\right)^2$$

10. Note that the weights, w_i and w_j, come from the vectors **r** and **c** that were formed in step 3.

$$w_i = \{r_i\}$$

$$w_j = \{c_j\}$$

11. Compute the reported statistics as follows:

Statistic	**Formula**
Mass	w_i
Inertia	$\frac{w_i d_i^2}{\sum_k w_k^2 d_k^2}$
Distance	d_i^2
Row Factor	f_{ij}
Column Factor	g_{ij}
Row COR	$\frac{f_{ij}^2}{d_i^2}$
Column COR	$\frac{g_{ij}^2}{d_j^2}$

Statistic	Formula
Row CTR	$\dfrac{w_i f_{ij}^2}{v_i}$
Column CTR	$\dfrac{w_j g_{ij}^2}{v_j}$
Angle	$ArcCos\left(\sqrt{COR_{ij}}\right)$

Data Structure

We will use a set of data from Greenacre (1993) in the tutorial that follows. The table below shows the results of a survey relating the smoking habits of the employees of a fictitious company to their position within the company. These data are contained in the CORRES1 database .

The entries in the table are the counts of the number of employees falling into each cell.

CORRES1 dataset

None	Light	Medium	Heavy	Staff
4	2	3	2	(SM) Senior Managers
4	3	7	4	(JM) Junior Managers
25	10	12	4	(SE) Senior Employees
18	24	33	13	(JE) Junior Employees
10	6	7	2	(SE) Secretaries

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Two-Way Table Variables

Table Variables

Specify the variables containing the two-way table to be analyzed. Note that this is a previously tabulated table. The procedure cannot be used on a raw survey itself. The data must first be tabulated into a two-way table before this procedure can be applied.

Supplementary Variables (Not used in calculation of axes.)

Supplementary Variables

This is an optional list of one or more supplementary variables (columns) which may be specified. Supplementary columns are not used in the calculation of the axes, but are displayed on all plots and reports.

The actual values that you use do not matter since the entries are all standardized before they are used.

Row Description Variables

Row Type Variable

This is an optional variable that indicates the role each row of data will play in the analysis. Only the values 0, 1, and 2 are allowed in this variable. These values have the following interpretation by the program.

0 The row is completely ignored.

1 The row is used as a regular row of data.

2 The row is a supplementary row. It is not used in the formation of the axes, but it is displayed on all plots and reports.

Row Label Variable

Specify an optional variable containing labels for each of the rows.

Options

Number of Axes

This option specifies the number of axes (factors or coordinates) on which reports and plots should be generated. Usually you will keep only two or three axes.

Zero Replacement

This value replaces zeros in the data. Zeros can cause problems during the calculations. Changing zeros to a small positive number circumvents these problems without changing the results a great deal.

RANGE: 0 to 1E-300.

RECOMMENDED: 0.00000001

Row/Column Name

On the plots, this label is used to designate the rows and columns.

This allows you to assign a more meaningful name than just 'Rows' or 'Columns'.

Reports Tab

The following options control which reports are displayed.

Select Reports

Raw Data Report – Eigenvalue Report

Specify whether to display the indicated reports. Some of these options let you specify whether to display a separate report for the rows and columns.

Plot Report

Indicate whether to display this report or plot for rows only, columns only, or both.

Axis Report

Indicate whether to display this report or plot for rows only, columns only, or both.

Select Plots

Correspondence Plots

Specify whether to display the indicated plot.

Report Options

Scale Factor

Most of the output consists of proportions--decimal numbers between zero and one. Tables displaying these numbers may be more readable if they are multiple by a scale factor (such as 100 or 1000). This gets rid of the leading zero and often the decimal point as well. You select the scale factor here. Following are examples of how various numbers are displayed with the various scale factors.

Original	Scale Factor		
Number	1	100	1000
0.034321	0.034	3.4	34
0.923514	0.924	92.4	924
0.512345	0.512	51.2	512

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.

Plot Options

Size of Plots

This option controls the size of the plots that are displayed. You can select *small*, *medium*, or *large*. *Medium* and *large* are displayed one per line, while *small* are displayed two per line.

Correspondence Plot Tab

These options control the attributes of the correspondence plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis 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.

Plot Settings - Symbols

Row

Designate the plot symbol used for plotting rows.

Column

Designate the plot symbol used for plotting columns.

Plot Settings - Legend

Legend

Specify whether you want to view a legend.

Legend Text

Specifies the legend title.

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.

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 – Correspondence Analysis

This section presents an example of how to run an analysis of the data presented in the table above. These data are contained in the CORRES1 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Correspondence Analysis window.

1 Open the Corres1 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Corres1.s0**.
- Click **Open**.

2 Open the Correspondence Analysis window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Correspondence Analysis**. The Correspondence Analysis 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 Correspondence Analysis window, select the **Variables tab**.
- Double-click in the **Table Variables** box. This will bring up the variable selection window.
- Select **None** to **Heavy** from the list of variables and then click **Ok**. “None-Heavy” will appear in the Table Variables box.
- Double-click in the **Row Label Variable** box. This will bring up the variable selection window.
- Select **Staff** from the list of variables and then click **Ok**. “Staff” will appear in the Row Label Variable box.

4 Specify the reports.

- Select the **Reports tab**.
- Enter **Both** in the **Correspondence Plots** box.

5 Specify the plot.

- Select the **Correspondence Plot tab**.
- Enter **0.5** in the **Horizontal - Maximum** box.

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).

Raw Data Section

```
Raw Data Section
Staff        None    Light  Medium     Heavy      Total
SM              4        2       3         2         11
JM              4        3       7         4         18
SE             25       10      12         4         51
JE             18       24      33        13         88
SC             10        6       7         2         25
Total          61       45      62        25        193
```

This report displays the raw data. It documents the data that were used by the procedure so that you can check for data-entry errors.

Row Profiles Section

```
Row Profiles Section

Staff        None    Light  Medium     Heavy      Total
SM          36.36    18.18   27.27     18.18     100.00
JM          22.22    16.67   38.89     22.22     100.00
SE          49.02    19.61   23.53      7.84     100.00
JE          20.45    27.27   37.50     14.77     100.00
SC          40.00    24.00   28.00      8.00     100.00
Total       31.61    23.32   32.12     12.95     100.00
```

This report shows the row profiles (percentages). These are the values the will be plotted on the row oriented plot. Note that since there are five rows, these data would require five dimensions to be plotted in the standard fashion. CA investigates the differences between each individual row profile and the average row profile (the row labeled “Total”).

Column Profiles Section

```
Column Profiles Section

Staff        None     Light  Medium    Heavy     Total
SM           6.56      4.44    4.84     8.00      5.70
JM           6.56      6.67   11.29    16.00      9.33
SE          40.98     22.22   19.35    16.00     26.42
JE          29.51     53.33   53.23    52.00     45.60
SC          16.39     13.33   11.29     8.00     12.95
Total      100.00    100.00  100.00   100.00    100.00
```

This report shows the column profiles (percentages). These are the values the will be plotted in a column oriented CA plot. Note that since there are four columns, these data would require four dimensions to be plotted in the standard fashion. CA investigates the differences between each individual column profile and the average column profile (the column labeled “Total”).

Eigenvalue Section

```
Eigenvalue Section

Factor                      Individual   Cumulative
No.      Eigenvalue            Percent      Percent        Bar Chart
1          0.074759              87.76        87.76        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2          0.010017              11.76        99.51        |||||||||
3          0.000414               0.49       100.00        |
Total      0.085190
```

Since CA projects the row (or column) profiles onto a two-dimensional subspace, a critical issue is how well this projection works. The eigenvalues gives us important information regarding this. The Cumulative Percent column tells us how much of the total information is reproduced by each number of dimensions.

In this example, the CA plot using the first two factors accounts for 99.5% of the variation. In other words, the dimension reduction is only costing us a 0.5% loss in information. We can be confident that the patterns we see in the CA plot represent the patterns that we would see if we could peer into n-dimensional space.

Factor No.

This is the number of the factor (coordinate or dimension) that is reported about on this row of the report.

Eigenvalue

This is the eigenvalue associated with this dimension. It gives a relative size (importance) of this dimension.

Individual and Cumulative Percents

The first column gives the percentage of the total of the eigenvalues accounted for by this dimension. The second column is the cumulative total of the percentage.

In ideal situations, the first two dimensions will account for over 90% of the variation. If the cumulative percentage is less than 50%, CA is not appropriate.

Bar Chart

This is a rough bar plot of the eigenvalues. It enables you to quickly note the relative size of each eigenvalue.

Plot Detail Section

```
Plot Detail Section for Rows
                                         --------   Axis1   --------   --------   Axis2   --------
Name      Quality   Mass    Inertia      Factor     COR       CTR      Factor     COR       CTR
1 SM        0.893   0.057     0.031       0.066     0.092     0.003    -0.194     0.800     0.214
2 JM        0.991   0.093     0.139      -0.259     0.526     0.084    -0.243     0.465     0.551
3 SE        1.000   0.264     0.450       0.381     0.999     0.512    -0.011     0.001     0.003
4 JE        1.000   0.456     0.308      -0.233     0.942     0.331     0.058     0.058     0.152
5 SC        0.999   0.130     0.071       0.201     0.865     0.070     0.079     0.133     0.081
```

This report provides the information you need to interpret a correspondence plot correctly. A similar report is generated for each CA plot.

This report is used as follows. First, for each axis, look down the CTR column to determine which profiles contribute highly to the axis. This is useful in finding possible interpretations of the axis. Next, look across the COR values to identify which of the axes represent the profile well. Finally, the Quality column shows how well the profile is reproduced in the subspace defined by the two axes.

Axis1, Axis 2

These are the two axes (coordinates or dimensions) that are reported on here.

Name

The name of the dimension (profile) being reported about on this line of the report.

Quality

This is the sum of the two COR values. It is the proportion of the variation in this profile that is reproduced by the two factors being reported on here.

In this example, we see that all of the profiles are above 89%. In fact, all but the SM profile are over 99%. We can feel confident that the points shown in this plot are not distorted by the projection process.

Mass

The mass (or weight) is the proportion of the whole table that is in the category represented by this row. It is the ratio of the row count to the total table count. You will find the masses also reported as percentages in the last column of the Column Profile Section.

Inertia

The inertia of the whole table is a function of the Chi-square statistic, χ^2. If

$$\chi^2 = \sum_{all\ i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where O_{ij} is the count of row *i* and column *j* of the table, E_{ij} is the value expected under the assumption of row-by-column independence, and *N* is the total table count, then the total inertia of the table is given by

$$Total\ Inertia = \frac{\chi^2}{N}$$

The inertia value reported is the proportion of the total inertia that is due to this profile.

Another way to interpret the inertia is that it is the weighted average of the Chi-square distances between the row profiles and their average profile.

Factor

The coordinate of the profile along this axis. This is the value of the row profile projected onto the line defined by this axis. It is the value that is plotted.

COR

This is the correlation between this profile and the axis. It allows you to determine which of the axes represent the profile well. This is the proportion of the variance in a profile explained by the axis.

This is the contribution of this axis to the inertia of this profile. The formula used to compute this was given earlier.

CTR

The contribution of this profile to the inertia of this axis. This is the proportion of variance in the axis accounted for by this profile. The formula used to compute this was given earlier.

Principal Coordinate Section

```
Principal Coordinate Section for Rows - Axis 1
Name    Mass   Inertial   Distance   Factor     COR      CTR    Angle   Eigenvalue
1 SM    0.057     0.031      0.047    0.066   0.092    0.003     72.3     0.000247
2 JM    0.093     0.139      0.127   -0.259   0.526    0.084     43.5     0.006254
3 SE    0.264     0.450      0.145    0.381   0.999    0.512      1.8     0.038277
4 JE    0.456     0.308      0.058   -0.233   0.942    0.331     13.9     0.024743
5 SC    0.130     0.071      0.047    0.201   0.865    0.070     21.5     0.005238

Principal Coordinate Section for Rows - Axis 2
Name    Mass   Inertial   Distance   Factor     COR      CTR    Angle   Eigenvalue
1 SM    0.057     0.031      0.047   -0.194   0.800    0.214     26.5     0.002139
2 JM    0.093     0.139      0.127   -0.243   0.465    0.551     47.0     0.005521
3 SE    0.264     0.450      0.145   -0.011   0.001    0.003     88.4     0.000030
4 JE    0.456     0.308      0.058    0.058   0.058    0.152     76.1     0.001520
5 SC    0.130     0.071      0.047    0.079   0.133    0.081     68.6     0.000807
```

This report provides all information about each axis (dimension or factor). Much of the information is duplicated in the Plot Detail Section (see above) and will not be redefined here. We will present only those items that were not defined in the last report.

Distance

This is the weighted distance of the row profile from the average row profile. It is provided for completeness.

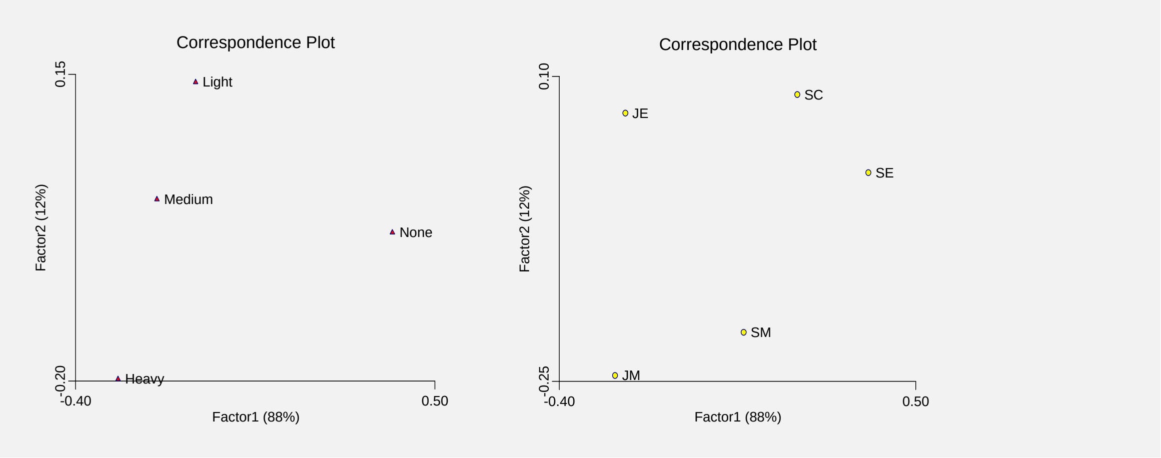
Angle

This is the angle between the axis and the profile.

Eigenvalue

If we partition the eigenvalue associated with this axis into separate parts for each profile, this is the absolution amount of the eigenvalue that is due to this profile. This value is provided more for completeness than interpretation.

Plots Section

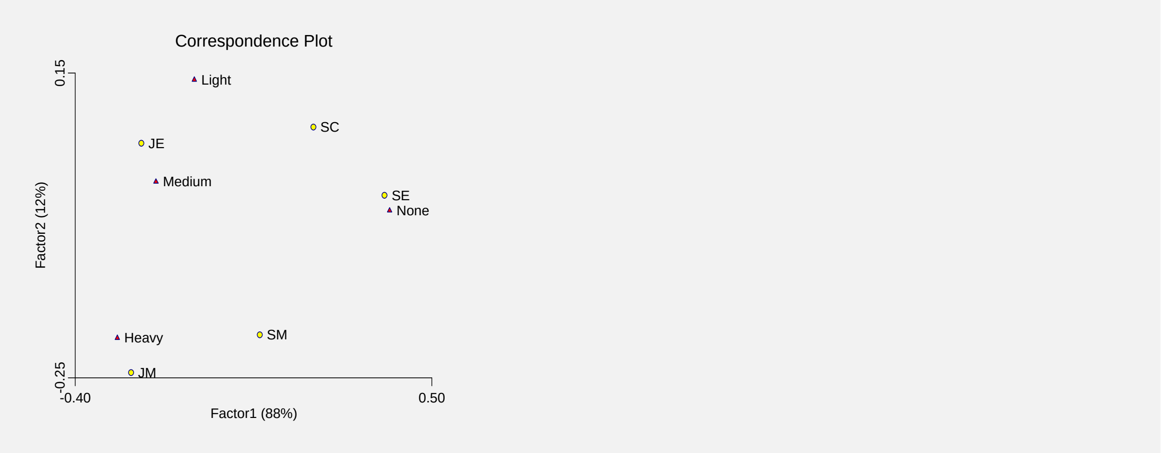


This plot is the main objective of a CA. The plot on the left shows the column profiles and the plot on the right shows the row profiles. It is important to remember that each point represents a profile projected onto the plane defined by the two axes.

Lets begin by discussing the left plot, the one presenting the four column profiles. These profiles represented the proportions belonging to each staff category. We can see from this plot that the first factor seems to separate those who smoke from those who do not. The second factor seems to separate the three types of smokers: light, medium, and heavy.

The right plot presents the five row profiles. The first axis appears to separate junior people from senior people. The second axis seems to separate managers (near the bottom) from non-managers (near the top).

Note that the distances between points on these plots are Chi-square distances between the profiles those of two points. Hence, the closer two points appear, the closer their profile patterns are to each other.



Finally, we come to the most popular CA plot in which we overlay the two plots shown above onto one plot. Extreme caution must be used when interpreting this plot. The critical point to remember is that this is a combination of two independent plots. Distances between the row profile points and column profile points are not defined. Hence, the distance between the categories SE and None (although this appear near each other on the plot) is not defined. Here's why: The point SE is a projection of the SE profile from the four dimensional space to the two-dimensional subspace defined by our axes. The point None is a projection of the None profile from the five dimensional space to the two-dimensional subspace defined by the two axes. The original spaces are different. They represent different things. In the case of the row profiles, each of the four axes represents a smoking pattern (none, light, medium, and heavy). In the case of column profiles, each of the five axes represents a staff category. The point is that the meaning of the original spaces was completely different. Their axes have completely different definitions. This is the classical apples and oranges situation. As we view a subspace from each overlaid onto one plot, what is the connection between these two subspaces?

To understand why analysts like to plot the row and column profiles on one graph, we will create a new version of our row profile plot with the addition of supplementary rows. The following table presents the data being analyzed plus addition “supplementary” rows.

CORRES1 dataset with supplementary rows

None	Light	Medium	Heavy	Staff	RowType
4	2	3	2	(SM) Senior Managers	**1**
4	3	7	4	(JM) Junior Managers	**1**
25	10	12	4	(SE) Senior Employees	**1**
18	24	33	13	(JE) Junior Employees	**1**
10	6	7	2	(SE) Secretaries	**1**
100	**0**	**0**	**0**	**N**	**2**
0	**100**	**0**	**0**	**L**	**2**
0	**0**	**100**	**0**	**M**	**2**
0	**0**	**0**	**100**	**H**	**2**

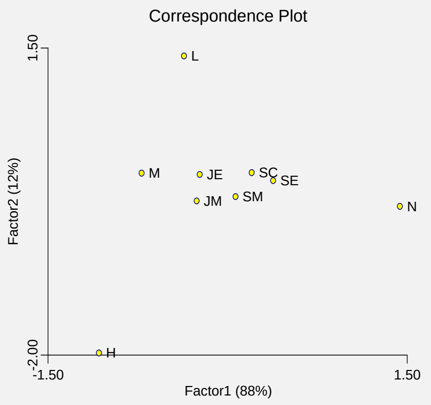
Note the addition of the RowType variable with its 1’s and 2’s. This is used to indicate which rows contain data and which rows are supplementary.

Now, if you consider the four rows that have been added to the bottom, you will see that each has a value of 100 in one column and 0 in all the rest. Hence each row represents a particular type of smoker. N represents the None group, L represents the Light group, M represents the Medium group, and H represents the Heavy group. If we could peer into four dimensional space, we would see that each of the points fall on the corresponding axis. That is, the four supplementary rows represent the four axes.

Incidentally, we could have entered “1” in each position instead of “100” since the program rescales these values.

Now let’s take a look at the row profile CA plot with these supplementary rows.

Row Profile Plot with Supplemental Axis Points



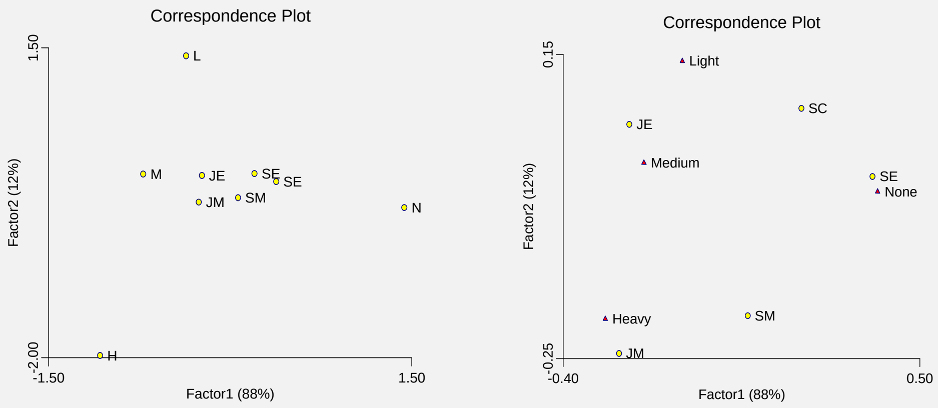
Studying this plot, we note that supplemental points (the N, L, M, and H) seem to surround the regular points. This is because each of these points defines an edge point, or vertex, to the four-dimensional space we are considering.

Now, a point near one of these supplemental points is one whose profile is similar to that point. For example, it appears that both JE and JM are near M (Medium). If you look at the Row Profile Section, you will see that they had 37.5% and 38.9% in the medium category, respectively. These are much larger than the other groups.

We see that points closer to one of our supplementary points tend to have higher than normal values for that category. However, since none of the row profiles had more than 50% in any one category, none of the profiles is right next to the vertex point (as defined by the supplementary row).

We are now ready to see why we can legitimately overlay the row profile and column profile plots. We will redisplay the last two plots side by side.

Row Profile Plot with Supplemental Axis Points and Overlaid CA Plot



The plot on the left is the regular row profile CA plot with supplementary points. Compare the relative positions of the L, M, H, and N with those of Light, Medium, Heavy, and None in the overlay plot on the right. You can see that these points maintain their relative position. They just shrink inward toward the center.

That this is the case can be shown mathematically. The message is clear. When the two plots are overlaid, the points from one space (row or column) represent the vertices of the other space, except they have been shrunken in towards the center. Hence, as we analyze the right plot from the row profile context, we must mentally move the column profile points out from the middle. That is, we must realize that each point represents the direction of the end point of that axis, but the point is not at the actual position of the end point.

Personally, I believe that interpretation is easier if you always construct two plots: one for the row profiles and another for the column profiles. The axes of each space are shown as supplementary rows (or columns). This avoids the temptation to see points from two spaces as being "near" each other.

This concludes our discussion of correspondence analysis. We again encourage you to obtain the workbook by Greenacre (1993) if you want to study this technique in more depth.

Chapter 435

Multidimensional Scaling

Introduction

Multidimensional scaling (MDS) is a technique that creates a map displaying the relative positions of a number of objects, given only a table of the distances between them. The map may consist of one, two, three, or even more dimensions. The program calculates either the metric or the non-metric solution. The table of distances is known as the *proximity* matrix. It arises either directly from experiments or indirectly as a correlation matrix.

To understand how the proximity matrix may be observed directly, consider the following marketing research example. Suppose ten subjects rate the similarities of six automobiles. That is, each subject rates the similarity of each of the fifteen possible pairs. The ratings are on a scale from 1 to 10, with "1" meaning that the cars are identical in every way and "10" meaning that the cars are as different as possible. The ratings are averaged across subjects, forming a similarity matrix. MDS provides the marketing researcher with a map (scatter plot) of the six cars that summarizes the results visually. This map shows the perceived differences between the cars.

The program offers two general methods for solving the MDS problem. The first is called *Metric*, or *Classical*, *Multidimensional Scaling (CMDS)* because it tries to reproduce the original metric or distances. The second method, called *Non-Metric Multidimensional Scaling (NMMDS)*, assumes that only the ranks of the distances are known. Hence, this method produces a map which tries to reproduce these ranks. The distances themselves are not reproduced.

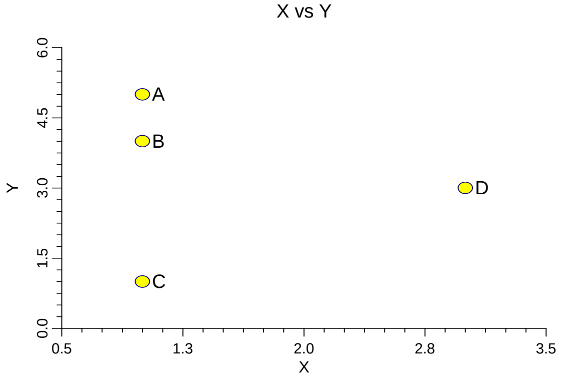
Discussion

The following example will help explain what MDS does. Consider the following set of data.

Original Data Matrix

Label	X	Y
A	1	5
B	1	4
C	1	1
D	3	3

A scatter plot of these data appears as follows:



Notice that the scatter plot lets us visually assess the distance between each pair of points. We can see that A is near B, but far from C and D. We can also see that C and D each seem to be by themselves. The actual distance between two points *i* and *j* may be computed numerically using the Euclidean distance formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}$$

where *p* is the number of dimensions (which is 2 in our example), d_{ij} is the distance, and x_{ik} is the data value of the i^{th} row and k^{th} column. This formula is an simple extension of the famous Pythagorean Theorem. Note that this formula allows for an unlimited number of dimensions. That is, although we are only plotting the points in two-dimensional space, the formula computes the distance in p-dimensional space, where p can be greater than two.

For example, the distance from A to D is calculated as follows:

$$2.82843 = \sqrt{(1-3)^2 + (5-3)^2}\ 1$$

These distances are arranged in matrix format as follows:

Computed Distance Matrix

	A	B	C	D
A	0.00000	1.00000	4.00000	2.82843
B	1.00000	0.00000	3.00000	2.23607
C	4.00000	3.00000	0.00000	2.82843
D	2.82843	2.23607	2.82843	0.00000

Note that since the distance from A to D is the same as the distance from D to A, the distance matrix is symmetric. We only need to consider half of the matrix. In the program, we only require the upper half. The final distance matrix will be:

Upper-Triangular Distance Matrix

	A	B	C	D
A	0.00000	1.00000	4.00000	2.82843
B		0.00000	3.00000	2.23607
C			0.00000	2.82843
D				0.00000

The task attempted by MDS is that given only a distance matrix, find the original data so that a map (scatter plot) of the data may be drawn.

Some of the difficulties facing MDS may be seen even in this simple example. First, as the number of objects increases, the possible number of dimensions increases as well. If you have three objects, these will at most define a two-dimensional plane. With four objects, you will usually find a three-dimensional space. And so on, with each new object adding one more possible dimension.

Also, notice that if the data are shifted in such a way that their positions relative to each other are maintained (rotated, translated, or transposed), the computed distance matrix will be the same. Hence, the distance matrix could have come from numerous sets of data.

A third challenge comes when the distances themselves are not actually known. You might only be given knowledge of their relative size.

MDS techniques have proved useful because circumstances often occur where the actual coordinates of the objects are not known, but some type of distance matrix is available. This is especially the case in psychology where people cannot draw an overall picture of a group of objects, but they can express how different individual pairs of objects are. From these pair-wise differences MDS often can provide a useful picture.

Goodness-of-Fit

As in any data analysis problem, an expression is needed to express how well a particular set of data are represented by the model that the analysis imposes. In the case of MDS, you are trying to model the distances. Hence, the most obvious choice for a goodness-of-fit statistic is one based on the differences between the actual distances and their predicted values. Such a measure is called *stress* and is calculated as values:

$$stress = \sqrt{\frac{\sum \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum d_{ij}^{\;2}}}$$

Here $\hat{d}_{ij}$ is predicted distance based on the MDS model. Note that this predicted value depends on the number of dimensions kept and the algorithm that you used (metric versus non-metric).

As you can see from this equation, MDS fits with stress values near zero are the best.

In his original paper on MDS, Kruskal (1964) gave following advise about stress values based on his experience:

Stress	Goodness-of-fit
0.200	poor
0.100	fair
0.050	good
0.025	excellent
0.000	perfect

More recent articles caution against using a table like this since acceptable values of stress depends on the quality of the distance matrix and the number of objects in that matrix.

Number of Dimensions

One of the main tasks the analyst has is determining the number of dimensions in the MDS model. Each dimension represents a different underlying factor. One of the goals of the MDS analysis is to keep the number of dimensions as small as possible. Usually, the analyst will anticipate select two or, at most, three dimensions. If more are required, you may decide that MDS is not appropriate for your data.

The usual technique is to solve the MDS problem for a number of dimension values and adopt the smallest number of dimensions that achieves a reasonably small value of stress. The program displays a simple bar chart of the stress values to aid in the selection of the number of dimensions.

Some researchers also consider the relative size of the eigenvalues that are generated during the solution process. These eigenvalues are then used to determine the number of dimensions just as they are used in factor analysis to determine the number of factors.

Proximity Measures

Proximity measures quantify how “close” two objects are. The program accepts three forms of proximity values: dissimilarities, similarities, and correlations.

Dissimilarities represent the distance between two objects. They may be measured directly, as in the distance between two towns, or approximated, as in “Bill is five points different from Joe on a ten-point scale.” MDS algorithms use the dissimilarities directly. A dissimilarity matrix is symmetrical.

Similarities represent how close (in some sense) two objects are. The program lets you enter a similarity measure for each pair of objects. Similarities must obey the rule: $similarity_{ij}$ <= $similarity_{ii}$ and $similarity_{jj}$ for all *i* and *j*. Similarity matrices are symmetrical.

Similarities are converted to dissimilarities using the formula:

$$d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$$

where d_{ij} represents a dissimilarity and s_{ij} represents a similarity.

When your data consists of standard measures rather than dissimilarities or similarities, you can create a dissimilarity matrix by first creating the correlation matrix and then using the above formula to convert the correlations to dissimilarities. The program automatically calculates pair-wise correlations for the variable you specify.

Comparison of Metric and Non-Metric MDS

Although the computations are simpler for the metric method than for the non-metric method, both seem to yield similar results when applied to well-known examples. When you have true distance data, the classical method yields a solution that can be used directly. When you only have dissimilarities, the non-metric approach is somewhat more appealing.

Metric MDS

Classical MDS procedures stem back to Torgerson (1952), who was one of the pioneers of the technique. His algorithm is explained next.

Suppose a distance matrix ***D*** approximates the inter-point distances of a configuration of points ***X*** in a space of low dimensionality *p* (usually *p* = 1, 2, or 3). That is, the elements of **D**, denoted d_{ij}, may be calculated from ***X*** using the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2}$$

The steps in the classical MDS algorithm are as follows:

1. From ***D*** calculate ***A*** = $\left\{-\tfrac{1}{2}d_{ij}^{\;2}\right\}$.

2. From ***A*** calculate ***B*** = $\left\{a_{ij} - a_{i.} - a_{.j} + a_{..}\right\}$, where $a_{i.}$ is the average of all a_{ij} across *j*.

3. Find the *p* largest eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ of ***B*** and corresponding eigenvectors $L = \left(L_{(1)}, L_{(2)}, \cdots L_{(p)}\right)$ which are normalized so that $L'_{(i)}L_{(i)} = \lambda_i$. (We are assuming that *p* is selected so that the eigenvalues are all relatively large and positive.)

4. The coordinates of the objects are the *rows* of ***L***.

The classical solution is optimal in the least-squares sense. That is, when a direct solution is possible (i.e., when ***D*** is truly a Euclidean distance matrix), the solution, ***L,*** minimizes the sum of squared differences between the actual d_{ij}'s (elements of ***D***) and the $\hat{d}_{ij}$'s based on ***L***. Another way of saving this is that it minimizes the value of *stress*, where *stress* was defined above.

Non-Metric MDS

Implicit in the above is the assumption that there is a true configuration in *p* dimensions, i.e., that ***D*** is a distance matrix. Often, however, it is more realistic to assume a less stringent relationship between the observed distances (or dissimilarities) d_{ij} and the true distances, denoted δ_{ij}. That is, suppose we assume that

$$d_{ij} = f\left(\delta_{ij} + e_{ij}\right)$$

where e_{ij} represents errors of measurements, distortions, etc. Also, we assume that *f(x)* is an unknown, monotonically-increasing function.

For this model, the only information we can use is the rank order of the d_{ij}. Usually, this approach is used when ***D*** is simply a dissimilarity matrix rather than a true distance matrix. This assumption is often more plausible in practical situations.

An algorithm to produce a solution based only on the rank order information was provided by Kruskal (1964). It is involved, so we will not reproduce it here. We note that Kruskal's algorithm minimizes stress.

Kruskal's algorithm uses steepest descent to find a local minimum from a given starting configuration. The choice of the starting configuration is important to finding the global rather than a local minimum. Many authors recommend using the solution of the metric MDS as the starting configuration. This is the default starting configuration in this program. You may also select several random starting configurations and compare the resulting stress values.

Data Structure

The data are may be entered in three formats. The first format is the standard row-column format from which the correlations have be calculated. The MDS conducted on the correlations in an attempt to determine which of the variables are similar. The second format is the upper-triangular portion of a distance matrix. The third format is the upper-triangular portion of a similarity matrix.

An example of an upper-triangular distance matrix is contained in the MDS2 database. We suggest that you open this database now so that you can follow along with the example.

MDS2 dataset

Sport	Hockey	Football	Basketball	Tennis	Golf	Croquet
Hockey	0	2	3	4	5	5
Football		0	3	5	6	5
Basketball			0	5	4	6
Tennis				0	4	3
Golf					0	2
Croquet						0

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Input Variables

Input Variables

Specify the variables containing the upper-triangular distance, or similarity, matrix or the variables from which the correlation matrix is to be calculated. Note that filter may not be used with the upper-triangular matrices.

Input Data Type

You can specify which of the three possible types of input data you have: *dissimilarities*, *similarities*, or *correlation*.

Dissimilarities and similarities signal the program that an upper-triangular matrix is to be read in. When the correlation option is selected, the program generates a correlation matrix from the standard variable by observation data format that is used throughout the rest of the program.

Options

Solution Type

This option specifies whether you want the *Metric* or the *Non-Metric* algorithm used.

Options – Dimensions Reported

Minimum and Maximum Dimensions

These options specify a range of dimensions to be reported on. The MDS algorithm will be run on each of these dimensions and the stress value will be calculated.

Dimensions Used

This option specifies the number of dimensions on which all results are based. This is the number of dimensions that will be plotted.

Options – Non-Metric MDS Options

Initial Configuration

This option specifies which starting configuration you want when using NMMDS. You can specify either *Random* or *Metric*. We suggest the Metric option since it has been more widely recommend. You should only use the Random option when the Metric option fails to produce good results.

Max Iterations/Dimension

This option specifies the maximum number of iterations run during the solution of a NMMDS problem for a particular number of dimensions.

Max Iterations/Min Phase

This option specifies the maximum number of iterations run during minimization phase.

Min Stress Value

This option specifies the value of stress that must achieved in order to stop iterations during a run of NMMDS.

Min Stress Change

This option specifies the change in stress from one iteration to the next that must achieved in order to stop iterations during a run of NMMDS.

Min Gradient Sum

This option specifies the value of the gradient sum that must achieved in order to stop iterations during a run of NMMDS.

Reports Tab

The following options control which reports are displayed.

Select Reports

Eigenvalue Report - Dissimilarity Report

Specify whether to display the indicated report.

Select Plots

MDS Map – Dissimilarity

Specify whether to display the indicated 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 select whether to display variable names, variable labels, or both.

Plot Options

Size of Plots

This option controls the size of the plots that are displayed. You can select *small*, *medium*, or *large*. *Medium* and *large* are displayed one per line, while *small* are displayed two per line.

MDS Map Plot Tab and Dissimilarities Plot Tab

These options control the attributes of the two plots.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis 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

Designate the plot symbol used for plotting rows.

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.

Storage Tab

Optionally specify variables to contain the coordinate values generated by the program.

Data Storage Variables

Coordinate Values

A list of variables into which the coordinate values (the values that are plotted) are stored.

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 – Metric Multidimensional Scaling

This section presents an example of how to run an analysis of the data contained in the MDS2 database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Multidimensional Scaling window.

1 Open the MDS2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MDS2.s0**.
- Click **Open**.

2 Open the Multidimensional Scaling window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Multidimensional Scaling**. The Multidimensional Scaling 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 Multidimensional Scaling window, select the **Variables tab**.
- Double-click in the **Input Variables** box. This will bring up the variable selection window.
- Select **Hockey** to **Croquet** from the list of variables and then click **Ok**. "Hockey-Croquet" will appear in the Input Variables box.

4 Specify the plot.

- Select the **MDS Map Plot tab**.
- Enter **-4** in the **Horizontal - Minimum** box.
- Enter **4** in the **Horizontal - Maximum** box.

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).

Eigenvalue Section

```
Eigenvalue Section
Dim                               Individual    Cumulative
No.        Eigenvalue                Percent       Percent        Bar Chart
1               30.73                  54.28         54.28        |IIIIIIIIIIIIIIIIIIIIIIIIIII
2 (Used)        12.85                  22.69         76.97        |IIIIIIIIIII
3                6.38                  11.27         88.24        |IIIIII
4                1.68                   2.97         91.21        |I
5                0.00                   0.00         91.21        |
6               -4.98                   8.79        100.00        |IIII
Total           56.62
```

This report is produced by CMDS.

In this particular example, the first two dimensions account for 77% of the variation while the first three dimensions account for 88%. We would probably use two or perhaps three dimensions.

Eigenvalues

These are the eigenvalues found during CMDS. The eigenvalues are helpful in determining the number of dimensions that are necessary to represent the dissimilarity matrix accurately. As in factor analysis, the task is to select enough dimensions to approximate the data, but few enough to keep the interpretation simple. The eigenvalue report allows you to quickly determine the impact of each new dimension.

In MDS, some of the eigenvalues can be negative. Do not keep these dimensions. The basic rule is to find the number of relatively large, positive eigenvalues. This report provides a bar graph and percentages to help you determine the number of dimensions.

Individual and Cumulative Percents

The first column gives the percentage of the total of the absolute value of the eigenvalues accounted for by this dimension. The second column is the cumulative total of the percentage.

Bar Chart

This is a rough bar plot of the eigenvalues. It enables you to quickly note the relative size of each eigenvalue. Many authors recommend it as a method of determining how many dimensions to retain.

Fit Summary Section

```
Fit Summary Section (Metric Solution)

No.          Squared                            Pseudo
Dim's    Differences          Stress         R-Squared
1          37.105982        0.364035              0.00
2           6.947666        0.157522             70.73
3           2.413305        0.092838             89.83
4           2.468686        0.093897             89.60

Number of Dissimilarities                                  15
Mean of Dissimilarities                                    4.133333
Sum of Squared Dissimilarities                             280.000000
Mean Corrected Sum of Squared Dissimilarities              23.733333
```

This report provides information useful in determining the number of dimensions that are necessary and assessing the goodness-of-fit of the CMDS model.

No. Dim's

The number of dimensions used in calculating this row of statistics.

Squared Differences

The sum of the squared differences between the actual dissimilarity values and those predicted by the solution.

Stress

This is the value of the stress goodness-of-fit statistic. It is equal to the square root of the Squared Differences divided by the square root of the Sum of the Squared Dissimilarities. It is one of the most popular measures of accuracy of the fit. A value below 0.05 is acceptable. A value below 0.01 is considered good.

Pseudo R-Squared

This is an index, similar to the R-squared value in regression analysis, which indicates what percentage of the sum of squared dissimilarities (corrected for the mean) is accounted for by this number of dimensions. A value above 80% is hoped for.

Number of Dissimilarities

This is the number of dissimilarity values.

Mean of Dissimilarities

This is the mean of the dissimilarity values

Sum of Squared Dissimilarities

This is the sum of the squared dissimilarities. It is the denominator of the stress statistic.

Mean Corrected Sum of Squared Dissimilarities

This is the sum of the squared dissimilarities about their mean. It is the denominator of the Pseudo R-Squared statistic.

Solution Section

```
Solution Section

Variables       Dim1        Dim2        Dim3        Dim4
Hockey        1.9301     -0.6756      0.3818      1.0441
Football      2.6179     -1.1281     -1.1303     -0.4680
Basketball    2.1119      2.0914      0.4168     -0.4032
Tennis       -1.4786     -1.3608      1.8070     -0.3940
Golf         -2.3836      2.0059     -0.2743      0.2351
Croquet      -2.7976     -0.9328     -1.2011     -0.0140
```

This report presents the solution of the MDS procedure. These are the data that are plotted in the MDS map. They have been scaled so that the sum of squares for each column is equal to the eigenvalue for that dimension.

Note that these data were constructed so that the distance between two rows is close to the original dissimilarity value.

Although some interpretation of these numbers may be made directly, usually the data are displayed on scatter plots.

Dissimilarity Section

```
Dissimilarity Section
                                    Actual   Predicted      Actual     Percent
Row            Column        Dissimilarity Dissimilarity Difference  Difference
1 Hockey       2 Football        2.000000    0.823324    1.176676       58.83
5 Golf         6 Croquet         2.000000    2.967759   -0.967759      -48.39
2 Football     3 Basketball      3.000000    3.259039   -0.259039       -8.63
4 Tennis       6 Croquet         3.000000    1.386718    1.613282       53.78
1 Hockey       3 Basketball      3.000000    2.772988    0.227012        7.57
1 Hockey       4 Tennis          4.000000    3.476854    0.523146       13.08
3 Basketball   5 Golf            4.000000    4.496329   -0.496329      -12.41
4 Tennis       5 Golf            4.000000    3.486229    0.513771       12.84
2 Football     6 Croquet         5.000000    5.419049   -0.419049       -8.38
3 Basketball   4 Tennis          5.000000    4.980893    0.019107        0.38
2 Football     4 Tennis          5.000000    4.103106    0.896894       17.94
1 Hockey       6 Croquet         5.000000    4.734691    0.265309        5.31
1 Hockey       5 Golf            5.000000    5.079231   -0.079231       -1.58
3 Basketball   6 Croquet         6.000000    5.766223    0.233777        3.90
2 Football     5 Golf            6.000000    5.902321    0.097679        1.63

Dimensions                        2
Sum of Squared Dissimilarities    280.000000
Sum of Squared Differences        6.947666
Stress                            0.157522
Pseudo R-Squared                  70.726127
```

You might think of this as a residual analysis report since it highlights the differences between the actual and the predicted dissimilarities. It will let you focus on those dissimilarities that are not fit well by the model.

Row

The variable associated with this row of the dissimilarity matrix.

Column

The variable associated with this column of the dissimilarity matrix.

Actual Dissimilarity

The value from the input (or calculated) dissimilarity matrix for this row and column.

Predicted Dissimilarity

The predicted dissimilarity value based on the number of dimensions that you have selected.

Actual Difference

The Actual Dissimilarity minus the Predicted Dissimilarity. This value shows the size of the error in predicting this element of the dissimilarity matrix.

Percent Difference

The percentage the Actual Difference is of the Actual Dissimilarity. This value highlights the outliers--those dissimilarities that are not fit well by the MDS model.

Dimensions

The number of dimensions used in calculating the statistics.

Sum of Squared Dissimilarities

This is the sum of the squared dissimilarities. It is the denominator of the stress statistic.

Sum of Squared Differences

This is the sum of the squared differences. It is the numerator of the stress statistic.

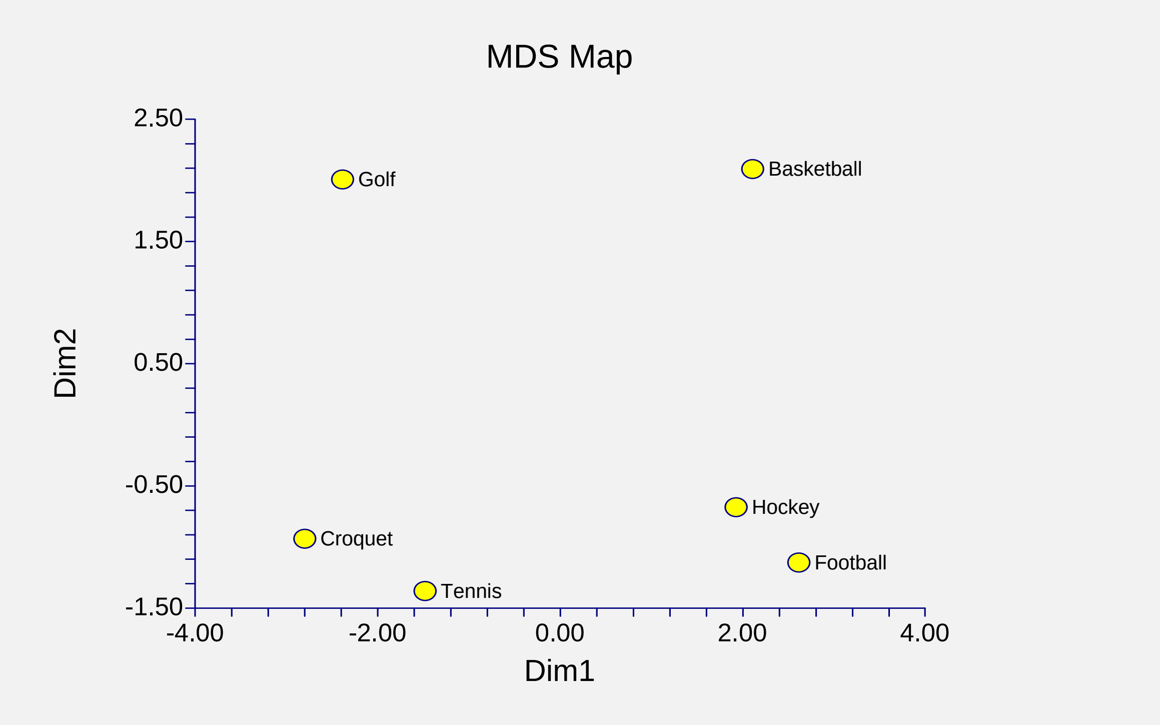
Stress

This is the value of the stress goodness-of-fit statistic. It is equal to the Squared Differences divided by the Sum of the Squared Dissimilarities. It is one of the most popular measures of accuracy of the fit. A value below 0.05 is acceptable. A value below 0.01 is considered good.

Pseudo R-Squared

This is an index, similar to the R-squared value in regression analysis, which indicates what percentage of the sum of squared dissimilarities (corrected for the mean) is accounted for by this number of dimensions. A value above 80% is hoped for.

MDS Map



This plot is the chief objective of an MDS analysis. It is often referred to as the MDS *map*. It allows you to interpret the dissimilarity matrix on a two-dimensional scatter plot.

There is no real orientation to this map. You could legitimately rotate the values around the plot's center. The main characteristics of interest are the relative positions of the points and any clusters that are apparent.

In this example, we see that the respondents considered hockey and football to be similar. They also considered croquet and tennis to be quite similar. Football appears quite different from golf. And so on. Notice how easy it is to draw conclusions about the similarities among the sports.

A second task of the MDS analyst is to find the underlying factors that respondents used when they created these dissimilarities. For example, a vertical line down the center of the plot would divide team sports on the right from individual sports on the left. We would hypothesize this as one interpretation of the Dim1 (horizontal) axis.

Example 2 – Non-Metric Multidimensional Scaling

This section presents an example of how to run an analysis of the data contained in the MDS2 database using NMMDS.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Multidimensional Scaling window.

1 Open the MDS2 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **MDS2.s0**.
- Click **Open**.

2 Open the Multidimensional Scaling window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Multidimensional Scaling**. The Multidimensional Scaling procedure will be displayed.

3 Specify the variables.

- On the Multidimensional Scaling window, select the **Variables tab**.
- Double-click in the **Input Variables** box. This will bring up the variable selection window.
- Select **Hockey** to **Croquet** from the list of variables and then click **Ok**. “Hockey-Croquet” will appear in the Input Variables box.
- Enter **Non-Metric** in the **Solution Type** 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).

Eigenvalue Section

```
Eigenvalue Section
Dim                            Individual  Cumulative
No.        Eigenvalue             Percent     Percent         Bar Chart
1               30.73               54.28       54.28         |IIIIIIIIIIIIIIIIIIIIIIIIIII
2 (Used)        12.85               22.69       76.97         |IIIIIIIIIII
3                6.38               11.27       88.24         |IIIIII
4                1.68                2.97       91.21         |I
5                0.00                0.00       91.21         |
6               -4.98                8.79      100.00         |IIII
Total           56.62
```

This report is produced by CMDS which was used as the starting configuration. Its definitions were given above and will not be repeated here.

Non-Metric Iteration Summary Section

```
Non-Metric Iteration Summary Section
No.     Percent Rank                   Why               Bar Chart
Dim's      Maintained        Stress    Terminated        of Stress
1               57.14      0.212628    Stress Change     |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII<<I
2               64.29      0.052276    Stress Change     |IIIIIIIIIIIIIIIIIIIIIIIIII
3               78.57      0.000344    Max Iterations    |
4               64.29      0.000003    Min Stress        |
```

This report provides information about the number of dimensions that are necessary and the goodness-of-fit of the solution.

No. Dim’s

The number of dimensions used in calculating this row of statistics.

Percent Rank Maintained

The non-metric solution tries to maintain the rank ordering of the dissimilarities. This is the percentage of the dissimilarities whose rank order was maintained. The higher this value is, the better the quality of the solution.

Stress

Defined earlier, this is one of the most popular measures of accuracy of the fit. A value below 0.05 is acceptable. A value below 0.01 is considered good.

Why Terminated

This field explains which stopping rule caused the iterative procedure to stop. This is important to watch since the solution is not optimal if the maximum iterations were reached before the algorithm converged. When this happens, you should change some of the iteration control parameters, especially the number of iterations.

Bar Chart of Stress

This column graphically portrays the stress values. You want to choose the fewest number of dimensions that give you a small stress value.

Solution Section

```
Solution Section

Variables       Dim1         Dim2
Hockey        0.3306      -0.1495
Football      0.4234      -0.0227
Basketball    0.2674       0.2837
Tennis       -0.2542      -0.3538
Golf         -0.3473       0.2227
Croquet      -0.4199       0.0196
```

This report presents the final configuration of the NMMDS procedure. These are the data that are plotted in the MDS map.

Note that these data were not constructed so that the distance between two rows is close to the original dissimilarity value. Instead, the non-metric solution attempts to maintain the same rank ordering of the calculated distances as occur in the original dissimilarity matrix. Although some interpretation of these numbers may be made directly, usually the data are displayed on scatter plots.

Dissimilarity Section

```
Dissimilarity Section
                                     Actual     Predicted
Row            Column          Dissimilarity  Dissimilarity
1 Hockey       2 Football           2.000000       0.157123
5 Golf         6 Croquet            2.000000       0.215646
2 Football     3 Basketball         3.000000       0.343893
4 Tennis       6 Croquet            3.000000       0.408445
1 Hockey       3 Basketball         3.000000       0.437798
1 Hockey       4 Tennis             4.000000       0.619391
3 Basketball   5 Golf               4.000000       0.617695
4 Tennis       5 Golf               4.000000       0.583869
2 Football     6 Croquet            5.000000       0.844336
3 Basketball   4 Tennis             5.000000       0.823650
2 Football     4 Tennis             5.000000       0.754122
1 Hockey       6 Croquet            5.000000       0.769237
1 Hockey       5 Golf               5.000000       0.773283
3 Basketball   6 Croquet            6.000000       0.736258
2 Football     5 Golf               6.000000       0.808819
```

You might think of this as a residual analysis report since it highlights the differences between the actual and the predicted dissimilarities. It will let you focus on those dissimilarities that are not fit well by the model.

This report presents the details of how well the rank ordering of the dissimilarity values is preserved in the final configuration. Note that the predicted values are quite different from the actual values since all the algorithm was attempting to do was maintain the ordering.

Row

The variable associated with this row of the dissimilarity matrix.

Column

The variable associated with this column of the dissimilarity matrix.

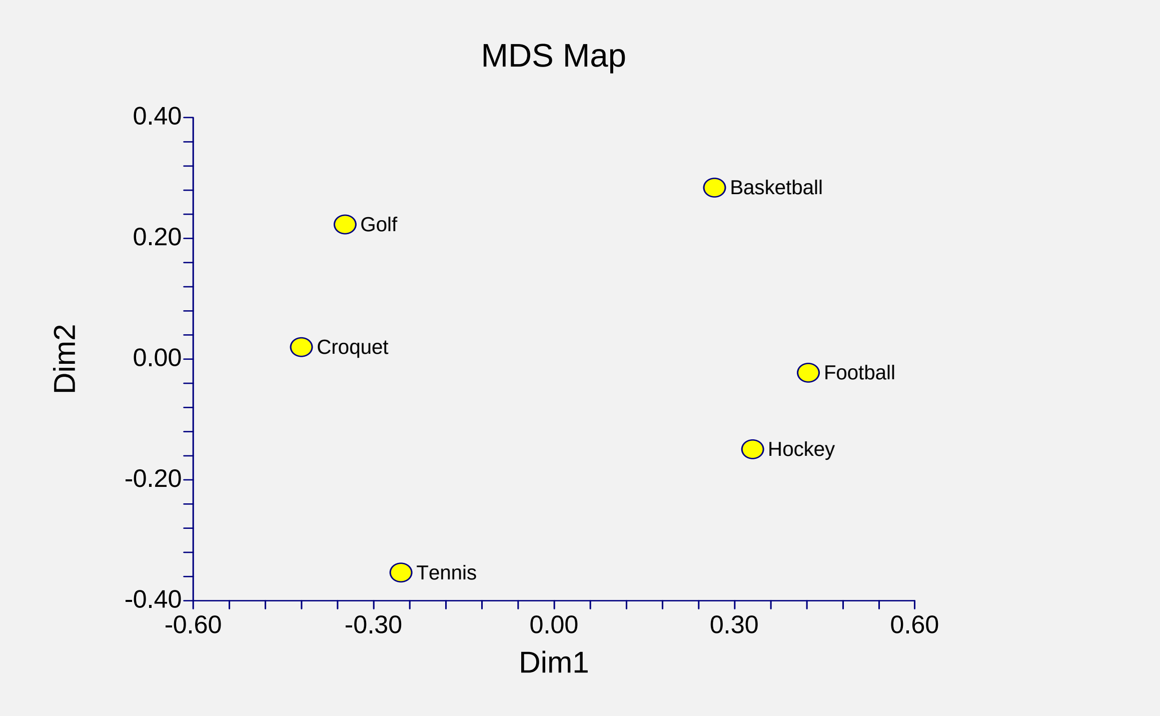
Actual Dissimilarity

The value from the input (or calculated) dissimilarity matrix for this row and column.

Predicted Dissimilarity

The predicted dissimilarity value based on the number of dimensions that you have selected. Note that this is not predicting the actual dissimilarity value, but some unknown function of the dissimilarity value. It is not usually necessary to determine the function. We are mainly interested in how well the ordering of the actual values is maintained by these predicted values.

MDS Map



This plot is the chief objective of an MDS analysis. It is often referred to as the MDS *map*. It allows you to interpret the dissimilarity matrix on a two-dimensional scatter plot.

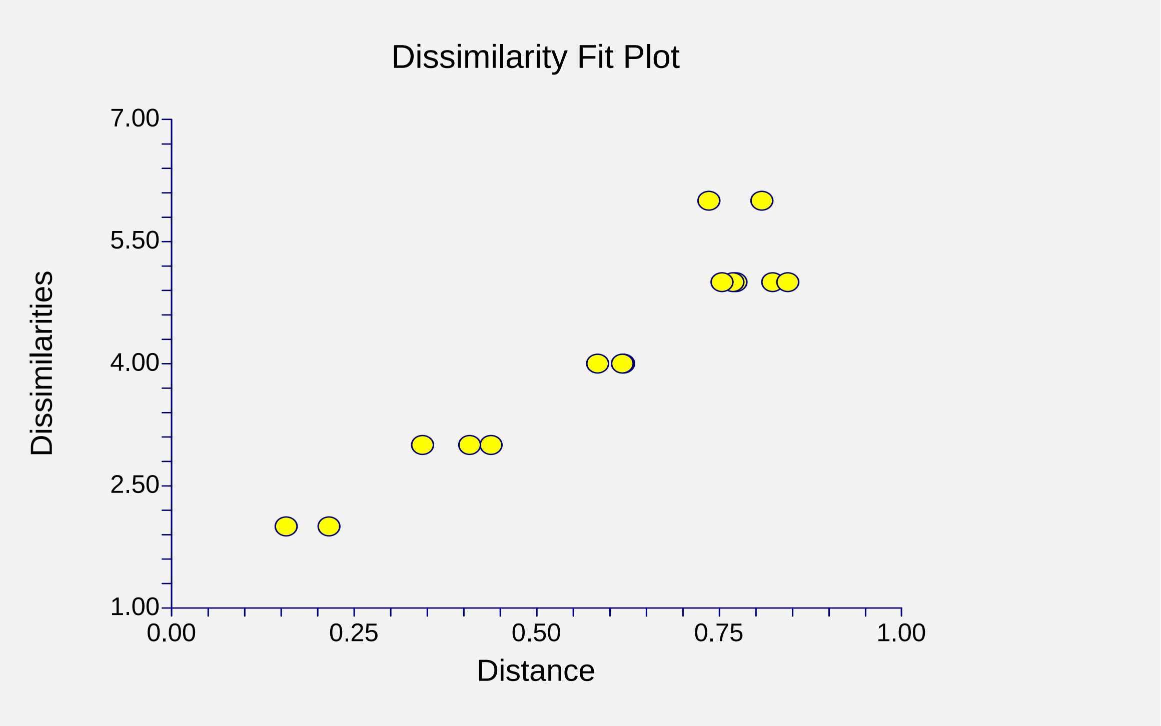
There is no real orientation to this map. You could legitimately rotate the values around the plot's center. The main characteristics of interest are the relative positions of the points and any clusters that are apparent.

In this example, we see that the respondents considered hockey and football to be similar. They also considered croquet and golf to be similar. Football appears quite different from croquet. And so on. Notice how easy it is to draw conclusions about the similarities among the sports.

A second task of the MDS analyst is to find the underlying factors that respondents used when they created these dissimilarities. For example, a vertical line down the center of the plot would divide team sports on the right from individual sports on the left. We might hypothesize this as one interpretation of the Dim1 (horizontal) axis.

It is interesting to compare this map with the map produced by the metric solution. The main difference appears to be that golf and croquet are now much closer together (as they were rated in the original data). Again, football and basketball appear to be closer together in this plot as we might expect from the original data. In this case, the NMMDS map appears to be more accurate than the CMDS map. This is as we might expect since, the NMMDS procedure refined the CMDS map.

Dissimilarity Fit Plot



This graph plots the dissimilarity values on the vertical axis against the predicted dissimilarity values on the horizontal axis. The caliber of the solution depends upon this plot showing an upward-sloping trend. If the solution was perfect, then as you move across the plot from left to right, you would never go down from one point to the next.

We notice in this case that the solution confuses the large distances. This may be due to the large number of ties in this area (look at the Dissimilarity Section to see all the 5’s and 6’s).

Chapter 440

Discriminant Analysis

Introduction

Discriminant Analysis finds a set of prediction equations based on independent variables that are used to classify individuals into groups. There are two possible objectives in a discriminant analysis: finding a predictive equation for classifying new individuals or interpreting the predictive equation to better understand the relationships that may exist among the variables.

In many ways, discriminant analysis parallels multiple regression analysis. The main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. The methodology used to complete a discriminant analysis is similar to regression analysis. You plot each independent variable versus the group variable. You often go through a variable selection phase to determine which independent variables are beneficial. You conduct a residual analysis to determine the accuracy of the discriminant equations.

The mathematics of discriminant analysis are related very closely to the one-way MANOVA. In fact, the roles of the variables are simply reversed. The classification (factor) variable in the MANOVA becomes the dependent variable in discriminant analysis. The dependent variables in the MANOVA become the independent variables in the discriminant analysis.

Technical Details

Suppose you have data for K groups, with N_k observations per group. Let N represent the total number of observations. Each observation consists of the measurements of p variables. The i^{th} observation is represented by X_{ki}. Let M represent the vector of means of these variables across all groups and M_k the vector of means of observations in the k^{th} group.

Define three sums of squares and cross products matrices, S_T, S_W, and S_A, as follows

$$S_T = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(X_{ki} - M\right)\left(X_{ki} - M\right)'$$

$$S_W = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\left(X_{ki} - M_k\right)\left(X_{ki} - M_k\right)'$$

$$S_A = S_T - S_W$$

Next, define two degrees of freedom values, *df1* and *df2*:

$$df1 = K - 1$$

$$df2 = N - K$$

A discriminant function is a weighted average of the values of the independent variables. The weights are selected so that the resulting weighted average separates the observations into the groups. High values of the average come from one group, low values of the average come from another group. The problem reduces to one of finding the weights which, when applied to the data, best discriminate among groups according to some criterion. The solution reduces to finding the eigenvectors, *V*, of $S_W^{-1} S_A$. The canonical coefficients are the elements of these eigenvectors.

A goodness-of-fit parameter, Wilks' lambda, is defined as follows:

$$\Lambda = \frac{|S_W|}{|S_T|} = \prod_{j=1}^{m} \frac{1}{1+\lambda_j}$$

where λ_j is the *jth* eigenvalue corresponding to the eigenvector described above and *m* is the minimum of *K-1* and *p*.

The canonical correlation between the j^{th} discriminant function and the independent variables is related to these eigenvalues as follows:

$$r_{c_j} = \sqrt{\frac{\lambda_j}{1+\lambda_j}}$$

Various other matrices are often considered during a discriminant analysis.

The overall covariance matrix, *T*, is given by:

$$T = \left(\frac{1}{N-1}\right) S_T$$

The within-group covariance matrix, *W*, is given by:

$$W = \left(\frac{1}{N-K}\right) S_W$$

The among-group (or between-group) covariance matrix, *A*, is given by:

$$A = \left(\frac{1}{K-1}\right) S_A$$

The linear discriminant functions are defined as:

$$LDF_k = W^{-1} M_k$$

The standardized canonical coefficients are given by:

$$v_{ij} \sqrt{w_{ij}}$$

where v_{ij} are the elements of *V* and w_{ij} are the elements of *W*.

The correlations between the independent variables and the canonical variates are given by:

$$Corr_{jk} = \frac{1}{\sqrt{w_{jj}}} \sum_{i=1}^{p} v_{ik} w_{ji}$$

Discriminant Analysis Checklist

Tabachnick (1989) provides the following checklist for conducting a discriminant analysis. We suggest that you consider these issues and guidelines carefully.

Unequal Group Size and Missing Data

You should begin by screening your data. Pay particular attention to patterns of missing values. When using discriminant analysis, you should have more observations per group than you have independent variables. If you do not, there is a good chance that your results cannot be generalized, and future classifications based on your analysis will be inaccurate.

Unequal group size does not influence the direct solution of the discriminant analysis problem. However, unequal group size can cause subtle changes during the classification phase. Normally, the sampling frequency of each group (the proportion of the total sample that belongs to a particular group) is used during the classification stage. If the relative group sample sizes are not representative of their sizes in the overall population, the classification procedure will be erroneous. (You can make appropriate adjustments to prevent these erroneous classifications by adjusting the prior probabilities.)

NCSS ignores rows with missing values. If it appears that most missing values occur in one or two variables, you might want to leave these out of the analysis in order to obtain more data and hence more accuracy.

Multivariate Normality and Outliers

Discriminant analysis does not make the strong normality assumptions that MANOVA does because the emphasis is on classification. A sample size of at least twenty observations in the smallest group is usually adequate to ensure robustness of any inferential tests that may be made.

Outliers can cause severe problems that even the robustness of discriminant analysis will not overcome. You should screen your data carefully for outliers using the various univariate and multivariate normality tests and plots to determine if the normality assumption is reasonable. *You should perform these tests on one group at a time.*

Homogeneity of Covariance Matrices

Discriminant analysis makes the assumption that the group covariance matrices are equal. This assumption may be tested with Box’s M test in the Equality of Covariances procedure or looking for equal slopes in the Probability Plots. If the covariance matrices appear to be grossly different, you should take some corrective action. Although the inferential part of the analysis is robust, the classification of new individuals is not. These will tend to be classified into the groups with larger covariances. Corrective action usually includes the close screening for outliers and the use of variance-stabilizing transformations such as the logarithm.

Linearity

Discriminant analysis assumes linear relations among the independent variables. You should study scatter plots of each pair of independent variables, using a different color for each group.

Look carefully for curvilinear patterns and for outliers. The occurrence of a curvilinear relationship will reduce the power and the discriminating ability of the discriminant equation.

Multicollinearity and Singularity

Multicollinearity occurs when one predictor variable is almost a weighted average of the others. This collinearity will only show up when the data are considered one group at a time. Forms of multicollinearity may show up when you have very small group sample sizes (when the number of observations is less than the number of variables). In this case, you must reduce the number of independent variables.

Multicollinearity is easily controlled for during the variable selection phase. You should only include variables that show an R^2 with other X's of less than 0.99.

See the chapter on Multiple Regression for a more complete discussion of multicollinearity.

Data Structure

The data given in the table below are the first eight rows (out of the 150 in the database) of the famous "iris data" published by Fisher (1936). These data are measurements in millimeters of sepal length, sepal width, petal length, and petal width of fifty plants for each of three varieties of iris: (1) Iris setosa, (2) Iris versicolor, and (3) Iris virginica. Note that Iris versicolor is a polyplid hybrid of the two other species. Iris setosa is a diploid species with 38 chromosomes, Iris virginica is a tetraploid, and Iris versicolor is a hexaploid with 108 chromosomes.

Discriminant analysis finds a set of prediction equations, based on sepal and petal measurements, that classify additional irises into one of these three varieties. Here Iris is the dependent variable, while SepalLength, SepalWidth, PetalLength, and PetalWidth are the independent variables.

FISHER dataset (subset)

SepalLength	SepalWidth	PetalLength	PetalWidth	Iris
50	33	14	2	1
64	28	56	22	3
65	28	46	15	2
67	31	56	24	3
63	28	51	15	3
46	34	14	3	1
69	31	51	23	3
62	22	45	15	2

Missing Values

If missing values are found in any of the independent variables being used, the row is omitted. If they occur only in the dependent (categorical) variable, the row is not used during the calculation of the prediction equations, but a predicted group (and scores) is calculated. This allows you to classify new observations.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Group Variable

Y: Group Variable

This is the dependent, Y, grouping, or classification variable. It must be discrete in nature. That means it can only have a few unique values. Each unique value represents a separate group of individuals. The values may be text or numeric.

Independent Variables

X's: Independent Variable

These are the set of independent variables. Although the probability statements used in discriminant analysis assume that these variables are continuous (and normal), the technique is robust enough that it can tolerate a few discrete variables (assuming that they are numeric).

Estimation Options

Estimation Method

This option designates the classification method used.

- **Linear Discriminant Function**

  Signifies that you want to classify using the linear-discriminant functions (assumes multivariate normality with equal covariance matrices). This is the most popular technique.

- **Regression Coefficients**

  Indicates that you want to classify using multiple regression coefficients (no special assumptions). This method develops a multiple regression equation for each group, ignoring the discrete nature of the dependent variable. Each of the dependent variables is constructed by using a 1 if a row is in the group and a 0 if it is not.

Estimation Options – Linear Discriminant Function Options

Prior Probabilities

Allows you to specify the prior probabilities for linear-discriminant classification. If this option is left blank, the prior probabilities are assumed equal. This option is not used by the regression classification method. The numbers should be separated by blanks or commas. They will be adjusted so that they sum to one. For example, you could use "4 4 2" or "2 2 1" when you have three groups whose population proportions are 0.4, 0.4, and 0.2, respectively.

Variable Selection Options

Variable Selection

This option specifies whether a stepwise variable-selection phase is conducted.

- **None**

  All independent variables are used in the analysis. No variable selection is conducted.

- **Stepwise**

  A stepwise variable-selection is performed using the “in” and “out” probabilities specified next.

Variable Selection Options – Stepwise Selection Options

Maximum Iterations

(Automatic Selection only). This options sets the maximum number of steps that are used in the stepwise procedure. It is possible to set the above probabilities so that one or two variables are alternately entered and removed, over and over. We call this an infinite loop. To avoid such an occurrence, you can set the maximum number of steps permitted.

Probability Enter

(Stepwise only.) This option sets the probability level for tests used to determine if a variable may be brought into the discriminant equation. At each step, the variable (not in the equation) with the smallest probability level below this cutoff value is entered.

Probability Removed

(Stepwise only.) This option sets the probability level for tests used to determine if a variable should be removed from the discriminant equation. At each step, the variable (in the equation) with the largest probability level above this cutoff value is removed.

Reports Tab

The following options control the format of the reports.

Select Reports

Group Means – Canonical Scores

These options let you specify which reports you want displayed. This is especially useful if you have a lot of data, since some of the reports produce a separate report row for each data row. You may want to omit these reports. The row-wise reports are Predicted Classification, Linear Discriminant Function Scores, Regression Scores, and Canonical Scores.

Select Plots

LD-Score Plots - Canonical-Score Plots

These options let you specify which plots you want displayed.

Report Options

Report Format

This option lets you specify whether to output a “brief” or “verbose” report during the variable-selection phase. Normally, you would only select “verbose” if you have fewer than ten independent variables.

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.

Plot Options

Legend

Specify whether you want to view a legend of the group values.

Legend Text

Specifies the legend title. Note that if you use the {G} symbol, this value will automatically be replaced with the Group Variable’s name or label.

LD-Score Plot, Reg-Score Plot, and Canonical-Score Plot Tabs

These panels specify the pair-wise plots of the scores generated for each set of functions. A separate function is generated for each group. A separate plot is constructed for each pair of functions.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis 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.

Titles

Plot Title

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

Symbols Tab

This section specifies the plot symbols.

Plotting Symbols

Group 1-15

Specifies the plotting symbols used for each of the first fifteen groups.

Storage Tab

These options let you specify where to store various row-wise statistics.

Warning: Any data already in a variable is replaced by the new data. Be careful not to specify variables that contain data.

Data Storage Variables

Predicted Group

You can automatically store the predicted group for each row into the variable specified here. The predicted group is generated for each row of data in which all independent variable values are nonmissing.

Linear Discriminant Scores

You can automatically store the linear-discriminant scores for each row into the variables specified here. These scores are generated for each row of data in which all independent variable values are nonmissing. Note that a variable must be specified for each group.

Linear Discriminant Probabilities

You can automatically store the linear-discriminant probabilities for each row into the variables specified here. These probabilities are generated for each row of data in which all independent variable values are nonmissing. Note that a variable must be specified for each group.

Regression Coefficient Scores

You can automatically store the regression coefficient scores for each row into the variables specified here. These scores are generated for each row of data in which all independent variable values are nonmissing. Note that a variable must be specified for each group.

Canonical Scores

You can automatically store the canonical scores for each row into the variables specified here. These scores are generated for each row of data in which all independent variable values are nonmissing. Note that the number of variables specified should be one less that the number of groups.

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 – Discriminant Analysis

This section presents an example of how to run a discriminant analysis. The data used are shown in the table above and found in the FISHER database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Discriminant Analysis window.

1 **Open the Fisher dataset.**

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

2 **Open the Discriminant Analysis window.**

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Discriminant Analysis**. The Discriminant Analysis 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 Discriminant Analysis window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Y: Group Variable box.
- Double-click in the **X's: Independent Variables** text box. This will bring up the variable selection window.
- Select **SepalLength** through **PetalWidth** from the list of variables and then click **Ok**. “SepalLength-PetalWidth” will appear in the X's: Independent Variables box.

4 **Specify the reports.**

- Select the **Reports tab**.
- Enter **Labels** in the **Variable Names** box.
- Enter **Value Labels** in the **Value Labels** box.
- Check all reports and plots. Normally you would only view a few of these reports, but we are selecting them all 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).

Group Means Report

```
Group Means
                     Iris
Variable             Setosa      Versicolor    Virginica    Overall
Sepal Length         50.06       59.36         65.88        58.43333
Sepal Width          34.28       27.7          29.74        30.57333
Petal Length         14.62       42.6          55.52        37.58
Petal Width          2.46        13.26         20.26        11.99333
Count                50          50            50           150
```

This report shows the means of each of the independent variables across each of the groups. The last row shows the count (number of observations) in the group. Note that the column headings come from the use of value labels for the group variable.

Group Standard Deviations Report

```
Group Standard Deviations

                     Iris
Variable             Setosa      Versicolor    Virginica    Overall
Sepal Length         3.524897    5.161712      6.358796     8.280662
Sepal Width          3.790644    3.137983      3.224966     4.358663
Petal Length         1.73664     4.69911       5.518947     17.65298
Petal Width          1.053856    1.977527      2.7465       7.622377
Count                50          50            50           150
```

This report shows the standard deviations of each of the independent variables across each of the groups. The last row shows the count or number of observations in the group.

Discriminant analysis makes the assumption that the covariance matrices are identical for each of the groups. This report lets you glance at the standard deviations to check if they are about equal.

Total Correlation\Covariance Report

```
Total Correlation\Covariance

                    Variable
Variable            Sepal Length    Sepal Width     Petal Length     Petal Width
Sepal Length        68.56935        -4.243401       127.4315         51.62707
Sepal Width         -0.117570       18.99794        -32.96564        -12.16394
Petal Length        0.871754        -0.428440       311.6278         129.5609
Petal Width         0.817941        -0.366126       0.962865         58.10063
```

This report shows the correlation and covariance matrices that are formed when the grouping variable is ignored. Note that the correlations are on the lower left and the covariances are on the upper right. The variances are on the diagonal.

Between-Group Correlation\Covariance Report

Between-Group Correlation\Covariance

Variable	Variable Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	3160.607	-997.6334	8262.42	3563.967
Sepal Width	-0.745075	567.2466	-2861.98	-1146.633
Petal Length	0.994135	-0.812838	21855.14	9338.7
Petal Width	0.999768	-0.759258	0.996232	4020.667

This report displays the correlations and covariances formed using the group means as the individual observations. The correlations are shown in the lower-left half of the matrix. The within-group covariances are shown on the diagonal and in the upper-right half of the matrix. Note that if there are only two groups, all correlations will be equal to one since they are formed from only two rows (the two group means).

Within-Group Correlation\Covariance Report

Within-Group Correlation\Covariance

Variable	Variable Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	26.50082	9.272109	16.75143	3.840136
Sepal Width	0.530236	11.53878	5.524354	3.27102
Petal Length	0.756164	0.377916	18.51878	4.266531
Petal Width	0.364506	0.470535	0.484459	4.188163

This report shows the correlations and covariances that would be obtained from data in which the group means had been subtracted. The correlations are shown in the lower-left half of the matrix. The within-group covariances are shown on the diagonal and in the upper-right half of the matrix.

Variable Influence Report

Variable Influence Section

Variable	Removed Lambda	Removed F-Value	Removed F-Prob	Alone Lambda	Alone F-Value	Alone F-Prob	R-Square Other X's
Sepal Length	0.938463	4.72	0.010329	0.381294	119.26	0.000000	0.858612
Sepal Width	0.766480	21.94	0.000000	0.599217	49.16	0.000000	0.524007
Petal Length	0.669206	35.59	0.000000	0.058628	1180.16	0.000000	0.968012
Petal Width	0.743001	24.90	0.000000	0.071117	960.01	0.000000	0.937850

This report analyzes the influence of each of the independent variables on the discriminant analysis.

Variable

The name of the independent variable.

Removed Lambda

This is the value of a Wilks' lambda computed to test the impact of removing this variable.

Removed F-Value

This is the F-ratio that is used to test the significance of the above Wilks' lambda.

Removed F-Prob

This is the probability (significance level) of the above F-ratio. It is the probability to the right of the F-ratio. The test is significant (the variable is important) if this value is less than the value of alpha that you are using, such as 0.05.

Alone Lambda

This is the value of a Wilks' lambda that would be obtained if this were the only independent variable used.

Alone F-Value

This is an F-ratio that is used to test the significance of the above Wilks' lambda.

Alone F-Prob

This is the probability (significance level) of the above F-ratio. It is the probability to the right of the F-ratio. The test is significant (the variable is important) if this value is less than the value of alpha that you are using, such as 0.05.

R-Squared OtherX's

This is the R-Squared value that would be obtained if this variable were regressed on all other independent variables. When this R-Squared value is larger than 0.99, severe multicollinearity problems exist. You should remove variables (one at a time) with large R-Squared and rerun your analysis.

Linear Discriminant Functions Report

```
Linear Discriminant Functions Section

                    Iris
Variable            Setosa        Versicolor     Virginica
Constant            -85.20985     -71.754        -103.2697
Sepal Length        2.354417      1.569821       1.244585
Sepal Width         2.358787      0.707251       0.3685279
Petal Length        -1.643064     0.5211451      1.276654
Petal Width         -1.739841     0.6434229      2.107911
```

This report presents the linear discriminant function coefficients. These are often called the discriminant coefficients. They are also known as the "plug-in" estimators, since the true variance-covariance matrices are required but their estimates are plugged-in. This technique assumes that the independent variables in each group follow a multivariate-normal distribution with equal variance-covariance matrices across groups. Studies have shown that this technique is fairly robust to departures from either assumption.

The report represents three classification functions, one for each of the three groups. Each function is represented vertically. When a weighted average of the independent variables is formed using these coefficients as the weights (and adding the constant), the discriminant scores result. To determine which group an individual belongs to, select the group with the highest score.

Regression Coefficients Report

```
Regression Coefficients Section

                        Iris
Variable                Setosa          Versicolor      Virginica
Constant                0.1182229       1.577059        -0.6952819
Sepal Length            6.602977E-03    -2.015369E-03   -4.587608E-03
Sepal Width             2.428479E-02    -4.456162E-02   2.027684E-02
Petal Length            -2.246571E-02   2.206692E-02    3.987911E-04
Petal Width             -5.747273E-03   -4.943066E-02   5.517793E-02
```

This report presents the regression coefficients. These coefficients are determined as follows:

1. Create three indicator variables, one for each of the three varieties of iris. Each indicator variable is set to one when the row belongs to that group and zero otherwise.
2. Fit a multiple regression of the independent variables on each of the three indicator variables.
3. The regression coefficients obtained are those shown in this table.

Hence, predicted values generated by these coefficients will be between zero and one. To determine which group an individual belongs to, select the group with the highest score.

Classification Count Table Report

```
Classification Count Table for Iris

              Predicted
Actual        Setosa      Versicolor  Virginica   Total
Setosa        50          0           0           50
Versicolor    0           48          2           50
Virginica     0           1           49          50
Total         50          49          51          150

Reduction in classification error due to X's = 97.0%
```

This report presents a matrix that indicates how accurately the current discriminant functions classify the observations. If perfect classification has been achieved, there will be zeros on the off-diagonals. The rows of the table represent the actual groups, while the columns represent the predicted group.

Percent Reduction

The percent reduction is the classification accuracy achieved by the current discriminant functions over what is expected if the observations were randomly classified. The formula for the Reduction in classification error is [Sum of diagonal minus N/k]/[N - N/k].

Misclassified Rows Report

```
Misclassified Rows Section
                                          Percent Chance of Each Group

Row        Actual       Predicted      Pcnt1    Pcnt2    Pcnt3
5          Virginica    Versicolo      0.0      72.9     27.1
9          Versicolo    Virginica      0.0      25.3     74.7
12         Versicolo    Virginica      0.0      14.3     85.7
```

This report shows the actual group and the predicted group of each observation that was misclassified. It also shows 100 times the estimated probability, P(i), that the row is in each group. For easier viewing, we have multiplied the probabilities by 100 to make this a percent probability (between 0 and 100) rather than a regular probability (between 0 and 1). A value near 100 gives a strong indication that the observation belongs in that group.

P(i)

If the linear discriminant classification technique was used, these are the estimated probabilities that this row belongs to the i^{th} group. See James (1985), page 69, for details of the algorithm used to estimate these probabilities. This algorithm is briefly outlined here.

Let f_i (i = 1, 2, ..., K) be the linear discriminant function value. Let *max(f_k)* be the maximum score of all groups. Let *P(G_i)* be the overall probability of classifying an individual into group *i*. The values of *P(i)* are generated using the following equation:

$$P(i) = \frac{\exp[\, f_i - \max(\, f_k)]P(G_i)}{\sum_{j=1}^{K} \exp[\, f_j - \max(\, f_k)]P(G_j)}$$

If the regression classification technique was used, this is the predicted value of the regression equation. The implicit *Y* value in the regression equation is one or zero, depending on whether this observation is in the i^{th} group or not. Hence, a predicted value near zero indicates that the observation is not in the i^{th} group, while a value near one indicates a strong possibility that this observation is in the i^{th} group. There is nothing to prevent these predicted values from being greater than one or less than zero. They are not estimated probabilities.

You can store these values for further analysis by listing variables in the appropriate *Storage Tab* options.

Predicted Classification Report

```
Predicted Classification Section
                                          Percent Chance of Each Group
Row        Actual       Predicted        Pcnt1     Pcnt2     Pcnt3
1          Setosa       Setosa           100.0       0.0       0.0
2          Virginica    Virginica          0.0       0.0     100.0
3          Versicolo    Versicolo          0.0      99.6       0.4
4          Virginica    Virginica          0.0       0.0     100.0
5          Virginica    Versicolo          0.0      72.9      27.1
6          Setosa       Setosa           100.0       0.0       0.0
7          Virginica    Virginica          0.0       0.0     100.0
8          Versicolo    Versicolo          0.0      96.0       4.0

(report continues for all 150 rows)
```

This report shows the actual group, the predicted group, and the percentage probabilities of each row. The definitions are given above in the *Misclassified Rows Report*.

Canonical Variate Analysis Report

```
Canonical Variate Analysis Section

    Inv(W)B       Ind'l    Total      Canon      Canon             Numer   Denom     Prob      Wilks'
Fn  Eigenvalue     Pcnt     Pcnt       Corr      Corr2  F-Value       DF      DF    Level      Lambda
1   32.191929      99.1     99.1   0.984821   0.969872    199.1      8.0   288.0   0.0000    0.023439
2   0.285391        0.9    100.0   0.471197   0.222027     13.8      3.0   145.0   0.0000    0.777973
The F-value tests whether this function and those below it are significant.
```

This report provides a canonical correlation analysis of the discriminant problem. Recall that canonical correlation analysis is used when you want to study the correlation between two sets of variables. In this case, the two sets of variables are defined in the following way. The independent variables comprise the first set. The group variable defines another set, which is generated by creating an indicator variable for each group except the last one.

Inv(W)B Eigenvalue

The eigenvalues of the matrix $W^{-1}B$. These values indicate how much of the total variation explained is accounted for by the various discriminant functions. Hence, the first discriminant function corresponds to the first eigenvalue, and so on. Note that the number of eigenvalues is the minimum of the number of variables and *K-1*, where *K* is the number of groups.

Ind'l Prcnt

The percent that this eigenvalue is of the total.

Total Prcnt

The cumulative percent of this and all previous eigenvalues.

Canon Corr

The canonical correlation coefficient.

Canon Corr2

The square of the canonical correlation. This is similar to R-Squared in multiple regression.

F-Value

The value of the approximate F-ratio for testing the significance of the Wilks' lambda corresponding to this row and those below it. Hence, in this example, the first F-value tests the significance of both the first and second canonical correlations, while the second F-value tests the significance of the second correlation only.

Num DF

The numerator degrees of freedom for this F-test.

Denom DF

The denominator degrees of freedom for this F-test.

Prob Level

The significance level of the F-test. This is the area under the F-distribution to the right of the F-value. Usually, a value less than 0.05 is considered significant.

Wilks' Lambda

The value of Wilks' lambda for this row. This Wilks' lambda is used to test the significance of the discriminant function corresponding to this row and those below it. Recall that Wilks' lambda is a multivariate generalization of R^2. The above F-value is an approximate test of this Wilks' lambda.

Canonical Coefficients Report

```
Canonical Coefficients Section

                   Canonical Variate
Variable           Variate1        Variate2
Constant           -2.105106       6.661473
Sepal Length       -0.082938       -0.002410
Sepal Width        -0.153447       -0.216452
Petal Length       0.220121        0.093192
Petal Width        0.281046        -0.283919
```

This report gives the coefficients used to create the canonical scores. The canonical scores are weighted averages of the observations, and these coefficients are the weights (with the constant term added).

Canonical Variates at Group Means Report

```
Canonical Variates at Group Means Section

                   Canonical Variate
Iris               Variate1        Variate2
Setosa             -7.6076         -0.215133
Versicolor         1.82505         0.7278996
Virginica          5.78255         -0.5127666
```

This report gives the results of applying the canonical coefficients to the means of each of the groups.

Std. Canonical Coefficients Report

```
Std. Canonical Coefficients Section

                   Canonical Variate
Variable           Variate1        Variate2
Sepal Length       -0.426955       -0.012408
Sepal Width        -0.521242       -0.735261
Petal Length       0.947257        0.401038
Petal Width        0.575161        -0.581040
```

This report gives the standardized canonical coefficients.

Variable-Variate Correlations Report

```
Variable-Variate Correlations Section

                   Canonical Variate
Variable           Variate1        Variate2
Sepal Length       0.222596        -0.310812
Sepal Width        -0.119012       -0.863681
Petal Length       0.706065        -0.167701
Petal Width        0.633178        -0.737242
```

This report gives the loadings (correlations) of the variables on the canonical variates. That is, each entry is the correlation between the canonical variate and the independent variable. This report can help you interpret a particular canonical variate.

Linear Discriminant Scores Report

```
Linear Discriminant Scores Section

Row      Iris          Score1          Score2          Score3
1        Setosa        83.86837        38.65921        -6.790054
2        Virginica     1.230765        91.857          104.5692
3        Versicolo     32.19471        83.71141        78.29187
4        Virginica     11.89069        99.97506        113.6244
5        Virginica     19.27056        83.17749        82.18597
6        Setosa        75.06965        33.7306         -9.291955
7        Virginica     26.55469        99.86555        107.6224

(report continues for all 150 rows)
```

This report gives the individual values of the linear discriminant scores. Note that this information may be stored on the database using the Data Storage options.

Regression Scores Report

```
Regression Scores Section

Row      Iris          Score1          Score2          Score3
1        Setosa        0.923755        0.215832        -0.139588
2        Virginica     -0.163732       0.348623        0.815109
3        Versicolo     0.107759        0.471953        0.420288
4        Virginica     -0.082564       0.110031        0.972533
5        Virginica     -0.017776       0.586318        0.431458
6        Setosa        0.915881        0.129902        -0.045782
7        Virginica     0.048718        0.045096        0.906186

(report continues for all 150 rows)
```

This report gives the individual values of the predicted scores based on the regression coefficients. Even though these values are predicting indicator variables, it is possible for a value to be less than zero or greater than one. Note that this information may be stored on the database using the *Data Storage* options.

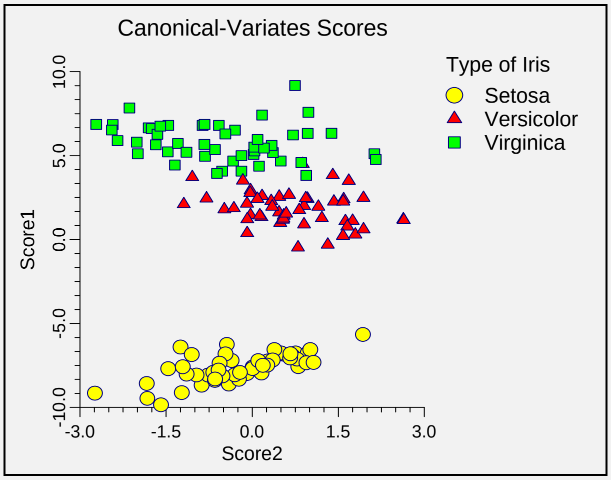
Canonical Scores Report

```
Canonical Scores Section

Row       Iris              Score1            Score2
1         Setosa            -7.671967         0.134894
2         Virginica         6.800150          -0.580895
3         Versicolo         2.548678          0.472205
4         Virginica         6.653087          -1.805320
5         Virginica         3.815160          0.942986
6         Setosa            -7.212618         -0.355836
7         Virginica         5.105559          -1.992182
(report continues for all 150 rows)
```

This report gives the scores of the canonical variates for each row. Note that this information may be stored on the database using the Data Storage options.

Scores Plot(s)

You may select plots of the linear discriminant scores, regression scores, or canonical scores to aid in your interpretation. These plots are usually used to give a visual impression of how well the discriminant functions are classifying the data. (Several charts are displayed. Only one of these is displayed here.)



This chart plots the values of the first and second canonical scores. By looking at this plot you can see what the classification rule would be. Also, it is obvious from this plot that only the first canonical function is necessary in discriminating among the varieties of iris since the groups can easily be separated along the vertical axis.

Example 2 – Automatic Variable Selection (Brief Report)

The tutorial we have just concluded was based on all four of the independent variables. A common task in discriminant analysis is variable selection. Often you have a large pool of possible independent variables from which you want to select a smaller set (up to about eight variables) which will do almost as well at discriminating as the complete set. ***NCSS*** provides an automatic procedure for doing this, which will be described next.

The automatic variable selection is run by changing the Variable Selection option to Stepwise. The program will conduct a stepwise variable selection. It will first find the best discriminator and then the second best. After it has found two, it checks whether the discrimination would be almost as good if one were removed. This stepping process of adding the best remaining variable and then checking if one of the active variables could be removed continues until no new variable can be found whose F-value has a probability smaller than the Probability Enter value.

An alternative procedure is to use the Multivariate Variable Selection procedure described elsewhere in this manual. If you have more than two groups, you must create a set of dummy (indicator) variables, one for each group. You ignore the last dummy variable, so if there are *K* groups, you analyze *K-1* dummy variables. The Multivariate Variable Selection program will always find a subset of your independent variables that is at least as good (and usually better) as the stepwise procedure described in this section. Once a subset of independent variables has been found, they can then be analyzed using the Discriminant Analysis program described here.

Once the variable selection has been made, the program provides the reports that were described in the previous tutorial. Note that two report formats may be called for during the variable selection phase: brief and verbose. We will now provide an example of each type of report.

You may follow along here by making the appropriate entries or load the completed template **Example2** from the Template tab of the Discriminant Analysis window.

1 Open the Fisher dataset.

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

2 Open the Discriminant Analysis window.

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Discriminant Analysis**. The Discriminant Analysis procedure will be displayed.

3 Specify the variables.

- On the Discriminant Analysis window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Y: Group Variable box.
- Double-click in the **X’s: Independent Variables** text box. This will bring up the variable selection window.
- Select **Sepal Length** through **PetalWidth** from the list of variables and then click **Ok**. “SepalLength-PetalWidth” will appear in the X’s: Independent Variables.
- Enter **Stepwise** in the **Variable Selection** box.

4 **Specify the reports.**

- Select the **Reports tab**.
- Enter **Labels** in the **Variable Names** box.
- Enter **Value Labels** in the **Value Labels** box.
- Enter **Brief** in the **Output** box.
- Uncheck all reports and plots. We will only view the Variable Selection Report.

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).

Variable-Selection Summary Report

```
Variable-Selection Summary Section

               Action   Independent     Pct Chg In                   Prob       Wilks'
Iteration   This Step      Variable         Lambda    F-Value       Level       Lambda
0                None                                                         1.000000
1             Entered  Petal Length          94.14    1180.16    0.000000     0.058628
2             Entered   Sepal Width          37.09      43.04    0.000000     0.036884
3             Entered   Petal Width          32.23      34.57    0.000000     0.024976
4             Entered  Sepal Length           6.15       4.72    0.010329     0.023439
```

This report shows what action was taken at each step.

Iteration

This gives the number of this step.

Action This Step

This tells what action (if any) was taken during this step. “Entered” means that the variable was entered into the set of active variables. “Removed” means that the variable was removed from the set of active variables.

Pct Chg In Lambda

This is the percentage decrease in lambda that resulted from this step. Note that Wilks’ lambda is analogous to 1 - R-Squared in multiple regression. Hence, we want to *decrease* Wilks’ lambda to improve our model. For example, going from iteration 2 to iteration 3 results in lambda decreasing from 0 .036884 to 0.024976. This is a 32.29% decrease in lambda.

F-Value

This is the F-ratio for testing the significance of this variable. If the variable was “Entered,” this tests the hypothesis that the variable should be added. If the variable was “Removed,” this tests whether the variable should be removed.

Prob Level

The significance level of the above F-Value.

Wilks’ Lambda

The multivariate extension of R-Squared. Wilks’ lambda reduces to 1-(R-Squared) in the two-group case. It is interpreted just backwards from R-Squared. It varies from one to zero. Values near one imply low predictability, while values close to zero imply high predictability. Note that this Wilks’ lambda value corresponds to the currently active variables.

Example 3 – Automatic Variable Selection (Verbose Report)

We will now rerun this example with the “verbose” option. We assume that the FISHER database is available and you are in the Discriminant Analysis procedure.

You may follow along here by making the appropriate entries or load the completed template **Example3** from the Template tab of the Discriminant Analysis window.

1 **Open the Fisher dataset.**

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

2 **Open the Discriminant Analysis window.**

- On the menus, select **Analysis**, then **Multivariate Analysis**, then **Discriminant Analysis**. The Discriminant Analysis procedure will be displayed.

3 **Specify the variables.**

- On the Discriminant Analysis window, select the **Variables tab**.
- Double-click in the **Y: Group Variable** box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. “Iris” will appear in the Y: Group Variable box.
- Double-click in the **X’s: Independent Variables** text box. This will bring up the variable selection window.
- Select **Sepal Length** through **PetalWidth** from the list of variables and then click **Ok**. “SepalLength-PetalWidth” will appear in the X’s: Independent Variables.
- Enter **Stepwise** in the Variable Selection box.

4 **Specify the reports.**

- Select the **Reports tab**.
- Enter **Labels** in the **Variable Names** box.
- Enter **Value Labels** in the **Value Labels** box.
- Enter **Verbose** in the **Output** box.
- Uncheck all reports and plots. We will only view the Variable Selection Report.

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).

Variable-Selection Detail Report

```
Variable-Selection Detail Section

Step        0
            Independent       Pct Chg In                      Prob      R-Squared
Status      Variable              Lambda      F-Value        Level      Other X’s
Out         Sepal Length         61.8706       119.26     0.000000       0.000000
Out         Sepal Width          40.0783        49.16     0.000000       0.000000
Out         Petal Length         94.1372      1180.16     0.000000       0.000000
Out         Petal Width          92.8883       960.01     0.000000       0.000000
Overall Wilks' Lambda           1.000000
Action this step:                   None

Step        1
            Independent       Pct Chg In                      Prob      R-Squared
Status      Variable              Lambda      F-Value        Level      Other X’s
In          Petal Length         94.1372      1180.16     0.000000       0.000000
Out         Sepal Length         31.9811        34.32     0.000000       0.759955
Out         Sepal Width          37.0882        43.04     0.000000       0.183561
Out         Petal Width          25.3317        24.77     0.000000       0.927110
Overall Wilks' Lambda           0.058628
Action this step:   Petal Length Entered

Step        2
            Independent       Pct Chg In                      Prob      R-Squared
Status      Variable              Lambda      F-Value        Level      Other X’s
In          Sepal Width          37.0882        43.04     0.000000       0.183561
In          Petal Length         93.8446      1112.95     0.000000       0.183561
Out         Sepal Length         14.4729        12.27     0.000012       0.840178
Out         Petal Width          32.2865        34.57     0.000000       0.929747
Overall Wilks' Lambda           0.036884
Action this step:    Sepal Width Entered

Step        3
            Independent       Pct Chg In                      Prob      R-Squared
Status      Variable              Lambda      F-Value        Level      Other X’s
In          Sepal Width          42.9479        54.58     0.000000       0.213103
In          Petal Length         34.8165        38.72     0.000000       0.933764
In          Petal Width          32.2865        34.57     0.000000       0.929747
Out         Sepal Length          6.1537         4.72     0.010329       0.858612
Overall Wilks' Lambda           0.024976
Action this step:    Petal Width Entered

Step        4
            Independent       Pct Chg In                      Prob      R-Squared
Status      Variable              Lambda      F-Value        Level      Other X’s
In          Sepal Length          6.1537         4.72     0.010329       0.858612
In          Sepal Width          23.3520        21.94     0.000000       0.524007
In          Petal Length         33.0794        35.59     0.000000       0.968012
In          Petal Width          25.6999        24.90     0.000000       0.937850
Overall Wilks' Lambda           0.023439
Action this step:   Sepal Length Entered
```

This report shows the details of each step.

Step

This gives the number of this step (iteration).

Status

This tells whether the variable is “in” or “out” of the set of active variables.

Pct Chg In Lambda

This is the percentage decrease in lambda that would result if the status of this variable were reversed.

F-Value

This is the F-ratio for testing the significance of changing the status of this variable.

Prob Level

The significance level of the above F-Value.

R-Squared Other X's

This is the R-Squared that would result if this variable were regressed on the other independent variables that are active (status = "In"). This provides a check for multicollinearity in the active independent variables.

Overall Wilks' Lambda

This is the value of Wilks' lambda for all active independent variables. A value near zero indicates an accurate model; a value near one indicates a poor model.

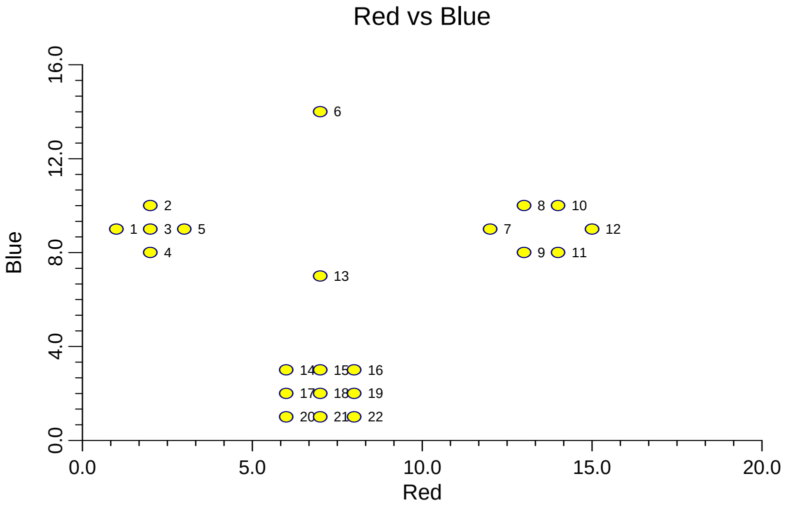
Chapter 445

Hierarchical Clustering (Dendrograms)

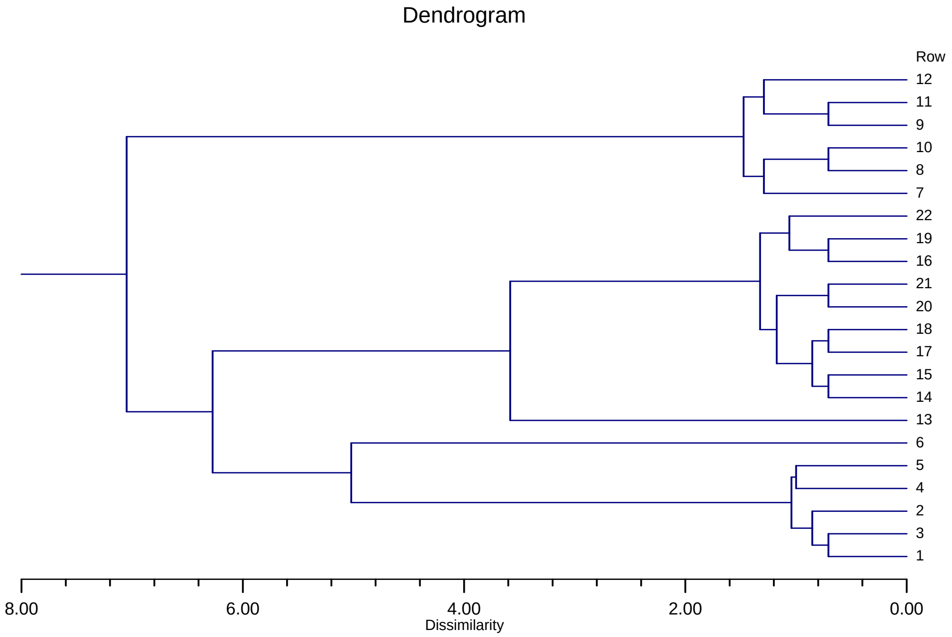
Introduction

The *agglomerative hierarchical clustering* algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a *dendrogram*. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated. The eight methods that are available represent eight methods of defining the similarity between clusters.

Suppose we wish to cluster the bivariate data shown in the following scatter plot. In this case, the clustering may be done visually. The data have three clusters and two singletons, 6 and 13.



Following is a dendrogram of the results of running these data through the Group Average clustering algorithm.



The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The dendrogram is fairly simple to interpret. Remember that our main interest is in similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters.

Looking at this dendrogram, you can see the three clusters as three branches that occur at about the same horizontal distance. The two outliers, 6 and 13, are fused in rather arbitrarily at much higher distances. This is the interpretation.

In this example we can compare our interpretation with an actual plot of the data. Unfortunately, this usually will not be possible because our data will consist of more than two variables.

Dissimilarities

The first task is to form the distances (dissimilarities) between individual objects. This is described in the Medoid Clustering chapter and will not be repeated here.

Hierarchical Algorithms

The algorithm used by all eight of the clustering methods is outlined as follows. Let the distance between clusters *i* and *j* be represented as d_{ij} and let cluster *i* contain n_i objects. Let ***D*** represent the set of all remaining d_{ij}. Suppose there are *N* objects to cluster.

1. Find the smallest element d_{ij} remaining in ***D***.

2. Merge clusters *i* and *j* into a single new cluster, *k*.
3. Calculate a new set of distances d_{km} using the following distance formula.

   $$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma \left| d_{im} - d_{jm} \right|$$

   Here *m* represents any cluster other than *k*. These new distances replace d_{im} and d_{jm} in ***D***.

   Also let $n_k = n_i + n_j$.

   Note that the eight algorithms available represent eight choices for $\alpha_i, \alpha_j, \beta,$ and γ .
4. Repeat steps 1 - 3 until ***D*** contains a single group made up off all objects. This will require *N-1* iterations.

We will now give brief comments about each of the eight techniques.

Single Linkage

Also known as *nearest neighbor* clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closest members. It often yields clusters in which individuals are added sequentially to a single group.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = -0.5.$

Complete Linkage

Also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.5.$

Simple Average

Also called the weighted pair-group method, this algorithm defines the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = 0.$

Centroid

Also referred to as the unweighted pair-group centroid method, this method defines the distance between two groups as the distance between their centroids (center of gravity or vector average). The method should only be used with Euclidean distances.

The coefficients of the distance equation are $\alpha_i = \frac{n_i}{n_k}, \alpha_j = \frac{n_j}{n_k}, \beta = -\alpha_i \alpha_j, \gamma = 0.$

Backward links may occur with this method. These are recognizable when the dendrogram no longer exhibits its simple tree-like structure in which each fusion results in a new cluster that is at a higher distance level (moves from right to left). With backward links, fusions can take place that result in clusters at a lower distance level (move from left to right). The dendrogram is difficult to interpret in this case.

Median

Also called the weighted pair-group centroid method, this defines the distance between two groups as the weighted distance between their centroids, the weight being proportional to the number of individuals in each group. Backward links (see discussion under Centroid) may occur with this method. The method should only be used with Euclidean distances.

The coefficients of the distance equation are $\alpha_i = \alpha_j = 0.5, \beta = -0.25, \gamma = 0.$

Group Average

Also called the unweighted pair-group method, this is perhaps the most widely used of all the hierarchical cluster techniques. The distance between two groups is defined as the average distance between each of their members.

The coefficients of the distance equation are $\alpha_i = \frac{n_i}{n_k}, \alpha_j = \frac{n_j}{n_k}, \beta = 0, \gamma = 0.$

Ward's Minimum Variance

With this method, groups are formed so that the pooled within-group sum of squares is minimized. That is, at each step, the two clusters are fused which result in the least increase in the pooled within-group sum of squares.

The coefficients of the distance equation are $\alpha_i = \frac{n_i + n_m}{n_k + n_m}, \alpha_j = \frac{n_j + n_m}{n_k + n_m}, \beta = \frac{-n_m}{n_k + n_m}, \gamma = 0.$

Flexible Strategy

Lance and Williams (1967) suggested that a continuum could be made between single and complete linkage. The program lets you try various settings of these parameters which do not conform to the constraints suggested by Lance and Williams.

The coefficients of the distance equation should conform to the following constraints $\alpha_i = 1 - \beta - \alpha_j, \alpha_j = 1 - \beta - \alpha_i, -1 \le \beta \le 1, \gamma = 0.$

One interesting exercise is to vary these values, trying to find the set that maximizes the cophenetic correlation coefficient.

Goodness-of-Fit

Given the large number of techniques, it is often difficult to decide which is best. One criterion that has become popular is to use the result that has largest *cophenetic correlation coefficient*. This is the correlation between the original distances and those that result from the cluster configuration. Values above 0.75 are felt to be good. The Group Average method appears to produce high values of this statistic. This may be one reason that it is so popular.

A second measure of goodness of fit called *delta* is described in Mather (1976). These statistics measure degree of distortion rather than degree of resemblance (as with the cophenetic correlation). The two delta coefficients are given by

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d^*_{jk}|^{1/A}}{\sum_{j<k} (d^*_{jk})^{1/A}} \right]^A$$

where A is either 0.5 or 1 and d^*_{ij} is the distance obtained from the cluster configuration. Values close to zero are desirable.

Mather (1976) suggests that the Group Average method is the safest to use as an exploratory method, although he goes on to suggest that several methods should be tried and the one with the largest cophenetic correlation be selected for further investigation.

Number of Clusters

These techniques do not let you explicitly set the number of clusters. Instead, you pick a distance value that will yield an appropriate number of clusters. This will be discussed further when we discuss the Dendrogram and the Linkage report.

Limitations and Criticisms

We have attempted problems with up to 1,000 objects. Running times will vary with computer speed, with larger problems running several hours. Problems with 100 objects or less should run in a few seconds.

Hierarchical clustering methods are popular because they are relatively simple to understand and implement. However, this simplicity yields one of their strongest criticisms. Once two objects are joined, they can never be separated. As Kaufman (1990) complains, “once the damage is done, it can never be repaired.”

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable.

The data given in the following table contain information on twelve superstars in basketball. The stats are on a per game basis for games played through the 1989 season.

BBALL dataset (subset)

Player	Height	FgPct	Points	Rebounds
Jabbar K.A.	86.0	55.9	24.6	11.2
Barry R	79.0	44.9	23.2	6.7
Baylor E	77.0	43.1	27.4	13.5
Bird L	81.0	50.3	25	10.2
Chamberlain W	85.0	54.0	30.1	22.9
Cousy B	72.5	37.5	18.4	5.2
Erving J	78.5	50.6	24.2	8.5

Missing Values

When an observation has missing values, appropriate adjustments are made so that the average dissimilarity across all variables with non-missing data is computed. Hence, rows with missing values are not omitted unless all variables have missing values. Note that the distances require that at least one variable have non-missing values for each pair of rows.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Interval Variables

Designates interval-type variables (if any) or the columns of the matrix if distance or correlation matrix input was selected. Interval variables are continuous measurements that may be either positive or negative and follow a linear scale. Examples include height, weight, age, price, temperature, and time.

In general, an interval should keep the same importance throughout the scale. For example, the length of time between 1905 and 1925 is the same as the length of time between 1995 and 2015.

Note that a nonlinear transformation of an interval variable is probably not an interval variable. For example, the logarithm of height is not an interval variable since the value of an interval along the scale changes depending upon where you are on the scale.

Ordinal Variables

Specifies the ordinal-type variables (if any). Ordinal variables are measurements that may be ordered according to magnitude. For example, a survey question may require you to pick one of five possible choices: strongly disagree (5), disagree (4), neutral (3), agree (2), or strongly agree (1). Interval variables are ordinal, but ordinal variables are not necessarily interval.

The original values of ordinal variables are replaced by their ranks. These ranks are then analyzed as if they were interval variables.

Symmetric-Binary Variables

Specifies the symmetric binary-type variables (if any). Symmetric binary variables have two possible outcomes, each of which carry the same information and weight. Examples include gender, marital status, or membership in a particular group. Usually, they are coded as 1 for yes or 0 for no, although this is not necessary.

These variables are analyzed using the number of matches between two individuals.

Ratio Variables

Specifies the ratio variables (if any). Ratio-type variables are positive measurements in which the distinction between two numbers is constant if their ratio is constant. For example, the distinction between 3 and 30 would have the same meaning as the distinction between 30 and 300. Examples are chemical concentration or radiation intensity. The logarithms of ratio variables are analyzed as if they were interval variables.

Nominal Variables

Specifies the nominal-type variables (if any). Nominal variables are those in which the number represents the state of the variable. Examples include gender, race, hair color, country of birth, or zipcode. If a nominal variable has only two categories, it is often called a binary variable.

Nominal variables are analyzed using the number of matches between two individuals.

Asymmetric-Binary Variables

Specifies the asymmetric binary-type variables (if any). Asymmetric binary-scaled variables are concerned with the presence or absences of a relatively rare event, the absence of which is unimportant.

These variables are analyzed using the number of matches in which both individuals have the trait of interest. Those cases in which both individuals do not have the trait are not of interest and are ignored.

Linkage Options

Linkage Type

This option specifies which of the eight possible hierarchical techniques is used. These methods were described earlier. The choices are

- **Single Linkage (Nearest Neighbor)**
- **Complete Linkage (Furthest Neighbor)**
- **Simple Average (Weighted Pair-Group)**
- **Group Average (Unweighted Pair-Group)**
- **Median (Weighted Pair-Group Centroid)**
  Requires the Distance Method to be Euclidean.
- **Centroid (Unweighted Pair-Group Centroid)**
  Requires the Distance Method to be Euclidean.
- **Ward’s Minimum Variance**
  Requires the Distance Method to be Euclidean.
- **Flexible Strategy**
  Requires the Distance Method to be Euclidean.

When in doubt, we suggest you try the Group Average method. It seems to be the most popular and most recommended in the cluster literature.

Linkage Options – Flexible Strategy Parameters

Alpha

Specifies the values of α_i and α_j when the Flexible Strategy method is selected. You may enter a number or the letters “NI/NK.” The “NI/NK” will cause this constant to be calculated and used as it is in the Centroid and Group Average methods.

Beta

Specifies the values of β when the Flexible Strategy method is selected. You may enter a number between -1 and 1 or the letters “NIJ/NK.” The “NIJ/NK” will cause this constant to be calculated and used as it is in the Centroid method.

Gamma

Specifies the values of γ when the Flexible Strategy method is selected. You may enter any number.

Clustering Options

Distance Method

This option specifies with Euclidean or Manhattan distance is used. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Scaling Method

Specify the type of scaling to be used from Interval, Ordinal, and Ratio variables. Possible choices are Standard Deviation, Average Absolute Deviation, Range, and None. These were discussed in the introduction to this chapter.

Cluster Cutoff

This is the cutoff point at which clusters are formed and stored if a Cluster Id variable is specified. Subgroups that join at a distance below this value are put in the same cluster. Subgroups that join at a distance greater than this value are placed in different clusters.

Note that usually you will have to run an analysis first to determine an appropriate value for this distance. This can be done by viewing the dendrogram and the Linkage Report.

Format Options

Label Variable

This is an optional variable containing identification for each row (object). These labels are used to enhance the interpretability of the reports. When used, they replace the row numbers on the right of the dendrogram.

Input Format

Specify the type of data format that you have. Your choices are

- **Raw Data**

  The variables are in the standard format in which each row represents an object and each column represents a variable.

- **Distances**

  The variables containing a distance matrix are specified in the Interval Variables option. Note that this matrix contains the distances between each pair of objects. Each object is represented by a row and the corresponding column. Also, the matrix must be complete. You cannot use only the lower triangular portion, for example.

- **Correlations 1**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

  $$d_{ij} = \frac{1 - r_{ij}}{2}$$

- **Correlations 2**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

  $$d_{ij} = 1 - \left| r_{ij} \right|$$

- **Correlations 3**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

  $$d_{ij} = 1 - r_{ij}^2$$

  Note that all three types of correlation matrices must be completely specified. You cannot specify only the lower or upper triangular portions. Also, the rows correspond to variables. That is, the values along the first row represent the correlations of the first variable with each of the other variables. Hence, you cannot rearrange the order of the matrix.

Reports Tab

The following options control the formatting of the reports.

Select Reports

Cluster Report - Dendrogram

Specify whether to display the indicated reports and 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.

Max Distance Items

This option specifies the maximum size of a distance matrix that will be displayed in the Distance Section report. Distance matrices with more items than this will not be displayed.

This option is here because for large datasets, the distance matrix may be very large.

Dendrogram Tab

These options control the attributes of the dendrogram.

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.

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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Dendrogram 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.

Label Space

This is a positioning parameter that specifies how much room is set aside on the right of the dendrogram for the row labels (or numbers).

Rows per Page

The maximum number of rows displayed on a single dendrogram. If you have more rows than this on your database, the dendrogram will be divided up into several sections. Each section will appear as a single dendrogram. This option allows you to divide up a dendrogram so that row labeling will not overlap.

Line Color

This is the color of the dendrogram's lines.

Line Width

This is the width of the dendrogram's lines.

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.

Storage Tab

These options let you specify where to store the cluster number of each row on the current database.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the cutoff value specified in the Cluster Cutoff option. Points that are unnumbered are those that cannot be placed in any cluster.

Warning: Any data already in this variable are replaced by the cluster number. Be careful not to specify variables that contain important data.

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 – Hierarchical Clustering

This section presents an example of how to run a cluster analysis of the basketball superstars data. The data are found in the BBALL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Hierarchical Clustering / Dendrograms window.

1 Open the BBall dataset.

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

2 Open the Hierarchical Clustering / Dendrograms window.

- On the menus, select **Analysis**, then **Clustering**, then **Hierarchical** or **Dendrograms**. The Hierarchical Clustering / Dendrograms 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 Hierarchical Clustering / Dendrograms window, select the **Variables** tab.
- Double-click in the **Interval Variables** box. This will bring up the variable selection window.
- Select **Height, FgPct, Points, Rebounds** from the list of variables and then click **Ok**. “Height, FgPct, Points, Rebounds” will appear in the Interval Variables box.
- Double-click in the **Label Variable** box. This will bring up the variable selection window.
- Select **Player** from the list of variables and then click **Ok**. “Player” will appear in the Label Variable box.

4 Specify the report.

- On the Hierarchical Clustering / Dendrograms window, select the **Reports tab**.
- Check the **Distance Report**. All reports should be selected.

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).

Cluster Detail Section

```
Cluster Detail Section

Row       Cluster     Player
1         1           Jabbar K.A.
8         1           Johnson M
2         2           Barry R
3         2           Baylor E
4         2           Bird L
7         2           Erving J
9         2           Jordan M
10        2           Robertson O
12        2           West J
5                     Chamberlain W
6                     Cousy B
11                    Russell B
```

This report displays the cluster number associated with each row. The report is sorted by row number within cluster number. The cluster number of rows that cannot be classified are left blank. The cluster configuration depends on the Cluster Cutoff value that was used.

Linkage Section

```
Linkage Section

          Number                 Distance   Distance                          Rows
Link      Clusters                  Value   Bar                               Linked
11        1                      1.822851   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIII    1,8,2,4,7,10,12,3,9,5,11,6
10        2                      1.780810   |IIIIIIIIIIIIIIIIIIIIIIIIIIIII     1,8,2,4,7,10,12,3,9,5,11
9         3                      1.642553   |IIIIIIIIIIIIIIIIIIIIIIIIIII       1,8,2,4,7,10,12,3,9,5
8         4                      1.199225   |IIIIIIIIIIIIIIIIIIII              1,8,2,4,7,10,12,3,9
7         5                      0.941566   |IIIIIIIIIIIIIII                   2,4,7,10,12,3,9
6         6                      0.919016   |IIIIIIIIIIIIIII                   1,8
5         7                      0.826883   |IIIIIIIIIIIIII                    2,4,7,10,12,3
4         8                      0.693822   |IIIIIIIIIII                       2,4,7,10,12
3         9                      0.579517   |IIIIIIIIII                        2,4,7,10
2         10                     0.470534   |IIIIIIII                          4,7,10
1         11                     0.325592   |IIIII                             7,10

Cophenetic Correlation           0.830472
Delta(0.5)                       0.171620
Delta(1.0)                       0.223057
```

This report displays the subgroup that is formed at each fusion that took place during the cluster analysis. The links are displayed in reverse order so that you can quickly determine an appropriate number of clusters to use. It displays the distance level at which the fusion took place. It will let you precisely determine the best value of the Cluster Cutoff value.

For example, looking down the Distance Value column of the report, you can see that the cutoff value that we used (the default value is 1.0) occurs between Links 7 and 8. Hence, the cutoff value of 1.0 results in five clusters. Looking at the Cluster Detail Section (above), you will see that we obtained two real clusters and three outliers. These outliers are called as clusters even though they consist of only one individual.

The cophenetic correlation and the two delta goodness of fit statistics are reported at the bottom of this report. As discussed earlier, these values let you compare the fit of various cluster configurations.

Link

This is the sequence number of the fusion.

Number Clusters

This is the number of clusters that would result if the Cluster Cutoff value were set to the corresponding Distance Value or higher. Note that this number includes outliers.

Distance Value

This is distance value between the two joining clusters that is used by the algorithm. Normally, this value is monotonically increasing. When backward linking occurs, this value will no longer exhibit a strictly increasing behavior.

As discussed above, these values are used to determine an appropriate number of clusters.

Distance Bar

This is a bar graph of the Distance Values. Choose the number of clusters by finding a jump in the decreasing pattern shown in this bar chart.

Rows Linked

These are the rows that were joined at this step. Remember that the links are presented in reverse order, so, in our example, rows 7 and 10 were joined first, row 4 was added, and so on.

Cophenetic Correlation

This is the Pearson correlation between the actual distances and the predicted distances based on this particular hierarchical configuration. A value of 0.75 or above needs to be achieved in order for the clustering to be considered useful.

Delta (0.5, 1)

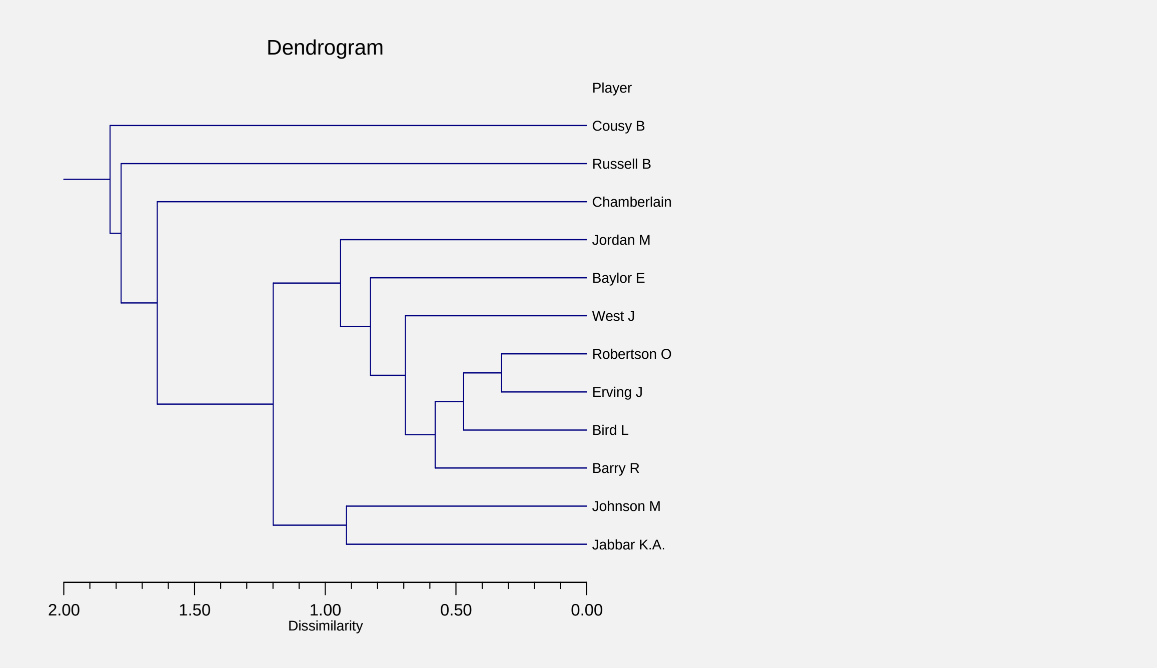
These are the values of the goodness of fit deltas. When comparing to clustering configurations, the configuration with the smallest delta value fits the data better.

Distance Section

```
Distance Section

First   Second      Actual      Dendrogram       Actual      Percent
Row     Row       Distance        Distance   Difference   Difference
1       2         1.427013        1.199225     0.227788        15.96
1       3         1.703276        1.199225     0.504050        29.59
1       4         0.833498        1.199225    -0.365727       -43.88
1       5         1.126296        1.642553    -0.516257       -45.84
1       6         2.575167        1.822851     0.752316        29.21
1       7         1.100763        1.199225    -0.098462        -8.94
1       8         0.919016        0.919016     0.000000         0.00
.       .                .               .            .            .
.       .                .               .            .            .
.       .                .               .            .            .
```

This report displays the actual and predicted distance for each pair of rows. It also includes their difference and percent difference. Since the report grows very long for even a modest number of rows, it is usually omitted.

Dendrogram Section



This report displays the dendrogram which visually displays a particular cluster configuration. Rows that are close together (have small dissimilarity) will be linked near the right side of the plot. For example, we notice the Oscar Robertson and Julius Erving are very similar.

Rows that link up near the left side are very different. For example, Bob Cousy appears to be quite different from any of the other players.

The number of clusters the will be formed at a particular Cluster Cutoff value may be quickly determined from this plot by drawing a vertical line at that value and counting the number of lines that the vertical line intersects. For example, you can see that if we draw a vertical line at the value 1.0, five clusters will result. One cluster will contain two objects, one will contain seven objects, and three clusters each will contain only one object.

We strongly recommend that you compare the dendrograms from several different methods and on several different datasets with known cluster patterns so that you can get the feel of the technique.

445-16 Hierarchical Clustering (Dendrograms)

Chapter 446

K-Means Clustering

Introduction

The k-means algorithm was developed by J.A. Hartigan and M.A. Wong of Yale University as a partitioning technique. It is most useful for forming a small number of clusters from a large number of observations. It requires variables that are continuous with no outliers. Discrete data can be included but may cause problems.

The objective of this technique is to divide *N* observations with *P* dimensions (variables) into *K* clusters so that the within-cluster sum of squares is minimized. Since the number of possible arrangements is enormous, it is not practical to expect the best solution. Rather, this algorithm finds a “local” optimum. This is a solution in which no movement of an observation from one cluster to another will reduce the within-cluster sum of squares. The algorithm may be repeated several times with different starting configurations. The optimum of these cluster solutions is then selected.

Technical Details

The k-means clustering algorithm is popular because it can be applied to relatively large sets of data. The user specifies the number of clusters to be found. The algorithm then separates the data into spherical clusters by finding a set of cluster centers, assigning each observation to a cluster, determining new cluster centers, and repeating this process.

Assume that you have *N* rows (observations), which are separated into *K* groups. The k^{th} cluster contains n_k observations. Each row consists of *P* variables. A missing value in the i^{th} variable of the j^{th} row of the k^{th} group is designated by δ_{ijk}.

The data are standardized by subtracting the variable mean and dividing by the standard deviation. The standardized data elements are referred to as z_{ij}.

Cluster Initialization

The method of initializing the clusters influences the final cluster solution. For each trial, **NCSS** randomly assigns each point to a cluster. This configuration is optimized using the k-means algorithm. Trying several random starting configurations will greatly increase the probability of finding the global optimum solution for a particular number of clusters.

Goodness-of-Fit Criterion

The goodness-of-fit criterion used to compare various cluster configurations is based on the within-cluster sum of squares, WSS_K, where

$$WSS_K = \left(\frac{NP}{NP-m}\right)\sum_{k=1}^{K}\sum_{i=1}^{P}\sum_{j=1}^{n_k}(1-\delta_{ijk})(z_{ij}-c_{ik})^2$$

where c_{ik} is the average (center) value of the i^{th} variable in the k^{th} cluster.

The percent of variation is defined as

$$PV_K = 100\frac{WSS_K}{WSS_1}$$

Data Structure

The data given in the following table contain information on twelve of the most famous superstars in basketball. The stats are on a per game basis for games played through the 1989 season.

BBALL dataset (subset)

Player	Height	FgPct	Points	Rebounds
Jabbar K.A.	86.0	55.9	24.6	11.2
Barry R	79.0	44.9	23.2	6.7
Baylor E	77.0	43.1	27.4	13.5
Bird L	81.0	50.3	25	10.2
Chamberlain W	85.0	54.0	30.1	22.9
Cousy B	72.5	37.5	18.4	5.2
Erving J	78.5	50.6	24.2	8.5
Johnson M	81.0	53.0	19.5	7.4

Missing Values

You control the fate of observations with missing values by setting a percent-missing parameter. Observations with more than the specified percentage of missing values are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Cluster Variables

Designates the variables to be clustered. Note that the k-means algorithm assumes that all of your variables are continuous with no outliers. If your data do not meet these requirements, use caution when applying this technique.

Label Variable

An optional variable containing row labels that you may want to use to document your output. You can use dates (like Jan-23-95) as labels. Here is how. First, enter your dates using the standard date format (like 06/20/93). In the Variable Info screen, change the format of the date variable to something like *mmm-dd-yyyy* or *mm-dd-yy*. The labels will be displayed as labels. Without changing the variable format, the dates will be displayed as long integer values.

Cluster Options

Minimum and Maximum Clusters

These options specify a minimum and maximum number of clusters to try. Although the k-means algorithm finds a cluster configuration for a fixed number of clusters, ***NCSS*** lets you specify a range of values to try for the number of clusters. Various goodness-of-fit tests help you determine the optimum number of clusters.

Often, values between two and five are used here, although your data might require more.

Reported Clusters

This is the number of clusters to use for reporting purposes. This is the so-called “optimum” number of clusters. Usually, you will have to make two passes through your data. On the first pass, you will determine the optimum number of clusters. On the second pass, you will obtain the information about the clusters.

Other Options

Random Starts

The first box specifies the number of random initial configurations to try for each value between the minimum and maximum cluster range. Since the k-means algorithm finds a local optimum, it is thought that trying several random, initial configurations will lead to the global optimum (or near optimum). Of course, as this value is increased, the program’s running time also increases.

Max Iterations

This option specifies the maximum number of retries before the algorithm is aborted.

Percent Missing

An observation with missing values may be clustered by using only the non-missing data. This option specifies the percentage of missing values to allow in an observation before it is skipped. For example, an observation with five variables, with two values missing, would be 60% complete. If the value of this option were 50, this observation would be kept, while an observation with three missing values would be skipped.

Reports Tab

The following options control the format of the reports.

Select Reports

Minimum Iteration Report - Bivariate Plots

These options specify which reports and plots are displayed.

Report Options

Precision

This allows you to 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.

Bivariate Plots Tab

These options control the attributes of the dendrogram.

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.

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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Bivariate 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 Numbers and Labels

Specify whether to display row numbers and labels on plots.

Bivariate Plot Settings - Legend

Show Legend

Specify whether you want to view a legend.

Legend Text

Specifies the legend title.

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.

Symbols Tab

These options let you specify the plotting symbols for the groups.

Plotting Symbols

Cluster (1-15)

Specifies the plotting symbols used for each of the first fifteen clusters.

Storage Tab

These options let you specify where to store various row-wise statistics.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the number of clusters specified in the Reported Clusters option.

Warning: Any data already in this variable is replaced by the cluster number. Be careful not to specify variables that contain important data.

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 – K-Means Clustering

This section presents an example of how to run a K-Means cluster analysis. The data used are shown above and found in the BBALL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the K-Means Clustering window.

1 Open the BBall dataset.

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

2 Open the K-Means Clustering window.

- On the menus, select **Analysis**, then **Clustering**, then **K-Means**. The K-Means Clustering 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 K-Means Clustering window, select the **Variables tab**.
- Double-click in the **Cluster Variables** box. This will bring up the variable selection window.
- Select **Height, FgPct, Points, Rebounds** from the list of variables and then click **Ok**. "Height, FgPct, Points, Rebounds" will appear in the Cluster Variables box.
- Double-click in the **Label Variable** box. This will bring up the variable selection window.

- Select **Player** from the list of variables and then click **Ok**. “Player” will appear in the Label Variable box.
- Enter **4** for the **Maximum Clusters**.

4 Specify the report.

- On the K-Means Clustering window, select the **Reports tab**.
- All reports and plots should be selected.

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).

Minimum Iteration Section

```
Minimum Iteration Section

Iteration        No. of        Percent of     Bar Chart
No.              Clusters      Variation      of Percent
2                2             66.09          ||||||||||||||||||||
4                3             46.48          ||||||||||||||
8                4             29.17          |||||||||
```

This report is to help you determine the optimum number of clusters.

Iteration No.

The iteration number from the Iteration Report.

No. of Clusters

The number of clusters reported on.

Percent of Variation

This gives the within sum of squares for the number of clusters reported on in this line as a percentage of the within sum of squares with no clustering. As more and more clusters are added, this value should fall. Select as the optimum number of clusters the point where this percentage fails to decrease dramatically.

Bar Chart of Percent

This gives a visual display of the Percent of Variation values.

Iteration Section

```
Minimum Iteration Section

Iteration      No. of       Percent of   Bar Chart
No.            Clusters     Variation    of Percent
1              2            72.02716     ||||||||||||||||||||||
2              2            66.08894     ||||||||||||||||||||
3              2            70.51944     ||||||||||||||||||||||
4              3            46.47721     ||||||||||||||
5              3            47.91439     |||||||||||||||
6              3            46.47721     ||||||||||||||
7              4            31.81219     ||||||||||
8              4            29.17248     |||||||||
9              4            32.96159     ||||||||||
```

This report is especially useful in helping you determine if you have selected enough random starting configurations. If you have specified enough starting configurations, two or three of them will be optimum (minimum percent variation) for each number of clusters. If this does not occur, you should increase the number of random starting configurations (Initial Configurations) and re-run the problem.

Iteration No.

The iteration number reported on this line.

No. of Clusters

The number of clusters in this configuration.

Percent of Variation

This gives the within sum of squares for the number of clusters reported on in this line as a percentage of the within sum of squares with no clustering. As more and more clusters are added, this value should fall. Select as the optimum number of clusters the point where this percentage fails to decrease dramatically.

Bar Chart of Percent

This gives a visual display of the Percent of Variation values.

Cluster Means

```
Cluster Means

Variables          Cluster1     Cluster2     Cluster3
Height             78.25        85.5         77
FGPct              48.6375      54.95        40.75
Points             25.575       27.35        16.75
Rebounds           8.225        17.05        13.9
Count              8            2            2
```

This report shows the means of each of the variables across each of the clusters. The last row shows the *count* or number of observations in the cluster.

Cluster Standard Deviations

Cluster Standard Deviations

Variables	Cluster1	Cluster2	Cluster3
Height	2.171241	0.7071068	6.363961
FGPct	3.357694	1.343503	4.596194
Points	3.770089	3.889087	2.333452
Rebounds	2.544321	8.273149	12.30366
Count	8	2	2

This report shows the standard deviations of each of the variables across each of the clusters. The last row shows the count (number of observations) in the cluster.

F-Ratio Section

F-Ratio Section

			Between	Within		Prob
Variables	DF1	DF2	Mean Square	Mean Square	F-Ratio	Level
Height	2	9	48.125	8.222222	5.85	0.023532
FGPct	2	9	101.6469	11.31653	8.98	0.007170
Points	2	9	72.7475	13.34056	5.45	0.028096
Rebounds	2	9	75.04459	29.46	2.55	0.132844

This report summarizes the results of performing a one-way ANOVA on each variable, using the currently defined clusters as the factor. This report helps you investigate the importance of each variable in the clustering process.

Caution should be used with this report since it ignores the correlation that exists among the variables. A better approach to reducing the number of variables would be to save the cluster configuration and run a Discriminant Analysis with variable selection, since this would account for the correlation among the variables.

Distance Section

Distance Section

Row	Cluster	Dist1	Dist2	Dist3
1 Jabbar K.A.	2	2.4609	1.1263	4.0315
2 Barry R	1	0.9139	3.1499	1.9940
3 Baylor E	1	1.4427	3.1724	2.2139
4 Bird L	1	0.8398	1.8867	2.7392
5 Chamberlain W	2	3.2456	1.1263	4.4712
6 Cousy B	3	2.9971	5.3790	1.9512
7 Erving J	1	0.4724	2.4891	2.5912
8 Johnson M	1	1.6497	2.5426	2.8064
9 Jordan M	1	1.5532	2.8939	4.0067
10 Robertson O	1	0.3409	2.9490	2.5629
11 Russell B	3	3.3878	3.5197	1.9512
12 West J	1	1.0971	3.6374	2.8439

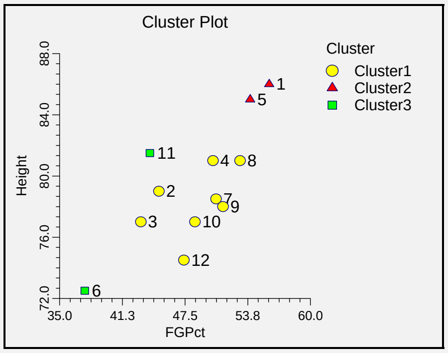
This report displays the relative distance of each row to the cluster centers. It is provided to help determine how sharp the clustering has been. If the distance from each point to its designated center is much less than the distance from the point to the other centers, the cluster configuration does a good job of clustering. However, if the smallest distance is close in value to the distance to one of the other clusters, there is ambiguity as to which cluster the point belongs. Such a solution is not as desirable.

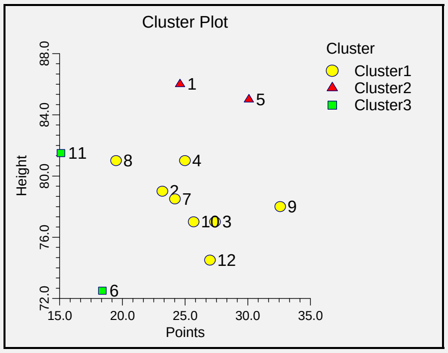
Individual Distance Section

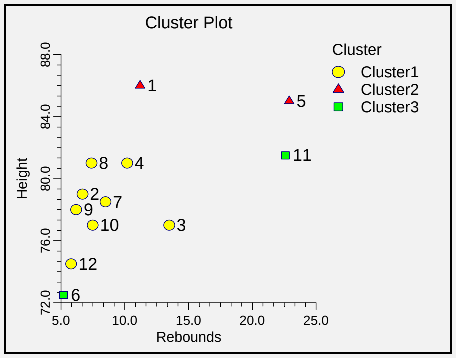
```
Distance Section for Cluster 1

Row                 Cluster      Dist1      Dist2      Dist3
2 Barry R                 1     0.9139     3.1499     1.9940
3 Baylor E                1     1.4427     3.1724     2.2139
4 Bird L                  1     0.8398     1.8867     2.7392
7 Erving J                1     0.4724     2.4891     2.5912
8 Johnson M               1     1.6497     2.5426     2.8064
9 Jordan M                1     1.5532     2.8939     4.0067
10 Robertson O            1     0.3409     2.9490     2.5629
12 West J                 1     1.0971     3.6374     2.8439
Count = 8

Distance Section for Cluster 2

Row                 Cluster      Dist1      Dist2      Dist3
1 Jabbar K.A.             2     2.4609     1.1263     4.0315
5 Chamberlain W           2     3.2456     1.1263     4.4712
Count = 2

Distance Section for Cluster 3

Row                 Cluster      Dist1      Dist2      Dist3
6 Cousy B                 3     2.9971     5.3790     1.9512
11 Russell B              3     3.3878     3.5197     1.9512

Count = 2
```

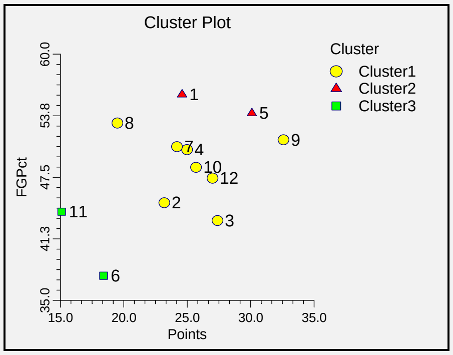
These sections show the same distances as in the previous distance report, except that the rows from only one cluster at a time are displayed. This makes it easier to see which items fell into each cluster.

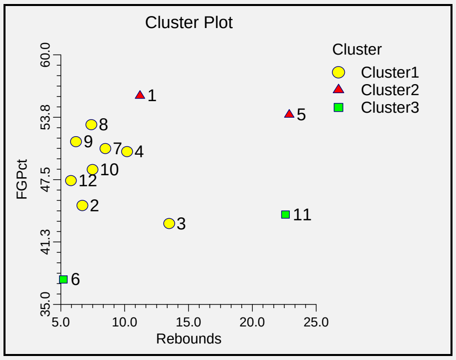
Bivariate Plots Section

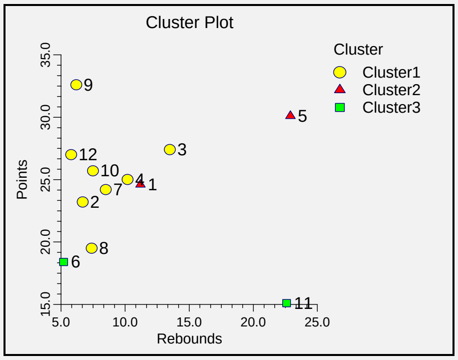












This series of scatter plots shows the data for each pair of variables with different clusters shown with different plotting symbols. The row numbers may be displayed at the side of plot symbols to help identify problem observations.

These plots will help you find outliers, anomalies, and various other problems. Note that because of the multivariate nature of the data, your cluster configuration may be good yet still show little pattern in these plots.

Chapter 447

Medoid Partitioning

Introduction

The objective of cluster analysis is to partition a set of objects into two or more clusters such that objects within a cluster are similar and objects in different clusters are dissimilar. The medoid partitioning algorithms presented here attempt to accomplish this by finding a set of representative objects called *medoids*. The *medoid* of a cluster is defined as that object for which the average dissimilarity to all other objects in the cluster is minimal. If *k* clusters are desired, *k* medoids are found. Once the medoids are found, the data are classified into the cluster of the nearest medoid.

Two algorithms are available in this procedure to perform the clustering. The first, from Spath (1985), uses random starting cluster configurations. The second, from Kaufman and Rousseeuw (1990), makes special use of silhouette statistics to help determine the appropriate number of clusters. Both of these algorithms will be explained in more later.

Dissimilarities

The fundamental value used in cluster analysis is the dissimilarity between two objects. This section discusses how the dissimilarity is computed for the various types of data.

For multivariate data, a critical issue is how the distance between individual variables is combined to form the overall dissimilarity. This depends on the variable type, scaling type, and distance type that is selected.

We begin with a brief discussion of the possible types of variables.

Types of Cluster Variables

Interval Variables

Interval variables are continuous measurements that follow a linear scale. Examples include height, weight, age, price, temperature, and time. These values may be positive or negative.

Ordinal Variables

Ordinal variables are measurements that may be ordered according to magnitude. For example, a survey question may require you to pick one of five possible choices: strongly disagree (5), disagree (4), neutral (3), agree (2), or strongly agree (1).

Ratio Variables

Ratio variables are positive measurements in which the distinction between two numbers is constant if their ratio is constant. For example, the distinction between 3 and 30 would have the same meaning as the distinction between 30 and 300. Examples are chemical concentration or radiation intensity.

Nominal Variables

Nominal variables are those in which the number represents the state of the variable, but does not represent magnitude. The number is used for identification purposes only. Examples include gender, race, hair color, city of birth, or zipcode.

Symmetric-Binary Variables

Symmetric-binary variables have two possible outcomes, each of which carry the same information and weight. Examples include gender, marital status, or membership in a particular group. Usually, they are coded as 1 for yes and 0 for no, although this is not necessary.

Assymetric-Binary Variables

Asymmetric-binary variables are concerned with the presence or absence of a relatively rare event, the absence of which is rather unimportant and uninformative. For example, if a person has a scar on his face, he might be more easily identified. But if you know the person does not have a scar, that will not help you identify him.

Distance Calculation

The dissimilarity (distance) between two objects is fundamental to cluster analysis since the techniques goal is to place similar objects in the same cluster and dissimilar objects in different clusters. Unfortunately, the measurement of dissimilarity depends on the type of variable. For interval variables, the distance between to objects is simply the difference in their values. However, how do you quantify the difference between males and females? Is it simply 1 - 0 = 1? How do you combine the difference between males and females with the difference in age to form an overall dissimilar? These are the questions that will be answered in this section. This discussion follows Kaufman and Rousseeuw (1990) very closely.

Assume that you have N rows (observations) which are separated to be clustered into K groups. Each row consists of P variables. Two types of distance measures are available in the program: Euclidean and Manhattan.

The *Euclidean distance* d_{jk} between rows j and k is computed using

$$d_{jk} = \sqrt{\frac{\sum_{i=1}^{P} \delta_{ijk}^{2}}{P}}$$

and *Manhattan distance* d_{jk} between rows *j* and *k* is computed using

$$d_{jk} = \frac{\sum_{i=1}^{P} \left| \delta_{ijk} \right|}{P}$$

where for interval, ordinal, and ratio variables

$$\delta_{ijk} = z_{ij} - z_{ik}$$

and for asymmetric-binary, symmetric-binary, and nominal variables

$$\delta_{ijk} = \begin{cases} 1 & if \ \ x_{ij} \neq x_{ik} \\ 0 & if \ \ x_{ij} = x_{ik} \end{cases}$$

with the exception that for asymmetric-binary, the variable is completely ignored (*P* is decreased by one for this row) if both x_{ij} and x_{ik} are equal to zero (the non-rare event).

The value of z_{ij} for interval, ordinal, and ratio variables is defined next.

Interval Variables

You most likely have variables with several different scales. For example, you might have percentages, ages, rates, income levels, and so on. In order to remove distortions due to these differences in scales, the data are transformed to a common scale.

Four types of scaling are available: absolute value, standard deviation, range, and none. Each of these have the general form:

$$z_{ij} = \frac{x_{ij} - A_i}{B_i}$$

where x_{ij} represents the original data value for variable *i* and row *j* and z_{ij} represents the corresponding scale value. The scaling choice determines the values used for A_i and B_i.

The following table shows the scaling mechanism used for each type of scaling.

Type of Scaling	**Value of A_i**	**Value of B_i**
Absolute Value	$\frac{\sum_{j=1}^{N} x_{ij}}{N}$	$\frac{\sum_{j=1}^{N} \left\| x_{ij} - A_i \right\|}{N}$
Standard Deviation	$\frac{\sum_{j=1}^{N} x_{ij}}{N}$	$\sqrt{\frac{\sum_{j=1}^{N} \left(x_{ij} - A_i \right)^2}{N - 1}}$
Range	$\underset{over\ j}{Min}\left(x_{ij} \right)$	$\underset{over\ j}{Max}\left(x_{ij} \right) - \underset{over\ j}{Min}\left(x_{ij} \right)$
None	0	1

Ordinal and Ratio Variables

The distance calculations for the ordinal and ratio variables are the same as for interval variables except that the values are transformed to an interval scale before distance calculations begin. The ranks of the ordinal variables and the natural logarithms of the ratio variables are substituted for the actual values. Once these transformations are made, the interval distance formulas are used.

Algorithm Details

Medoid Algorithm of Spath

The first medoid algorithm is presented in Spath (1985). The method minimizes an objective function by swapping objects from one cluster to another. Beginning at a random starting configuration, the algorithm proceeds to a local minimum by intelligently moving objects from one cluster to another. When no object moving would result in a reduction of the objective function, the procedure terminates. Unfortunately, this local minimum is not necessarily the global minimum. To overcome this limitation, the program lets you rerun the algorithm using several random starting configurations and the best solution is kept.

The objective function D is the total distance between the objects within a cluster. Mathematically, it is represented as follows:

$$D = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij}$$

where K is the number of clusters, d_{ij} is the distance between objects i and j, and C_k is the set of all objects in cluster k.

Medoid Algorithm of Kaufman and Rousseeuw

Kaufman and Rousseeuw (1990) present a medoid algorithm which they call PAM (Partition Around Medoids). This algorithm also attempts to minimize the total distance D (formula given above) between objects within each cluster. The algorithm proceeds through two phases.

In the first phase, a representative set of k objects is found. The first object selected has the shortest distance to all other objects. That is, it is in the center. An addition k-1 objects are selected one at a time in such a manner that at each step, they decrease D as much as possible.

In the second phase, possible alternatives to the k objects selected in phase one are considered in an iterative manner. At each step, the algorithm searches the unselected objects for the one that if exchanged with one of the k selected objects will lower the objective function the most. The exchange is made and the step is repeated. These iterations continue until no exchanges can be found that will lower the objective function.

Note that all potential swaps are considered and that the algorithm does not depend on the order of the objects on the database.

Silhouettes

Two of the most difficult tasks in cluster analysis are deciding on the appropriate number of clusters and deciding how to tell a bad cluster from a good one. Kaufman and Rousseeuw (1990) define a set of values called *silhouettes* that provide key information about both of these tasks. First, we will explain how these are calculated and then we will show how they are used.

Calculating Silhouettes

A silhouette value *s* is constructed for each object as follows.

1. Consider a particular object *i* which is in cluster *A*. Compute the value

   a = average dissimilarity of *i* to all other objects in *A*

   If *A* contains only one object, set *a* to zero.

2. For every other cluster not equal to *A*, find the cluster *B* that has the smallest average dissimilarity between its objects and *i*. Set

   b = average dissimilarity between *i* and the object in *B*.

   The cluster *B* is the nearest neighbor of object *i*.

3. Compute the silhouette *s* of object *i* as follows:

   If *A* contains only one object, set s = 0.

   If $a < b$, $s = 1 - a/b$.

   If $a > b$, $s = b/a - 1$.

   If $a = b$, $s = 0$.

Interpreting Silhouettes

A silhouette value is constructed for each object. The value can range from minus one to one. It measures how well an object has been classified by comparing its dissimilarity within its cluster to its dissimilarity with its nearest neighbor.

When *s* is close to one, the object is well classified. Its dissimilarity with other objects in its cluster is much less than its dissimilarity with objects in the nearest cluster.

When *s* is near zero, the object was just between clusters *A* and *B*. It was arbitrarily assigned to *A*.

When *s* is close to negative one, the object is poorly classified. Its dissimilarity with other objects in its cluster is much greater than its dissimilarity with objects in the nearest cluster. Why isn’t it in the neighboring cluster?

Hence, the silhouette value summarizes how appropriate each object’s cluster is.

Determining the Number of Clusters

One useful summary statistic is the average value of *s* across all objects. This summarizes how well the current configuration fits the data. An easy way to select the appropriate number of clusters is to choose that number of clusters which maximizes the average silhouette. We denote the maximum average silhouette across all values of *k* as *SC*.

Kaufman and Rousseeuw (1990) present the following table to aid in the interpretation of *SC*.

SC	Proposed Interpretation
0.71 to 1.00	A strong structure has been found.
0.51 to 0.70	A reasonable structure has been found.
0.26 to 0.50	The structure is weak and could be artificial. Try other methods on this database.
-1 to 0.25	No substantial structure has been found.

Finding Good Clusters

A bar chart of the silhouette values, sorted by cluster number and silhouette value, will show how well each cluster is doing. These charts will be discussed more in the output section.

Further Analysis

Once a cluster analysis has been run and an appropriate solution found, the cluster numbers should be saved to an empty variable so that the cluster solution can be further analyzed. What are some additional procedures that should be run? The most common is a discriminant analysis since it will let you study the impact of each of the variables on the solution. Discriminant analysis will also quantify how well the rows have been clustered. This will show up in the Wilks' lambda statistic.

In addition to discriminant analysis, you will want to produce various scatter plots in which the cluster number is used as a grouping variable. This will greatly increase your understanding of what the clusters that have been found look like.

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable. A discussion of the types of variables will be presented shortly.

The data given in the following table contain information on twelve superstars in basketball. The stats are on a per game basis for games played through the 1989 season.

BBALL dataset (subset)

Player	Height	FgPct	Points	Rebounds
Jabbar K.A.	86.0	55.9	24.6	11.2
Barry R	79.0	44.9	23.2	6.7
Baylor E	77.0	43.1	27.4	13.5
Bird L	81.0	50.3	25	10.2
Chamberlain W	85.0	54.0	30.1	22.9
Cousy B	72.5	37.5	18.4	5.2
Erving J	78.5	50.6	24.2	8.5

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Interval Variables

Designates interval-type variables (if any) or the columns of the matrix if distance or correlation matrix input was selected. Interval variables are continuous measurements that may be either positive or negative and follow a linear scale. Examples include height, weight, age, price, temperature, and time.

In general, an interval should keep the same importance throughout the scale. For example, the length of time between 1905 and 1925 is the same as the length of time between 1995 and 2015.

Note that a nonlinear transformation of an interval variable is probably not an interval variable. For example, the logarithm of height is not an interval variable since the value of an interval along the scale changes depending upon where you are on the scale.

Ratio Variables

Specifies the ratio variables (if any). Ratio-type variables are positive measurements in which the distinction between two numbers is constant if their ratio is constant. For example, the distinction between 3 and 30 would have the same meaning as the distinction between 30 and 300. Examples are chemical concentration or radiation intensity.

The logarithms of ratio variables are analyzed as if they were interval variables.

Ordinal Variables

Specifies the ordinal-type variables (if any). Ordinal variables are measurements that may be ordered according to magnitude. For example, a survey question may require you to pick one of five possible choices: strongly disagree (5), disagree (4), neutral (3), agree (2), or strongly agree (1). Interval variables are ordinal, but ordinal variables are not necessarily interval. The original values of ordinal variables are replaced by their ranks. These ranks are then analyzed as if they were interval variables.

Nominal Variables

Specifies the nominal-type variables (if any). Nominal variables are those in which the number represents the state of the variable. Examples include gender, race, hair color, country of birth, or zipcode. If a nominal variable has only two categories, it is often called a binary variable.

Nominal variables are analyzed using the number of matches between two individuals.

Symmetric-Binary Variables

Specifies the symmetric binary-type variables (if any). Symmetric binary variables have two possible outcomes, each of which carries the same information and weight. Examples include gender, marital status, or membership in a particular group. Usually, they are coded as 1 for yes or 0 for no, although this is not necessary. These variables are analyzed using the number of matches between two individuals.

Asymmetric-Binary Variables

Specifies the asymmetric binary-type variables (if any). Asymmetric binary-scaled variables are concerned with the presence or absences of a relatively rare event, the absence of which is unimportant.

These variables are analyzed using the number of matches in which both individuals have the trait of interest. Those cases in which both individuals do not have the trait are not of interest and are ignored.

Clustering Options

Cluster Method

This option specifies which of the two medoid algorithms to be used.

- **Spath**

  Perform the analysis using Spath's medoid partitioning algorithm. This algorithm was discussed in the introduction. When this option is selected, you must also set the Best Starting Configuration, Weighting Method, and Number Random Starts options.

- **Kaufman - Rousseeuw**

  Perform the analysis using Kaufman and Rousseeuw's medoid partitioning algorithm. This algorithm was discussed in the introduction.

Distance Method

This option specifies with Euclidean or Manhattan distance is used. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Scaling Method

Specify the type of scaling to be used from Interval, Ordinal, and Ratio variables. Possible choices are Standard Deviation, Average Absolute Deviation, Range, and None. These were discussed in the introduction to this chapter.

Max Iterations

This option sets a maximum number of iterations that are attempted before the algorithm terminates. This avoids the possible of the algorithm going into an infinite loop.

Clustering Options – Number of Clusters

Minimum Clusters

The minimum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters. The actual number of clusters used is set above by the Reported Clusters option.

Maximum Clusters

The maximum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters. The actual number of clusters used is set above by the Reported Clusters option.

Reported Clusters

This is the number of clusters to be reported on. Although the program can find results for a range of cluster sizes, this option set the size that is actually used. It is used in the Row Detail section and by the Storage Tab section.

Clustering Options – Spath Cluster Method Options

Number Random Starts

This is the number of random starting configurations that are attempted during Spath's algorithm. Usually, ten starting configurations should be enough.

Best Starting Configuration

This option applies to Method = Spath only. In Spath's algorithm, a number of random starting configurations are tried and the best in for each cluster size is retained. This option determines which statistic is used to indicate the best.

- **Mean Distance**

  The configuration with the smallest average dissimilarity is selected.

- **Silhouette**

  The configuration with the largest average silhouette is selected.

Weighting Method

This option designates which objective function is minimized during Spath's algorithm. Two types are possible.

- **Regular**

  Minimize the sum of the distances between all individuals within each cluster.

- **Weighted**

  Minimize the weighted sum of the distances between all individuals within each cluster. The weights are one over the number of objects in the cluster.

Format Options

Label Variable

This is an optional variable containing identification for each row (object). These labels are used to enhance the interpretability of the reports.

Input Format

Specify the type of data format that you have. Your choices are

- **Raw Data**

  The variables are in the standard format in which each row represents an object and each column represents a variable.

- **Distances**

  The variables containing a distance matrix are specified in the Interval Variables option. Note that this matrix contains the distances between each pair of objects. Each object is represented by a row and the corresponding column. Also, the matrix must be complete. You cannot use only the lower triangular portion, for example.

- **Correlations 1**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = \frac{1 - r_{ij}}{2}$$

- **Correlations 2**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - \left| r_{ij} \right|$$

- **Correlations 3**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - r_{ij}^2$$

  Note that all three types of correlation matrices must be completely specified. You cannot specify only the lower or upper triangular portions. Also, the rows correspond to variables. That is, the values along the first row represent the correlations of the first variable with each of the other variables. Hence, you cannot rearrange the order of the matrix.

Reports Tab

The following options control the formatting of the reports.

Select Reports

Iteration Report - Row Detail Report

Specify whether to display the indicated reports.

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.

Storage Tab

Storage Variable

These options let you specify where to store various row-wise statistics.

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the number of clusters specified in the Reported Clusters option.

Warning: Any data already in this variable are replaced by the cluster number. Be careful not to specify variables that contain important data.

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 – Medoid Partitioning

This section presents an example of how to run a medoid partitioning analysis. The data used were shown above and are found in the BBALL database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Medoid Partitioning window.

1 Open the BBall dataset.

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

2 Open the Medoid Partitioning window.

- On the menus, select **Analysis**, then **Clustering**, then **Medoid Partitioning**. The Medoid Partitioning 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 Medoid Partitioning window, select the **Variables tab**.
- Double-click in the **Interval Variables** box. This will bring up the variable selection window.
- Select **Height, Weight, FgPct, FtPct, Points, Rebounds** from the list of variables and then click **Ok**. “Height-Points,Rebounds” will appear in the Interval Variables box.
- Double-click in the **Label Variables** box. This will bring up the variable selection window.
- Select **Player** from the list of variables and then click **Ok**. “Player” will appear in the Label Variable box.
- Enter **2** for **Reported Clusters**.
- Enter **4** for **Number Random Starts**.

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).

Iteration Detail Section

```
Iteration Detail Section

              (Minimize This)     Adjusted  (Maximize This)
Number                Average      Average          Average
Clusters             Distance     Distance       Silhouette
2                   35.977405     5.996234         0.135735
2                   34.352873     5.725479         0.185579
2                   34.862052     5.810342         0.170356
2                   36.031237     6.005206         0.101405
3                   19.525066     4.881267         0.094407
3                   21.106317     5.276579         0.033435
3                   19.005957     4.751489         0.045621
3                   22.202362     5.550590        -0.026350
4                   12.547872     4.182624        -0.013869
4                   12.318440     4.106147         0.044989
4                   12.210147     4.070049         0.018876
4                   14.209356     4.736452        -0.097672
5                    9.344940     3.893725        -0.099737
5                   10.556815     4.398673        -0.189487
5                    8.274123     3.447551        -0.045335
5                    8.049819     3.354091        -0.004580
```

This report shows the values of the objective functions for each iteration and number of clusters. This report is only generated when the Method option is set to Spath.

The report is especially useful in determining if you have set the number of random starts correctly. If you can see that two or three configurations at the desired number of clusters are identical then you have set the Number Random Starts large enough. Otherwise, you should increase this value and rerun the analysis.

In this example, we will conclude that *k* is two (determined from a later report). However, we notice that we have not achieved the maximum silhouette value (0.185579) more than once. We should change the Number Replications options to ten and rerun the analysis.

Average Distance

This is the value of the average dissimilarity. It is computed using

$$D = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij}$$

Note that this value has been rescaled as a percentage from the maximum distance in the dissimilarity matrix to improve readability.

Adjusted Average Distance

This is the value of the adjusted average dissimilarity. It is computed using

$$D_{adjusted} = \frac{K}{N} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij}$$

Note that this value has been rescaled as a percentage from the maximum distance in the dissimilarity matrix to improve readability.

Average Silhouette

This is the average of the silhouette values of all rows.

Iteration Summary Section

```
Iteration Summary Section

                 (Minimize This)          Adjusted    (Maximize This)
Number                   Average           Average            Average
Clusters                Distance          Distance         Silhouette
2                      34.352873          5.725479           0.185579
3                      19.525066          4.881267           0.094407
4                      12.318440          4.106147           0.044989
5                       8.049819          3.354091          -0.004580
```

This report shows the values of the objective functions for each number of clusters.

This report is used to determine the appropriate number of clusters. The number selected corresponds to the maximum value of the last (Average Silhouette) column. Usually, the row selected will have a respectable value of the Adjusted Average Distance (this value should be near its minimum).

The definitions of the columns were given above and will not be repeated here.

Cluster Medoids Section

```
Cluster Medoids Section

Variable        Cluster1         Cluster2
Height          86               77
Weight          230              210
FgPct           55.9             48.5
FtPct           72.1             83.8
Points          24.6             25.7
Rebounds        11.2             7.5
Row             1 Jabbar K.A     10 Robertson
```

This report gives the medoid (most centrally located) of each cluster. It is provided to help you interpret and recognize each cluster. The last row of the report gives the row number (and label if designated) of the each cluster’s medoid.

Notice that the players in cluster one are typically nine inches taller and pull down about four more rebounds than the players in cluster two. Apparently, cluster one represents centers (or tall forwards) and cluster two represents other players.

Row Detail Section

```
Row Detail Section
                                             Average        Average
                              Nearest       Distance       Distance      Silhouette      Silhouette
Row              Cluster      Neighbor        Within       Neighbor           Value      Bar
5 Chamberlai     1            2                58.85          74.48          0.2098      |||||||||||
11 Russell B     1            2                58.40          57.83         -0.0098      |
1 Jabbar K.A     1            2                47.59          44.28         -0.0696      |
3 Baylor E       1            2                47.74          31.69         -0.3363      |
9 Jordan M       1            2                56.34          32.62         -0.4210      |
Cluster Average  1            (5)              53.79          48.18         -0.1254

2 Barry R        2            1                24.20          47.65          0.4921      ||||||||||||||||||||||||||
10 Robertson     2            1                21.94          42.43          0.4830      |||||||||||||||||||||||||
12 West J        2            1                29.13          51.74          0.4370      |||||||||||||||||||||||
7 Erving J       2            1                23.03          40.90          0.4369      |||||||||||||||||||||||
6 Cousy B        2            1                45.02          68.58          0.3435      ||||||||||||||||||
8 Johnson M      2            1                30.31          45.50          0.3339      ||||||||||||||||||
4 Bird L         2            1                27.19          40.44          0.3275      |||||||||||||||||
Cluster Average  2            (7)              28.69          48.18          0.4077

Overall Average               (12)             39.15          48.18          0.1856      = SC
```

This report displays information about each row that was clustered. The report is sorted by Silhouette Value within cluster.

Row

The row number and, if designated, label of this individual. Each row of the database is represented on this report.

Cluster

This is the number of the cluster into which this row was classified.

Nearest Neighbor

This is the identification number of the nearest cluster to this row (other than the one that it is in). This information is used in computing the silhouette value.

Average Distance Within

This is the average distance between this object and all other objects in the cluster. This is the value of *a* in the computation of the silhouette.

Average Distance Neighbor

This is the average distance between this object and the objects in the nearest neighbor. This is the value of *b* in the computation of the silhouette.

Silhouette Value

This is the value of the silhouette. Its interpretation was presented in the introduction and will not be repeated here. We note that the value should be positive and most rows should be greater than 0.50. The fact that several of the rows in this analysis have negative silhouette values would cause us to toss out this cluster configuration and look for a better one.

Chapter 448

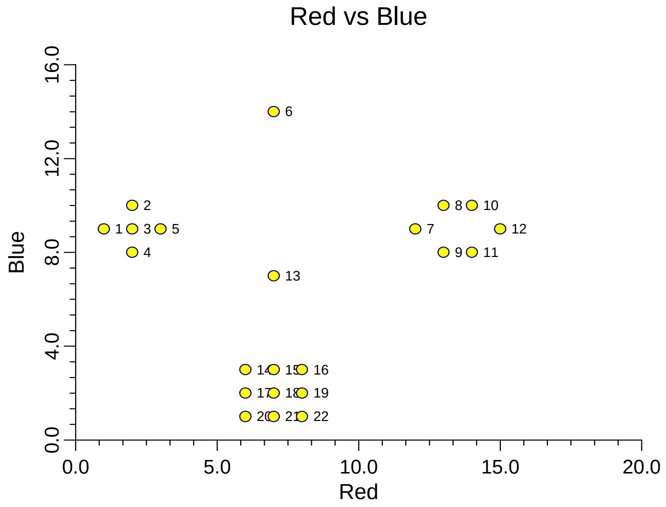
Fuzzy Clustering

Introduction

Fuzzy clustering generalizes partition clustering methods (such as k-means and medoid) by allowing an individual to be partially classified into more than one cluster. In regular clustering, each individual is a member of only one cluster. Suppose we have *K* clusters and we define a set of variables $m_{i1}, m_{i2}, \cdots, m_{iK}$ that represent the probability that object *i* is classified into cluster *k*. In partition clustering algorithms, one of these values will be one and the rest will be zero. This represents the fact that these algorithms classify an individual into one and only one cluster.

In fuzzy clustering, the membership is spread among all clusters. The m_{ik} can now be between zero and one, with the stipulation that the sum of their values is one. We call this a *fuzzification* of the cluster configuration. It has the advantage that it does not force every object into a specific cluster. It has the disadvantage that there is much more information to be interpreted.

To understand the reason that fuzzy clustering was developed, consider the following two-variable database whose values are plotted below.



The data have three obvious clusters and two outlier points (6 and 13). A regular clustering algorithm searching for three clusters will force these two points into specific clusters. This may cause distortion in the final solution. Fuzzy clustering, however, will assign a probability of about 0.33 for each cluster. This equal membership probability signals that these two points are outliers.

When you only have two variables, you can plot your data and see what the clusters are. Unfortunately, most clustering projects come with more than two variables, so plotting is not possible. Hence, we must use techniques like fuzzy clustering to deal with the anomalies that can occur.

Dissimilarities

The formation of the distances (dissimilarities) was described in the Medoid Clustering chapter and is not repeated here.

Fuzzy Algorithm

The fuzzy algorithm used by this program is described in Kaufman (1990). It seeks to minimize the following objective function, *C*, made up of cluster memberships and distances.

$$C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} m_{ik}^{2} m_{jk}^{2} d_{ij}}{2 \sum_{j=1}^{N} m_{jk}^{2}}$$

where m_{ik} represents the unknown membership of the object *i* in cluster *k* and d_{ij} is the dissimilarity between objects *i* and *j*. The memberships are subject to constraints that they all must be non-negative and that the memberships for a single individual must sum to one. That is, the memberships have the same constraints that they would if they were the probabilities that an individual belongs to each group (and they may be interpreted as such).

Goodness-of-Fit

One of the most difficult tasks in cluster analysis is choose the appropriate number of clusters. In fuzzy clustering, the following coefficients are used in conjunction with the silhouette values that are defined in the Medoid Clustering chapter.

The amount of ‘fuzziness’ in a solution may be measured by *Dunn’s partition coefficient* which measures how close the fuzzy solution is to the corresponding hard solution. This *hard* solution is formed by classifying each object into the cluster which has the largest membership. The formula for Dunn’s partition coefficient is

$$F(U) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N} m_{ik}^{2}$$

This coefficient ranges from 1/*K* to 1. Its value is 1/*K* when all memberships are equal to 1/*K*. The value of one results when, for each object, the value of one membership is unity and the rest are zero.

Dunn’s partition coefficient may be normalized so that it varies from 0 (completely fuzzy) to 1 (hard cluster). The normalized version is

$$Fc(U) = \frac{F(U) - (1/K)}{1 - (1/K)}$$

Another partition coefficient, given in Kaufman (1990), is

$$D(U) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}(h_{ik} - m_{ik})^2$$

This coefficient ranges from 0 (hard clusters) to 1-1/*K* (completely fuzzy). The normalized version of this equation is:

$$Dc(U) = \frac{D(U)}{1 - (1 / K)}$$

Fc(U) and Dc(U) together give a good indication of an optimum number of clusters. You should choose *K* so that Fc(U) is large and Dc(U) is small.

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable.

The data given in the following table were shown on the scatter plot displayed earlier and are found in the FUZZY database. They are from a concocted database found in Kaufman (1990) designed specifically to show the usefulness of fuzzy clustering.

FUZZY dataset (subset)

Red	Blue	ID
1	9	1
2	10	2
2	9	3
2	8	4
3	9	5
7	14	6
12	9	7
13	10	8
13	8	9

Missing Values

When an observation has missing values, appropriate adjustments are made so that the average dissimilarity across all variables with data may be computed. That is, rows with missing values are not omitted unless all variables have missing values. Note that the distances require that at least one variable have non-missing values for each pair of rows.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Interval Variables

Designates interval-type variables (if any) or the columns of the matrix if distance or correlation matrix input was selected. Interval variables are continuous measurements that may be either positive or negative and follow a linear scale. Examples include height, weight, age, price, temperature, and time.

In general, an interval should keep the same importance throughout the scale. For example, the length of time between 1905 and 1925 is the same as the length of time between 1995 and 2015.

Note that a nonlinear transformation of an interval variable is probably not an interval variable. For example, the logarithm of height is not an interval variable since the value of an interval along the scale changes depending upon where you are on the scale.

Ratio Variables

Specifies the ratio variables (if any). Ratio-type variables are positive measurements in which the distinction between two numbers is constant if their ratio is constant. For example, the distinction between 3 and 30 would have the same meaning as the distinction between 30 and 300. Examples are chemical concentration or radiation intensity.

The logarithms of ratio variables are analyzed as if they were interval variables.

Ordinal Variables

Specifies the ordinal-type variables (if any). Ordinal variables are measurements that may be ordered according to magnitude. For example, a survey question may require you to pick one of five possible choices: strongly disagree (5), disagree (4), neutral (3), agree (2), or strongly agree (1). Interval variables are ordinal, but ordinal variables are not necessarily interval.

The original values of ordinal variables are replaced by their ranks. These ranks are then analyzed as if they were interval variables.

Nominal Variables

Specifies the nominal-type variables (if any). Nominal variables are those in which the number represents the state of the variable. Examples include gender, race, hair color, country of birth, or zipcode. If a nominal variable has only two categories, it is often called a binary variable.

Nominal variables are analyzed using the number of matches between two individuals.

Symmetric-Binary Variables

Specifies the symmetric binary-type variables (if any). Symmetric binary variables have two possible outcomes, each of which carry the same information and weight. Examples include gender, marital status, or membership in a particular group. Usually, they are coded as 1 for yes or 0 for no, although this is not necessary. These variables are analyzed using the number of matches between two individuals.

Asymmetric-Binary Variables

Specifies the asymmetric binary-type variables (if any). Asymmetric binary-scaled variables are concerned with the presence or absences of a relatively rare event, the absence of which is unimportant.

These variables are analyzed using the number of matches in which both individuals have the trait of interest. Those cases in which both individuals do not have the trait are not of interest and are ignored.

Clustering Options

Distance Method

This option specifies whether Euclidean or Manhattan distance is used. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Scaling Method

Specify the type of scaling to be used on Interval, Ordinal, and Ratio variables. Possible choices are Standard Deviation, Average Absolute Deviation, Range, and None. These were discussed in the introduction to this chapter.

Max Iterations

This option sets a maximum number of iterations that are attempted before the algorithm terminates. This avoids the possible of the algorithm going into an infinite loop.

Fuzzifier Constant

Specifies the exponent of the memberships in the objective function that is being minimized. Normally, this value is set to two. In some situations, you may want to change this value. The value must be strictly greater than one. As this value is decreased from two towards one, the final solution will appear less and less fuzzy. That is, the membership values will be closer to either zero or one. Also, values of this option near one cause the algorithm to converge more slowly.

Minimum Change

When the change in the objective function from one iteration to the next is less than this amount, the algorithm terminates.

Maximum row

The maximum number of rows that will be analyzed by this procedure.

Clustering Options – Numbers of Clusters

Minimum Clusters

The minimum value of *K* to search. A separate analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters.

Maximum Clusters

The maximum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters.

Reported Clusters

The is the cluster configuration that is stored on the database if the Cluster Id or Membership Out Variables options are specified..

Format Options

Label Variable

This is an optional variable containing identification for each row (object). These labels are used to enhance the interpretability of the reports.

Input Format

Specify the type of data format that you have. Your choices are

- **Raw Data**

  The variables are in the standard format in which each row represents an object and each column represents a variable.

- **Distances**

  The variables containing a distance matrix are specified in the Interval Variables option. Note that this matrix contains the distances between each pair of objects. Each object is represented by a row and the corresponding column. Also, the matrix must be complete. You cannot use only the lower triangular portion, for example.

- **Correlations 1**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = \frac{1 - r_{ij}}{2}$$

- **Correlations 2**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - \left| r_{ij} \right|$$

- **Correlations 3**

  The variables containing a correlation matrix are specified in the Interval Variables option. Correlations are converted to distances using the formula:

$$d_{ij} = 1 - r_{ij}^2$$

Note that all three types of correlation matrices must be completely specified. You cannot specify only the lower or upper triangular portions. Also, the rows correspond to variables. That is, the values along the first row represent the correlations of the first variable with each of the other variables. Hence, you cannot rearrange the order of the matrix.

Reports Tab

The following options control the formatting of the reports.

Select Reports

Membership Report - Summary Report

Specify whether to display the indicated reports.

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.

Storage Tab

These options let you specify where to store various row-wise statistics.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the number of clusters specified in the Reported Clusters option.

Warning: Any data already in this variable are replaced by the cluster number. Be careful not to specify variables that contain important data.

Store Membership Out in Variable

You can automatically store the row memberships into the variables specified here. The configuration stored is for the number of clusters specified in the Reported Clusters 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 – Fuzzy Clustering

This section presents an example of how to run a cluster analysis. The data used found in the FUZZY database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Fuzzy Clustering window.

1 Open the Fuzzy dataset.

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

2 Open the Fuzzy Clustering window.

- On the menus, select **Analysis**, then **Clustering**, then **Fuzzy**. The Fuzzy Clustering 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 Fuzzy Clustering window, select the **Variables tab**.
- Double-click in the **Interval Variables** box. This will bring up the variable selection window.
- Select **Red** and **Blue** from the list of variables and then click **Ok**. "Red-Blue" will appear in the Interval Variables 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).

Summary Section

```
Summary Section

Number      Average      Average
Clusters    Distance     Silhouette        F(U)       Fc(U)        D(U)       Dc(U)
2          24.294968     0.535378        0.6799      0.3598      0.1400      0.2800
3          11.366128     0.704072        0.7102      0.5653      0.0861      0.1291
4           8.594860     0.487322        0.5422      0.3896      0.2161      0.2881
5           6.768054     0.340839        0.4783      0.3479      0.2837      0.3546
```

This report actually appears last on the printout, but it is the first section that should be studied. This report lets you select the appropriate number of clusters. Select the number of clusters that maximizes the Average Silhouette and Fc(U) while minimizing Dc(U). In this case, three clusters are selected.

Average Distance

This is the value of the average dissimilarity. Note that this value has been rescaled as a percentage from the maximum distance in the dissimilarity matrix to improve readability.

Average Silhouette

This is the average of the silhouette values of all rows. The Silhouette statistic is discussed in the Medoid Partitioning chapter. It is used to aid in the search for the appropriate number of clusters by selecting the number of clusters that maximizes this value.

F(U), Fc(U), D(U), Dc(U)

The definitions of these statistics were presented earlier. Here we will note that we search for the number of clusters that maximizes Fc(U) and minimizes Dc(U). There will not always be an obvious choice as in this example.

Once the appropriate number of clusters has been determined, the solution can be studied in detail. Since three clusters is appropriate for this database, only the results for three clusters will be shown here.

Cluster Medoids Section

```
Cluster Medoids Section

Variable        Cluster1        Cluster2        Cluster 3
Red             2               14              7
Blue            9               10              2
Row             3               10              18
```

This report gives the medoid (most centrally located) of the nearest hard cluster configuration. It is provided to help you recognize and interpret cluster. The last row of the report gives the row number (and label if designated) of the each cluster's medoid.

Membership Summary Section

```
Membership Summary Section
                           Sum of          Bar of
                Cluster    Squared         Squared                      Silhouette  Silhouette
Row   Cluster   Membership Memberships     Memberships                  Amount      Bar
3     1         0.9362     0.8786          |IIIIIIIIIIIIIIIIIIIIIIIIII  0.7337      |IIIIIIIIIIIIIIIIIIIIII
2     1         0.8785     0.7793          |IIIIIIIIIIIIIIIIIIIIIII     0.7313      |IIIIIIIIIIIIIIIIIIIIII
5     1         0.8741     0.7722          |IIIIIIIIIIIIIIIIIIIIIII     0.6840      |IIIIIIIIIIIIIIIIIIIII
1     1         0.8677     0.7618          |IIIIIIIIIIIIIIIIIIIIIII     0.6957      |IIIIIIIIIIIIIIIIIIIII
4     1         0.8606     0.7507          |IIIIIIIIIIIIIIIIIIIIIII     0.6400      |IIIIIIIIIIIIIIIIIII
6     1         0.4205     0.3531          |IIIIIIIIIII                 0.1392      |IIII
10    2         0.8745     0.7727          |IIIIIIIIIIIIIIIIIIIIIII     0.8284      |IIIIIIIIIIIIIIIIIIIIIIIII
8     2         0.8718     0.7683          |IIIIIIIIIIIIIIIIIIIIIII     0.8168      |IIIIIIIIIIIIIIIIIIIIIIIII
11    2         0.8613     0.7517          |IIIIIIIIIIIIIIIIIIIIIII     0.8033      |IIIIIIIIIIIIIIIIIIIIIIII
9     2         0.8564     0.7439          |IIIIIIIIIIIIIIIIIIIIII      0.7854      |IIIIIIIIIIIIIIIIIIIIIIII
12    2         0.8386     0.7164          |IIIIIIIIIIIIIIIIIIIII       0.8023      |IIIIIIIIIIIIIIIIIIIIIIII
7     2         0.8188     0.6870          |IIIIIIIIIIIIIIIIIIIII       0.7523      |IIIIIIIIIIIIIIIIIIIIIII
18    3         0.9196     0.8489          |IIIIIIIIIIIIIIIIIIIIIIIII   0.8228      |IIIIIIIIIIIIIIIIIIIIIIIII
21    3         0.8668     0.7602          |IIIIIIIIIIIIIIIIIIIIIII     0.7976      |IIIIIIIIIIIIIIIIIIIIIIII
19    3         0.8599     0.7492          |IIIIIIIIIIIIIIIIIIIIII      0.7840      |IIIIIIIIIIIIIIIIIIIIIIII
17    3         0.8589     0.7478          |IIIIIIIIIIIIIIIIIIIIII      0.7790      |IIIIIIIIIIIIIIIIIIIIIII
15    3         0.8524     0.7375          |IIIIIIIIIIIIIIIIIIIIII      0.7834      |IIIIIIIIIIIIIIIIIIIIIIII
22    3         0.8226     0.6924          |IIIIIIIIIIIIIIIIIIIII       0.7630      |IIIIIIIIIIIIIIIIIIIIIII
20    3         0.8222     0.6921          |IIIIIIIIIIIIIIIIIIIII       0.7604      |IIIIIIIIIIIIIIIIIIIIIII
16    3         0.8012     0.6617          |IIIIIIIIIIIIIIIIIIII        0.7444      |IIIIIIIIIIIIIIIIIIIIII
14    3         0.7992     0.6593          |IIIIIIIIIIIIIIIIIIII        0.7342      |IIIIIIIIIIIIIIIIIIIIII
13    3         0.3734     0.3393          |IIIIIIIIII                  0.1086      |III
```

This report displays information about each row. The report is sorted by Silhouette Value within cluster. Notice how well the two outliers, rows six and thirteen, stand out on this report.

Row

The row number and, if designated, label of this individual. Each row of the database is represented on this report.

Cluster

This is the number of the cluster into which this row was classified.

Cluster Membership

This is the maximum of the memberships. It is the membership value for the cluster into which this row was assigned for the hard clustering.

Sum of Squared Memberships

All memberships for a given row are squared and summed. When a row is completely assigned to a single cluster, this value will be one. When the row is equally likely to be classified into each cluster, the value will be $1/K$. Hence, rows with high values here are near the center of a cluster. Rows with low values here are outliers.

Bar of Squared Memberships

This is a bar graph of the sum of squared membership values. It will help you to detect rows that are not well clustered.

Silhouette Amount

This is the value of the silhouette. Its interpretation was presented in the introduction to the Medoid Clustering chapter and will not be repeated here. We note that the value should be positive and most rows should be greater than 0.50.

Silhouette Bar

This is a bar graph of the silhouette values. It will help you to detect rows that are not well clustered.

Membership Matrix Section

```
Membership Matrix Section

Row      Cluster    Prob in 1    Prob in 2      Prob in 3
1        1          0.8677       0.0564         0.0759
2        1          0.8785       0.0551         0.0664
3        1          0.9362       0.0274         0.0364
4        1          0.8606       0.0562         0.0832
5        1          0.8741       0.0549         0.0709
6        1          0.4205       0.3545         0.2250
7        2          0.0849       0.8188         0.0963
.        .          .            .              .
.        .          .            .              .
.        .          .            .              .
```

This report displays the membership of each row in each cluster.

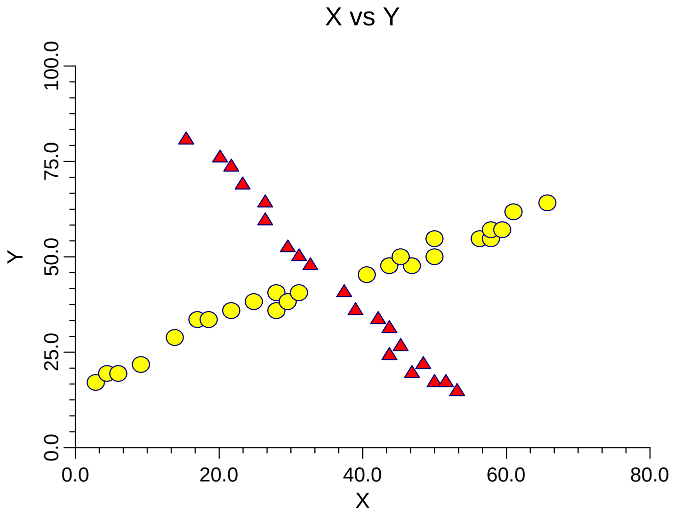
Chapter 449

Regression Clustering

Introduction

This algorithm provides for clustering in the multiple regression setting in which you have a dependent variable *Y* and one or more independent variables, the *X's*. The algorithm partitions the data into two or more clusters and performs an individual multiple regression on the data within each cluster. It is based on an exchange algorithm described in Spath (1985).

The following chart shows data that were clustered using this algorithm. Notice how the two clusters actually intersect.



Regression Exchange Algorithm

This algorithm is fairly simple to describe. The number of clusters, K, for a given run is fixed. The rows are randomly sorted into the groups to form K initial clusters. An exchange algorithm is applied to this initial configuration which searches for the rows of data that would produce a maximum decrease in a least-squares penalty function (that is, maximizing the increase in R-squared at each step). The algorithm continues until no beneficial exchange of rows can be found.

Our experience with this algorithm indicates that its success depends heavily upon the initial-random configuration. For this reason, we suggest that you try many different configurations. In one test, we found that the optimum resulted from only one in about fifteen starting configurations. Hence, we suggest that you repeat the process twenty-five or thirty times. The program lets you specify the number of repetitions.

Number of Clusters

A report is provided the gives the value of R-squared for each of the values of *K*. Select the value of *K* (number of clusters) that seems to maximize R-squared while minimizing *K*. Also, you should look at the plots of Y versus each X to help in determining the number of clusters. For example, the plot of the data on the previous page would suggest 2, 3, or 4 clusters.

Data Structure

The data are entered in the standard columnar format in which each column represents a single variable. One variable must be a dependent variable that will be regressed on the independent variables.

The data used in our tutorial, a portion of which is given in the following table, were generated with a large X pattern. They are plotted in the scatter plot that was shown above. The data are contained in the REGCLUS database.

REGCLUS dataset (subset)

Y	X
80.58823	15.4088
75.88235	20.12579
73.52941	21.69811
68.82353	23.27044
17.05882	2.830189
19.41177	4.402516
19.41177	5.974843
21.76471	9.119497

Missing Values

Rows with missing values are removed from the analysis.

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 a single, dependent variable. Remember that the dependent variable is predicted by the independent variables.

Independent Variables

X's: Independent Variables

Specify one or more independent variables. These are used to predict the dependent variable.

Include Intercept

Specifies whether you want to include the Y-intercept term in the regression model. Under most circumstances, you would.

Clustering Options

Number of Random Starts

This option specifies the number of different random configurations that should be run for each value of *K*. We suggest that about twenty-five repetitions be run since each initial configuration is completely random and the algorithm often converges to a non-optimal local optimum.

Because of execution time, you might want to set this value to three or four until you have found an appropriate value for *K* and then reset this value to twenty-five for a second run.

Maximum Iterations

This option sets the number of internal iterations before the algorithm is aborted. It is possible for a set of data to put the algorithm into an infinite loop. This option prevents this.

Minimum Rows Per Cluster

The third box lets you specify the minimum number of rows per cluster. Remember that in regression analysis, each cluster must contain at least one more row than there are independent variables.

Zero Exponent

This is the exponent of the value used as zero by the regression 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.

This box supplies the negative exponent. A value of 5 represents 1E-5 which is 0.00001.

Clustering Options – Number of Clusters

Minimum Clusters

The minimum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters. The actual number of clusters used is set above by the Number Clusters option.

Maximum Clusters

The maximum value of *K* to search. A separate cluster analysis is attempted for each value between the Minimum Clusters and the Maximum Clusters. The actual number of clusters used is set above by the Number Clusters option.

Reported Clusters

The is the number of clusters to be reported on. Although the program can find results for a range of cluster sizes, this option sets the size that is used in the detail and data storage sections.

Format Options

Label Variable

This is an optional variable containing identification for each row. These labels are used to enhance the interpretability of the reports.

Reports Tab

The following options control the formatting of the reports.

Select Reports

Iteration Detail Report - Cluster Report

Specify whether to display the indicated reports.

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.

Storage Tab

These options let you specify where to store the cluster number of each row on the current database.

Storage Variable

Store Cluster Id in Variable

You can automatically store the cluster identification number of each row into the variable specified here. The configuration stored is for the value of *K* specified by the Reported Clusters option.

Warning: Any data already in this variable are replaced by the cluster number. Be careful not to specify variables that contain important data.

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 – Regression Clustering

This section presents an example of how to run a cluster analysis of the data found in the REGCLUS database. This is a bivariate set of data generated to exhibit a large X pattern.

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

1 Open the RegClus dataset.

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

2 Open the Regression Clustering window.

- On the menus, select **Analysis**, then **Clustering**, then **Regression**. The Regression Clustering 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 Regression Clustering 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**. “Y” will appear in the Interval Variables box.
- Double-click in the **X’s: Independent Variables** box. This will bring up the variable selection window.
- Select **X** from the list of variables and then click **Ok**. “X” will appear in the X’s: Independent Variables box.
- Enter **5** for **Number of Random Starts**.
- Enter **4** for **Maximum Clusters**.
- Enter **2** for **Reported Clusters**.

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).

Iteration Detail Section

```
Iteration Detail Section
Number of      Replication     R-Squared   R-Squared
Clusters       Number              Value   Bar
2              1                0.997218   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2              2                0.997218   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2              3                0.997218   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2              4                0.997218   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2              5                0.961698   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

3              1                0.998170   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3              2                0.998170   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3              3                0.997952   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3              4                0.997952   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3              5                0.998343   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

4              1                0.999457   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
4              2                0.999015   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
4              3                0.999489   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
4              4                0.998505   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
4              5                0.998975   |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

This report displays the progress of the program through the various replications.

Number of Clusters

This column displays the number of clusters for the configurations presented on this row.

Replication Number

This column displays a sequence number for this replication.

R-Squared Value

This is the R-Squared that would result from fitting a separate regression of Y on X within each cluster. As this value approaches one, the fit of the regression is better.

R-Squared Bar

This is a bar chart of the R-Squared Value. This helps you visually determine the optimum value for the number of clusters.

Iteration Summary Section

```
Iteration Summary Section
Number of        R-Squared    R-Squared
Clusters             Value    Bar
2                 0.997218    |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3                 0.998343    |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
4                 0.999489    |IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

This section is the identical to the Iteration Detail Section except that only the row with the maximum value of R-Squared is displayed for each number of clusters. This report should help you determine the number of clusters by finding the first value of *K* where there is a large jump in the R-Squared value.

In this example, there is no jump. The value of *K* selected would be two.

Regression Coefficient Section

```
Regression Coefficient Section

Variable         Cluster 1    Cluster 2
Intercept        110.6421     18.26343
X                -1.866411    0.6797758
```

This report displays the coefficients of each regression equation for each cluster. For example, since we selected two clusters, there are two regression equations. These are

$$Y = 110.6421 - (1.866411)\ X$$

and

$$Y = 18.26343 + (0.6797758)\ X$$

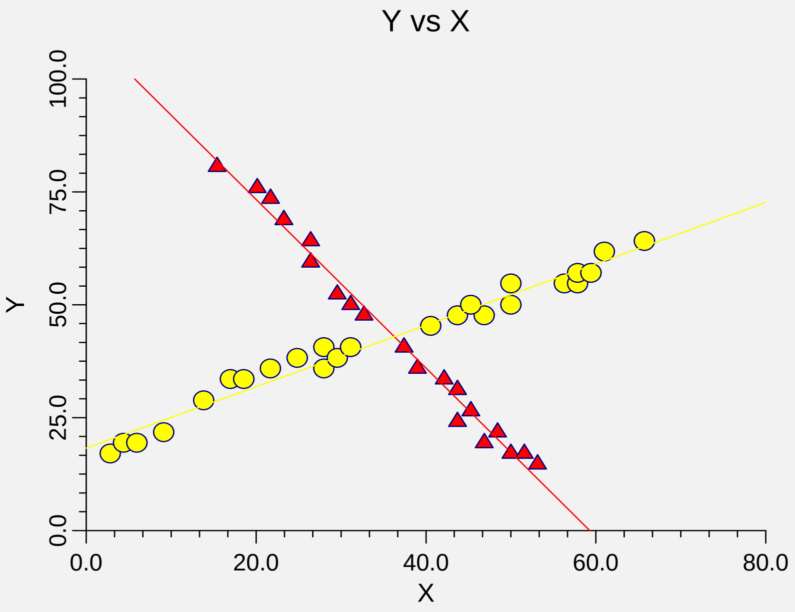
Cluster Section

```
Cluster Section
         Cluster
Row      Number            Y
1        1                 80.58823
2        1                 75.88235
3        1                 73.52941
4        1                 68.82353
5        2                 17.05882
6        2                 19.41177
7        2                 19.41177
8        2                 21.76471
.        .                 .
.        .                 .
.        .                 .
```

This report displays a report of which cluster each row is assigned to. The value of the dependent variable is also displayed to help you quickly identify a particular row. The cluster number may be stored directly on the database for further analysis and plotting.

Scatter Plot using Cluster Numbers

Once the cluster numbers are stored, you may use them as a grouping variable in the Scatter Plot program. This will provide a plot such as this:

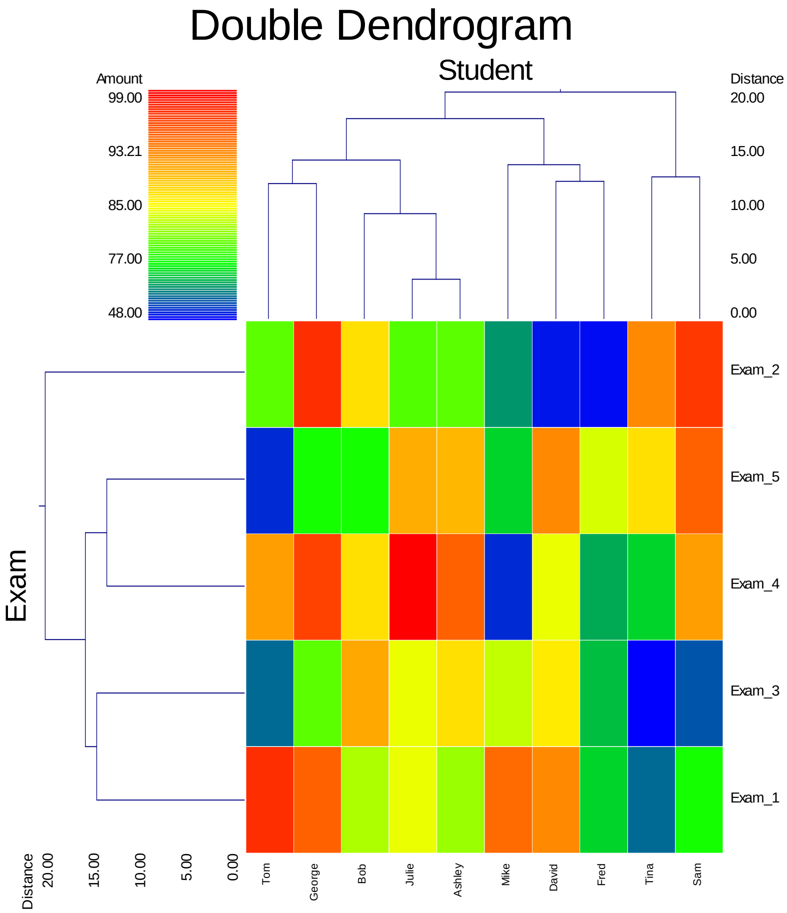


Chapter 450

Double Dendrograms

Introduction

This chapter describes how to obtain a double dendrograms using the ***NCSS***: Double Dendrograms procedure. Double dendrograms are dendrograms that cluster both the rows and the variables (columns) in a single graph. A set of eight hierarchical clustering algorithms are available including single linkage, complete linkage, and group average. The procedure outputs lists of the items in each cluster, linkage reports, and a double-dendrogram such as the following.



Hierarchical Cluster Algorithms

Chapter 445 of the ***NCSS*** manuals gives an introduction to hierarchical clustering. We suggest that you browse that chapter to get a basic overview of this technique. We will present here only highlights from that chapter.

We will first give brief comments about each of the eight hierarchical clustering techniques.

Single Linkage

Also known as *nearest neighbor* clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closest members. It often yields clusters in which individuals are added sequentially to a single group.

Complete Linkage

Also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

Simple Average

Also called the weighted pair-group method, this algorithm defines the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

Centroid

Also referred to as the unweighted pair-group centroid method, this method defines the distance between two groups as the distance between their centroids (center of gravity or vector average). The method should only be used with Euclidean distances.

Backward links may occur with this method. These are recognizable when the dendrogram no longer exhibits its simple tree-like structure in which each fusion results in a new cluster that is at a higher distance level (moves from right to left). With backward links, fusions can take place that result in clusters at a lower distance level (move from left to right). The dendrogram is difficult to interpret in this case.

Median

Also called the weighted pair-group centroid method, this defines the distance between two groups as the weighted distance between their centroids, the weight being proportional to the number of individuals in each group. Backward links (see discussion under Centroid) may occur with this method. The method should only be used with Euclidean distances.

Group Average

Also called the unweighted pair-group method, this is perhaps the most widely used of all the hierarchical cluster techniques. The distance between two groups is defined as the average distance between each of their members.

Ward's Minimum Variance

With this method, groups are formed so that the pooled within-group sum of squares is minimized. That is, at each step, the two clusters are fused which result in the least increase in the pooled within-group sum of squares.

Flexible Strategy

Lance and Williams (1967) suggested that a continuum could be made between single and complete linkage. The program lets you try various settings of these parameters which do not conform to the constraints suggested by Lance and Williams.

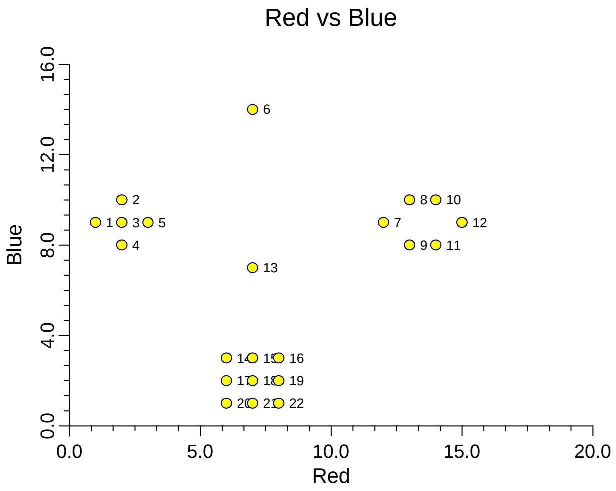
One interesting exercise is to vary these values, trying to find the set that maximizes the cophenetic correlation coefficient.

Dendrograms

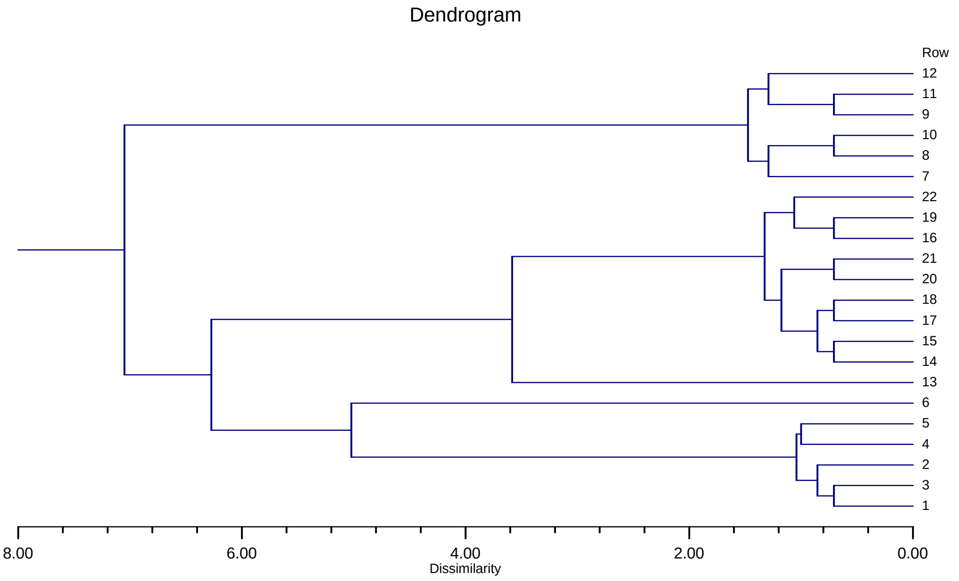
The *agglomerative hierarchical clustering* algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a *dendrogram*. The algorithm begins by placing each object in a separate cluster. Then, at each step, the two clusters that are most similar (according to a specific definition of similarity) are joined into a single new cluster. Once fused, objects are never separated. The eight clustering methods that are available represent eight methods of defining the similarity between clusters.

To help understand the dendrogram, consider the following example that has only two variables. Note that if we had only two variables, we perform the cluster analysis visually. The technique becomes useful once we have three or more variables to cluster.

Suppose we wish to cluster the bivariate data shown in the following scatter plot. In this case, the clustering may be done visually. The data seem to exhibit three clusters and two singletons, 6 and 13.



The following dendrogram was produced from the above data using popular the Group Average clustering algorithm.



The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The dendrogram is fairly simple to interpret. Remember that our main interest is in similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the dissimilarity between the two clusters.

Looking at this dendrogram, you can see the three clusters as three branches that occur at about the same horizontal distance. The two outliers, 6 and 13, are added in rather arbitrarily at much higher distances. This is the interpretation.

In this example we can compare our interpretation with an actual plot of the data. Unfortunately, this usually will not be possible because our data will consist of more than two variables.

Procedure Options

This section of the manual describes the function of each of the options on the panel windows.

Variables Tab

This panel specifies the variables and the clustering options.

Variables

These options specify the variables that will be used in the analysis.

Cluster Variables

Specify the variables (columns) that will be clustered. These variables should be interval-type variables, meaning the variable contain continuous measurements that may be either positive or negative and follow a linear scale. Examples include height, weight, age, price, temperature, and time.

A double dendrogram will be most useful when the cluster variables used are all on a similar scale.

Grouping Variable

This optional variable allows you to define a preliminary grouping pattern. By specifying a grouping variable here, you cause the clustering to take place within each group.

Row Label Variable

This optional variable contains labels that can be used for each row to aid in the interpretation. If this value is left blank, the row numbers will be used to identify the rows on the reports and dendrogram.

Linkage Options

Linkage Type (Clustering method)

This option specifies which of the eight possible hierarchical techniques is used. These methods were described earlier. The choices are

- **Single Linkage (Nearest Neighbor)**
- **Complete Linkage (Furthest Neighbor)**
- **Simple Average (Weighted Pair-Group)**
- **Group Average (Unweighted Pair-Group)**
- **Median (Weighted Pair-Group Centroid)**
  Requires the Distance Method to be Euclidean.
- **Centroid (Unweighted Pair-Group Centroid)**
  Requires the Distance Method to be Euclidean.

- **Ward's Minimum Variance**
  Requires the Distance Method to be Euclidean.

- **Flexible Strategy**
  Requires the Distance Method to be Euclidean.

Linkage Options – Flexible Strategy Parameters

Alpha

Specifies the values of α_i and α_j when the Flexible Strategy method is selected. You may enter a number or the letters "NI/NK." The "NI/NK" will cause this constant to be calculated and used as it is in the Centroid and Group Average methods.

Beta

Specifies the values of β when the Flexible Strategy method is selected. You may enter a number between -1 and 1 or the letters "NIJ/NK." The "NIJ/NK" will cause this constant to be calculated and used as it is in the Centroid method.

Gamma

Specifies the values of γ when the Flexible Strategy method is selected. You may enter any number.

Clustering Options

These options specify the cluster analysis technique.

Distance Method

This option specifies whether Euclidean or Manhattan distance is used. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Scaling Method

Specify the type of scaling to be used. Possible choices are Standard Deviation, Average Absolute Deviation, Range, and None. Since expression data is already scaled, we recommend selecting *None* here.

Four types of scaling are available: absolute value, standard deviation, range, and none. Each of these have the general form:

$$z_{ij} = \frac{x_{ij} - A_i}{B_i}$$

where x_{ij} represents the original data value for gene i and row j and z_{ij} represents the corresponding scale value. The scaling choice determines the values used for A_i and B_i.

The following table shows the scaling mechanism used for each type of scaling.

Type of Scaling	Value of A_i	Value of B_i
Absolute Value	$\frac{\sum_{j=1}^{N} x_{ij}}{N}$	$\frac{\sum_{j=1}^{N} \lvert x_{ij} - A_i \rvert}{N}$
Standard Deviation	$\frac{\sum_{j=1}^{N} x_{ij}}{N}$	$\sqrt{\frac{\sum_{j=1}^{N} \left(x_{ij} - A_i\right)^2}{N - 1}}$
Range	$\underset{over\ j}{Min}\left(x_{ij}\right)$	$\underset{over\ j}{Max}\left(x_{ij}\right) - \underset{over\ j}{Min}\left(x_{ij}\right)$
None	0	1

Reports Tab

The options on this panel control which reports and plots are generated.

Select Reports

These options designate which reports and plots to display.

Cluster Report - Dendrogram

Specify whether to display the indicated reports and plots.

Report Options

These options limit the cluster reports.

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.

Max Distance Items

This option specifies the maximum size of a distance matrix that will be displayed in the Distance Section report. Distance matrices with more items than this will not be displayed.

This option is here because for large datasets, the distance matrix may be very large.

Max Linkage Clusters

The Linkage Report can be long if the results for all links are printed. This parameter allows you to limit the number of links displayed so that only meaningful values are printed.

Reported Clusters

This option specifies the number of clusters used in the reports.

Dendrogram Tab

These options specify the double-dendrogram plot.

Heat Map

These options set the attributes of the heat map (the color scale).

Heat Map Colors

Click this box to bring up the Heat Map Window. The Heat Map window allows you to set the colors, the number of intervals (color gradations) in the heat map, and the type of scaling (regular, logarithmic, or percentile) that is to be used.

Heat Map - Legend

Label

This option defines the heading of the heat map legend. It sets the color, size, style, and value of the label text.

Show Legend

Check this box to display the heat map legend.

Number of Values

This option specifies the number of reference numbers that are to be displayed in the heat map legend.

Values Format

Click this button to bring up a window that sets the attributes of the legend reference values.

Miscellaneous

These are miscellaneous options for the double-dendrogram.

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.

Lines

This option sets the attributes of the dendrogram lines.

Show Trunk (Beginning Line)

Check this box to display the beginning line of the dendrograms. Since this line provides little information and sometimes adds to the clutter of the plot; it may be desirable to remove it.

Variables Dendrogram

These options set the attributes of the variables dendrogram—the dendrogram on the left.

Axis Label

This is the text of the label. Press the button on the right of the field to specify the font of the text.

Position

This option controls the placement of the axis label. You can place it either on the left or right edge of the plot.

Space

This option specifies the percentage of the horizontal space that is devoted to the variables dendrogram. A value near 30 is a good choice.

Label Settings…

This option specifies the characteristics of the variable names that are displayed on the right side of the plot. It displays a window that edits their font size and color. When you have fifty to one hundred genes, a font size of 4 or even 3 should be used.

Variables Dendrogram - Legend

Label

This option defines the heading of the dendrogram's legend. It sets the color, size, style, and value of the label text.

Show Legend

Check this box to display the variable dendrogram legend.

Number of Values

This option specifies the number of reference numbers that are to be displayed in the dendrogram's legend.

Value Format…

Click this button to bring up a window that sets the attributes of the legend reference values.

Rows Dendrogram

These options set the attributes of the rows dendrogram—the dendrogram at the top.

Axis Label

This is the text of the label. The character $\{X\}$ is replaced by the Row Label Variable's name. Press the button on the right of the field to specify the font of the text.

Position

This option controls the placement of the axis label. You can place it either on the bottom or top of the plot.

Space

This option specifies the percentage of the vertical space that is devoted to the rows dendrogram. A value near 30 is a good choice.

Label Settings…

This option specifies the characteristics of the row numbers or labels that are displayed at the bottom of the plot. It displays a window that edits their font size and color. When you have fifty to one hundred rows, a font size of 4 or even 3 should be used.

Rows Dendrogram - Legend

Label

This option defines the heading of the dendrogram's legend. It sets the color, size, style, and value of the label text.

Show Legend

Check this box to display the variable dendrogram legend.

Number of Values

This option specifies the number of reference numbers that are to be displayed in the dendrogram's legend.

Value Format…

Click this button to bring up a window that sets the attributes of the legend reference values.

Titles Tab

These options set the attributes of the attributes of the titles shown at the top and bottom of the plot.

Titles at the Top of the Plot

These options set the attributes of the attributes of the titles shown at the top of the plot.

Top Title 1

This sets the text and attributes of the top title line. 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.

Top Title 2

This sets the text and attributes of a second title line, usually have a smaller font size. 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.

Titles at the Bottom of the Plot

These options set the attributes of the attributes of the titles shown at the bottom of the plot.

Bottom Title 1

This sets the text and attributes of the first title line at the bottom of the plot. 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.

Bottom Title 2

This sets the text and attributes of a second title line at the bottom of the plot. 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.

Colors Tab

These options set the background colors of the dendrogram.

Group Background Colors (Used with Grouping Variable)

These options set the background colors of the rows dendrogram when a Grouping Variable (see Variables Tab above) has been specified.

Group 1 - Group 10

These options set the background colors of the region of the rows dendrogram devoted to rows within the corresponding group.

Background Colors of Various Plot Regions

Rows Dendrogram

This option sets the background color of the region of the plot that displays the rows dendrogram.

Variables Dendrogram

This option sets the background color of the region of the plot that displays the variables dendrogram.

Overall Background Color

This option sets the background color of the overall background of the dendrogram.

Storage Tab

The cluster id number for each row can be stored on the spreadsheet for further analysis. This option designates the column of the spreadsheet in which the cluster id's are stored.

Store the Cluster Id Number of Each Row in this Variable

The cluster id number for each row can be stored on the spreadsheet for further analysis. This option designates the column of the spreadsheet in which the cluster id's are stored.

WARNING: Existing data in this column will be replaced with the new values automatically when the procedure is run.

Cluster Id Variable

The cluster id number for each row is stored in this column. To omit the automatic storage of the cluster id's, leave this option blank.

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 – Creating a Double Dendrogram

This section presents an example of how to create a double dendrogram of the exam score data. The data are found in the EXAMS database.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Double Dendrograms window.

1 Open the EXAMS dataset.

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

2 Open the Double Dendrograms window.

- On the menus, select **Analysis**, then **Clustering**, then **Double Dendrograms**. The Double Dendrograms 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 Double Dendrograms window, select the **Variables tab**.
- Double-click in the **Cluster Variables** box. This will bring up the variable selection window.
- Select **Exam_1** through **Exam_5** from the list of variables and then click **Ok**. "Exam_1 - Exam_5" will appear in the Interval Variables box.
- Double-click in the **Label Variable** box. This will bring up the variable selection window.

- Select **Student** from the list of variables and then click **Ok**. "Student" will appear in the Label Variable box.
- Set **Scaling Method** to **None**.

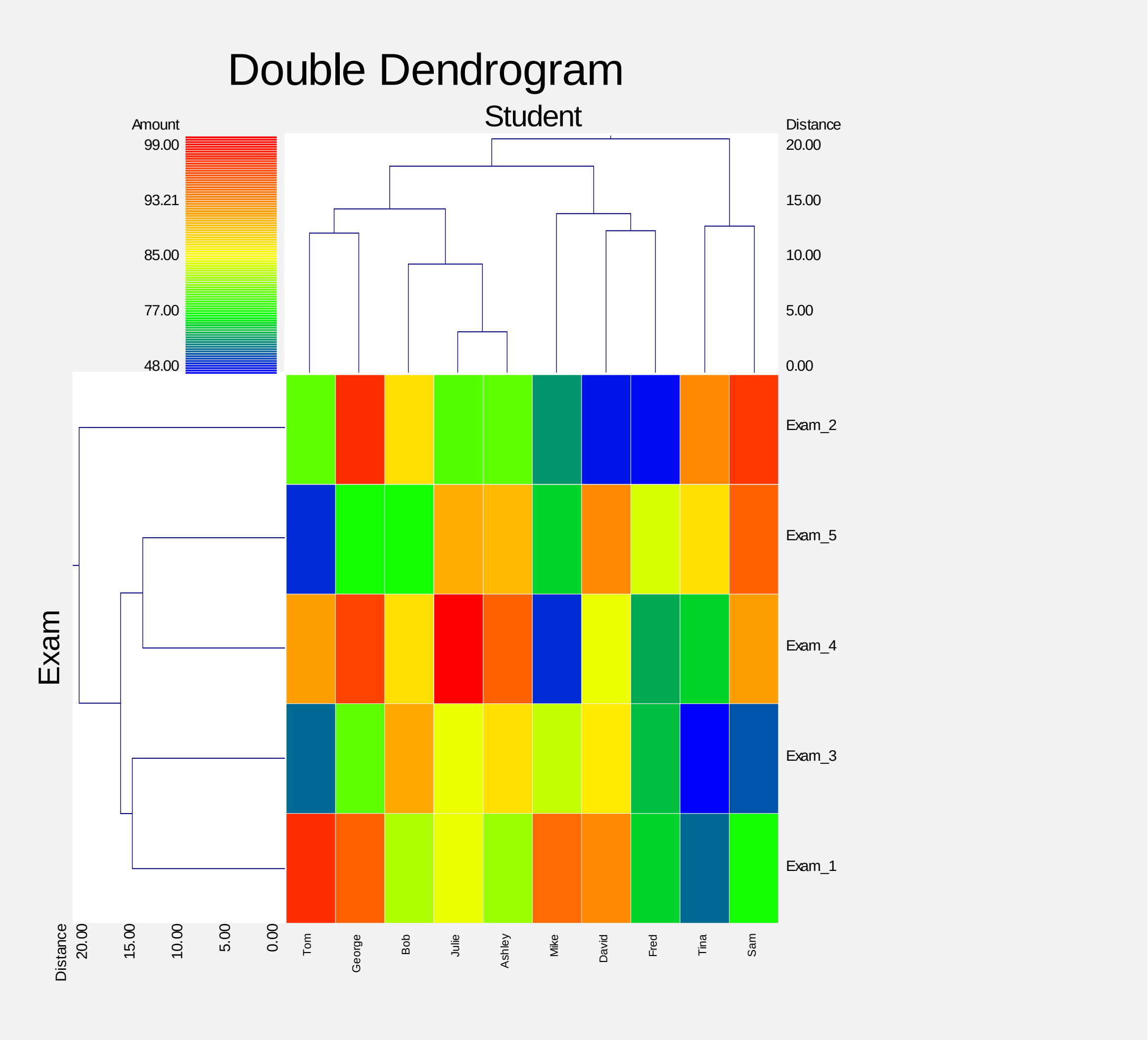
4 Specify the report.

- On the Double Dendrograms window, select the **Reports tab**.
- Check **Dendrogram**.
- Uncheck all other reports.

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).

Double Dendrogram Section



For examples of the other available reports, see Chapter 445, Hierarchical Clustering.

Chapter 455

Meta-Analysis of Means

Introduction

This module performs a meta-analysis on a set of two-group, continuous-scale, studies. These studies have a treatment group and a control group. Each study's result may be summarized by the sample size, mean, and standard deviation of each of the two groups. The program provides a complete set of numeric reports and plots to allow the investigation and presentation of the studies. The plots include the *forest plot*, *radial plot,* and *L'Abbe plot.* Both fixed-, and random-, effects models are available for analysis.

Meta-Analysis refers to methods for the systematic review of a set of individual studies with the aim to combine their results. Meta-analysis has become popular for a number of reasons:

1. The adoption of evidence-based medicine which requires that all reliable information is considered.
2. The desire to avoid narrative reviews which are often misleading.
3. The desire to interpret the large number of studies that may have been conducted about a specific treatment.
4. The desire to increase the statistical power of the results be combining many small-size studies.

The goals of meta-analysis may be summarized as follows. A meta-analysis seeks to systematically review all pertinent evidence, provide quantitative summaries, integrate results across studies, and provide an overall interpretation of these studies.

We have found many books and articles on meta-analysis. In this chapter, we briefly summarize the information in Sutton et al (2000) and Thompson (1998). Refer to those sources for more details about how to conduct a meta-analysis.

Treatment Effects

Suppose you have obtained the results for *k* studies, labeled $i = 1,\ldots,k$. Each study consists of a treatment group (T) and a control group (C). The results of each study are summarized by six statistics:

n_{Ti} the number of subjects in the treatment group.

n_{Ci} the number of subjects in the control group.

$\bar{x}_{Ti}$ the sample mean of the treatment group which estimates the treatment mean μ_{Ti}.

$\bar{x}_{Ci}$ the sample mean of the control group which estimates the control mean μ_{Ci}.

s_{Ti} the sample standard deviation of the treatment group.

s_{Ci} the sample standard deviation of the control group.

Mean Difference

The scales (e.g. blood pressure, pulse rate, volume, etc.) of all studies must be the same. If the logarithm has been taken in one study, it must be taken in all studies. You cannot combined studies with different scales using this program!

The measure of treatment effect for study *i* is

$$\theta_i = \mu_{Ti} - \mu_{Ci}$$

which is estimated by

$$\hat{\theta}_i = \bar{x}_{Ti} - \bar{x}_{Ci}$$

The standard deviation of the difference is

$$V\left(\hat{\theta}_i\right) = \sigma_i^2\left(\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}}\right)$$

The value of σ^2 is estimated by the pooled sample standard deviation given by

$$s_i^2 = \frac{\left(n_{Ti} - 1\right)s_{Ti}^2 + \left(n_{Ci} - 1\right)s_{Ci}^2}{n_{Ti} + n_{Ci}}$$

so that

$$\hat{V}\left(\hat{\theta}_i\right) = s_i^2\left(\frac{1}{n_{Ti}} + \frac{1}{n_{Ci}}\right)$$

Defining the Study Parameters

Confidence intervals based on the *t* distribution may be defined for θ_i in the usual manner.

$$\hat{\theta}_i \pm t_{n_{Ti}+n_{Ci}-2,1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}_i\right)}$$

It will be useful in the sequel to make the following definition of the weights.

$$v_i = \hat{V}\left(\hat{\theta}_i\right)$$

$$w_i = 1/v_i$$

Hypothesis Tests

Several hypothesis tests have be developed to test the various hypotheses that may be of interest. These will be defined next.

Overall Null Hypothesis

Two statistical tests have been devised to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0: \theta_i = 0 \quad i = 1, \cdots, k$$

Nondirectional Test

The nondirectional alternative hypothesis that at least one $\theta_i \neq 0$ may be tested by comparing the quantity

$$X_{ND} = \sum_{i=1}^{k} w_i \hat{\theta}_i^2$$

with a χ_k^2 distribution.

Directional Test

A test of the more interesting directional alternative hypothesis that $\theta_i = \theta \neq 0$ for all *i* may be tested by comparing the quantity

$$X_D = \frac{\left(\sum_{i=1}^{k} w_i \hat{\theta}_i \right)^2}{\sum_{i=1}^{k} w_i}$$

with a χ_1^2 distribution. Note that this tests the hypothesis that all effects are equal to the same nonzero quantity.

Effect-Equality (Heterogeneity) Test

When the overall null hypothesis is rejected, the next step is to test whether all effects are equal, that is, whether the effects are homogeneous. Specifically, the hypothesis is

$$H_0: \theta_i = \theta \quad i = 1, \cdots, k$$

versus the alternative that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction.

This hypothesis is tested using Cochran's Q test which is given by

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta} \right)^2$$

where

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

The test is conducted by comparing Q to a χ^2_{k-1} distribution.

Fixed versus Random Effects Combined Confidence Interval

If the effects are be assumed to be equal (homogeneous), either through testing or from other considerations, a *fixed effects model* may be used to construct a combined confidence interval. However, if the effects are heterogeneous, a *random effects model* should be used to construct the combined confidence interval.

Fixed Effects Model

The fixed effects model assumes homogeneity of study results. That is, it assumes that $\theta_i = \theta$ for all *i*. This assumption may not be realistic when combining studies with different patient pools, protocols, follow-up strategies, doses, durations, etc.

If the fixed effects model is adopted, the *inverse variance-weighted* method as described by Sutton (2000) page 58 is used to calculate the confidence interval for θ. The formulas used are

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

$$\hat{V}\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{k} w_i}$$

Random Effects Model

The random effects model assumes that the individual θ_i come from a random distribution with fixed mean $\bar{\theta}$ and variance σ^2. Sutton (2000) page 74 presents the formulas necessary to conduct a random effects analysis using the *weighted* method. The formulas used are

$$\hat{\hat{\theta}} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\hat{\theta}}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\bar{\theta}} = \frac{\sum_{i=1}^{k} \overline{w}_i \hat{\theta}_i}{\sum_{i=1}^{k} \overline{w}_i}$$

$$\hat{V}\left(\hat{\bar{\theta}}\right) = \frac{1}{\sum_{i=1}^{k} \overline{w}_i}$$

$$\overline{w}_i = \frac{1}{\frac{1}{w_i} + \hat{\tau}^2}$$

$$\hat{\tau}^2 = \begin{cases} \frac{Q-k+1}{U} & \text{if } Q > k-1 \\ 0 & \text{otherwise} \end{cases}$$

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2$$

$$U = (k-1)\left(\overline{w} - \frac{s_w^2}{k\overline{w}}\right)$$

$$s_w^2 = \frac{1}{k-1}\left(\sum_{i=1}^{k} w_i^2 - k\overline{w}^2\right)$$

$$\overline{w} = \frac{1}{k}\left(\sum_{i=1}^{k} w_i\right)$$

Graphical Displays

A number of plots have been devised to display the information in a meta-analysis. These include the forest plot, the radial plot, and the L'Abbe plot. More will be said about each of these plots in the Output section.

Data Structure

The data are entered into a dataset using one row per study. Six variables are required to hold the sample size, mean, and standard deviation of each study. In addition to these, an additional variable is usually used to hold a short (3 or 4 character) label. Another variable may be used to hold a grouping variable.

As an example, we will use data referred to in Sutton (2000) page 30 as the dental dataset. This dataset reviews nine randomized clinical trials that were conducted to study the effects of sodium fluoride (NaF) with sodium monofluorophosphate (SMFP). These nine studies were all on the same continuous scale, so their results could be analyzed using the meta-analysis techniques presented in this chapter. These data are contained in the SUTTON30 database. You should load this database to see how the data are arranged.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

The options on this screen control the variables that are used in the analysis.

Variables

Treatment N Variable

Specify the variable containing the sample size for the treatment group. Each row of data represents a separate study.

Treatment Mean Variable

Specify the variable containing the mean of the treatment group. Each row of data represents a separate study.

Treatment S.D. Variable

Specify the variable containing the standard deviation (not the standard error) of the treatment group. Each row of data represents a separate study.

Control N Variable

Specify the variable containing the sample size for the control group. Each row of data represents a separate study.

Control Mean Variable

Specify the variable containing the mean of the control group. Each row of data represents a separate study.

Control S.D. Variable

Specify the variable containing the standard deviation (not the standard error) of the control group. Each row of data represents a separate study.

Variables – Optional Variables

Label Variable

Specify an optional variable containing a label for each study (row) in the database. This label should be short (< 8 letters) so that it can fit on the plots.

Group Variable

Specify an optional variable containing a group identification value. Each unique value of this variable will receive its own plotting symbol on the forest plots. Some reports are sorted by these group values.

Combine Studies Method

Combine Studies Using

Specify the method used to combine treatment effects.

Use the **Fixed Effects** method when you do not want to account for the variation between studies.

Use the **Random Effects** method when you want to account for the variation between studies as well as the variation within the studies.

Reports Tab

The options on this screen control the appearance of the reports.

Select Reports

Summary Report - Outcome Detail Reports

Indicate whether to display the corresponding report.

Select Plots

Forest Plot – L’Abbe Plot

Indicate whether to display the corresponding plot.

Report Options

Alpha Level

This setting controls the confidence coefficient used in the confidence limits. Note that 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05, which results in 95% confidence intervals.

Show Notes

Indicate whether to show the notes at the end of reports. Although these notes are helpful at first, they may tend to clutter the output. This option lets you omit them.

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.

Report Options – Decimal Places

Probability Values - T Values

This setting controls the number of digits to the right of the decimal place that are displayed when showing this item.

Plot Options

Plot Options – Legend Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. Click the button on the right to specify the font size, color, and style of the legend text. The characters {G} are replaced with the name of the Group Variable.

Plot Options – Plot Symbol Options

Symbols Proportional to Sample Size

Check this box to cause the size of the plotting symbols on forest plots and L'Abbe plots to be proportional to relative study size. The larger the sample size, the larger the symbol. The range of the size of the symbol is controlled by the Size Min Pcnt and Size Max Pcnt options below.

Size Min Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the smallest sample size. The recommended value is 50. Typical values range from 20 to 99.

The formula for a symbol's size is

$$\text{Actual Symbol Size} = (\text{Normal Symbol Size})*\text{Radius}$$

where

$$\text{Radius} = [(\text{Min Pct}) + (\text{Max Pct} - \text{Min Pct})*(\text{Sample Size})/(\text{Max Sample Size})]/100$$

Size Max Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the largest sample size. The recommended value is 150. Typical values range from 101 to 200.

Forest Plot Tab

The options on this panel control the appearance of the forest plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forest 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 Forest style file is used. These style files are created in the Scatter Plot procedure.

Ref. Line

This option lets you indicate whether to display the reference line and the characteristics of that line.

Difference Ref. Value

This is the position of the reference line on the forest plot.

Titles

Plot Title

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

Radial Plot Tab

The options on this panel control the appearance of the radial plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Radial 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. These style files are created in the Scatter Plot procedure.

Symbol

Specify a symbol. Usually, no symbol is used.

Symbol Font Size

This option lets you specify the size of font used to display the row numbers or row labels.

Titles

Plot Title

This is the text of the title. The characters *{G}* are replaced by the output measure. Press the button on the right of the field to specify the font of the text.

L’Abbe Plot Tab

The options on this panel control the appearance of the L’Abbe plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

L'Abbe 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. These style files are created in the Scatter Plot procedure.

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 Tab

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fourteen groups may be used. When no Group Variable is specified, the options made for Symbol 1 are used to define the plot symbol. Following is an example of possible symbol settings for two groups.

Plotting Symbols

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol.

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

Data Storage Options – Select Items to Store on the Spreadsheet

Mean Difference - Weights

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 – Meta-Analysis of Means

This section presents an example of how to analyze the data contained in the SUTTON30 database. This dataset contains data for nine randomized clinical trials that were conducted to study the effect of fluoridation. The NaF data represent the control group and the SMFP represent the treated group.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Meta-Analysis of Means window.

1 Open the SUTTON30 dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sutton30.s0**.
- Click **Open**.

2 Open the Meta-Analysis of Means window.

- On the menus, select **Analysis**, then **Meta-Analysis**, then **Meta-Analysis of Means**. The Meta-Analysis of Means procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Variables tab**.
- Set the **Treatment N Variables** to **SMFPN**.
- Set the **Treatment Mean Variables** to **SMFPMean**.
- Set the **Treatment S.D. Variable**s to **SMFPSD**.
- Set the **Control N Variables** to **NaFN**.
- Set the **Control Mean Variables** to **NaFMean**.
- Set the **Control S.D. Variables** to **NaFSD**.
- Set the **Label Variable** to **Study**.

4 Specify the reports.

- Select the **Reports tab**.
- Check the **Summary Report** option box.
- Check the **Heterogeneity Tests** option box.
- Check the **Outcome Detail Reports** option box.
- Check the **Forest Plot (By Measure)** option box.
- Check **Radial Plot** option box.
- Check the **L'Abbe Plot** option box.

5 Specify the radial plot.

- Select the **Radial Plot tab**.
- Set the **Vertical Axis Minimum** to **-4**.
- Set the **Vertical Axis Maximum** to **4**.
- Set the **Horizontal Axis Minimum** to **0**.
- Set the **Horizontal Axis Maximum** to **6**.

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).

Run Summary Section

```
Parameter                   Value          Parameter                Value
N Treatment Variable        SMFPN          N Control Variable       NaFN
Mean Treatment Variable     SMFPMean       Mean Control Variable    NaFMean
SD Treatment Variable       SMFPSD         SD Control Variable      NaFSD
Group Variable              None           Number Groups            1
Row Label Variable          Study          Rows Processed           9
```

This report records the variables that were used and the number of rows that were processed.

Numeric Summary Section

```
Study    NT/NC         Mean T     Mean C   Difference        SD T        SD C
S1       113/134       6.8200     5.9600       0.8600      4.7200      4.2400
S2       151/175       5.0700     4.7400       0.3300      5.3800      4.6400
S3       140/137       2.5100     2.0400       0.4700      3.2200      2.5900
S4       179/184       3.2000     2.7000       0.5000      2.4600      2.3200
S5       169/174       5.8100     6.0900      -0.2800      5.1400      4.8600
S6       736/754       4.7600     4.7200       0.0400      5.2900      5.3300
S7       209/209      10.9000    10.1000       0.8000      7.9000      8.1000
S8       1122/1151     3.0100     2.8200       0.1900      3.3200      3.0500
S9       673/679       4.3700     3.8800       0.4900      5.3700      4.8500

[Combined]
Average                                        0.2835

Note: This report shows the input data for each study in the analysis. The 'Average' values are
actually weighted averages with weights based on the effects model that was selected.
```

This report summarizes the input data. You should scan it for any mistakes. Note that the ‘Average’ line provides the estimated group average.

NT/NC

These are the count values that were read from the database.

Mean T

These are the input treatment means.

Mean C

These are the input control means.

Difference

These are the computed values of Mean T minus Mean C. These difference values are the effects of interest in the analysis.

SD T

These are the input treatment standard deviations.

SD C

These are the input control standard deviations.

Nondirectional Zero-Effect Test

```
                Outcome                                 Prob
Rows            Measure          Chi-Square     DF      Level
Combined        Mean Difference     14.8417      9     0.0954
```

This reports the results of the nondirectional zero-effect chi-square test designed to test the null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0: \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative hypothesis is that at least one $\theta_i \neq 0$, that is, at least one study had a statistically significant result.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies.

Prob Level

This is the significance level of the test. If this value is less than the nominal value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Directional Zero-Effect Test

```
                Outcome                                 Prob
Rows            Measure          Chi-Square     DF      Level
Combined        Mean Difference      9.4395      1     0.0021
```

This reports the results of the directional zero-effect chi-square test designed to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0: \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative hypothesis is that $\theta_i = \theta \neq 0$ for all *i*, that is, that all effects are equal to the same, non-zero value.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal one.

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Effect-Equality (Heterogeneity) Test

Treatment	Outcome Measure	Cochran's Q	DF	Prob Level
Combined	Mean Difference	5.4022	8	0.7139

This reports the results of the effect-equality (homogeneity) test. This chi-square test was designed to test the null hypothesis that all treatment effects are equal. The null hypothesis is written

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

The alternative is that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction. This test may help you determine whether to use a Fixed Effects model (used for homogeneous effects) or a Random Effects model (heterogeneous effects).

Cochran's Q

This is the computed chi-square value for Cochran's Q statistic. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies minus one..

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Mean Difference Detail Section

Study	Mean Difference	Standard Error	95.0% Lower Confidence Limit	95.0% Upper Confidence Limit	Percent Random Effects Weight
S1	0.8600	0.5704	-0.2579	1.9779	2.6172
S2	0.3300	0.5549	-0.7577	1.4177	2.7648
S3	0.4700	0.3516	-0.2191	1.1591	6.8887
S4	0.5000	0.2509	0.0082	0.9918	13.5238
S5	-0.2800	0.5400	-1.3384	0.7784	2.9199
S6	0.0400	0.2752	-0.4993	0.5793	11.2455
S7	0.8000	0.7826	-0.7340	2.3340	1.3900
S8	0.1900	0.1337	-0.0720	0.4520	47.6529
S9	0.4900	0.2782	-0.0554	1.0354	10.9974
[Combined]					
Average	0.2835	0.0923	0.1026	0.4643	

This report displays results for the mean difference outcome measure.

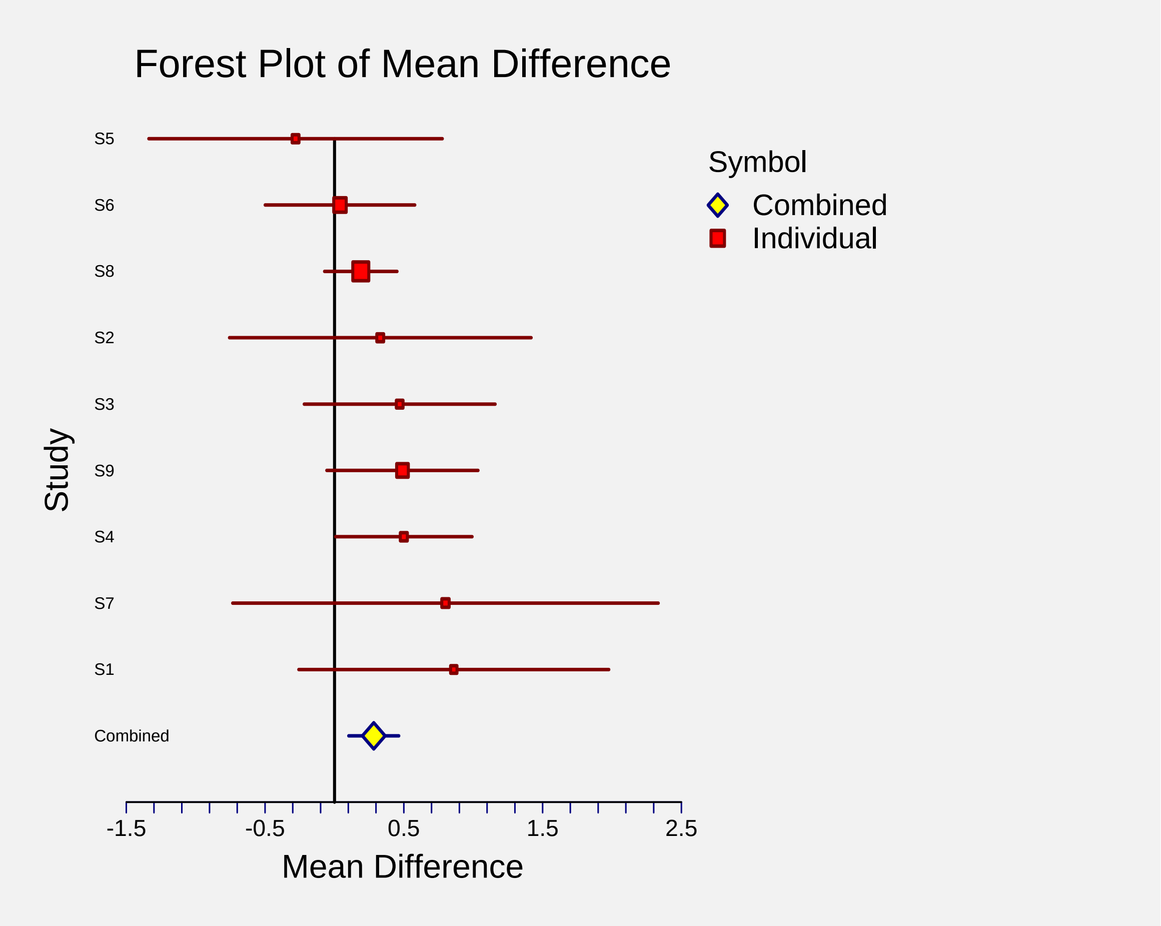
Confidence Limits

These are the lower and upper confidence limits (the formulas were given earlier in this chapter).

Weights

The last column gives the relative (percent) weight used in creating the weighted average. Using these values, you can decide how much influence each study has on the weighted average.

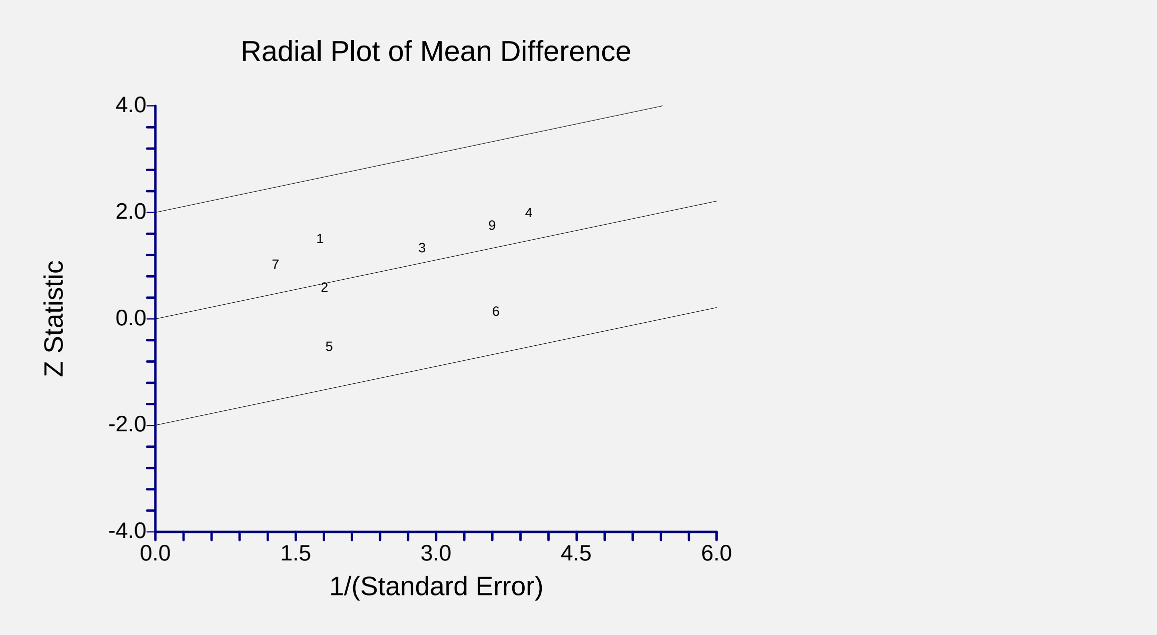
Forest Plot



This plot presents the results for each study on one plot. The size of the plot symbol is proportional to the sample size of the study. The points on the plot are sorted by the mean difference. The lines represent the confidence intervals about the mean differences. Note that the narrower the confidence limits, the better.

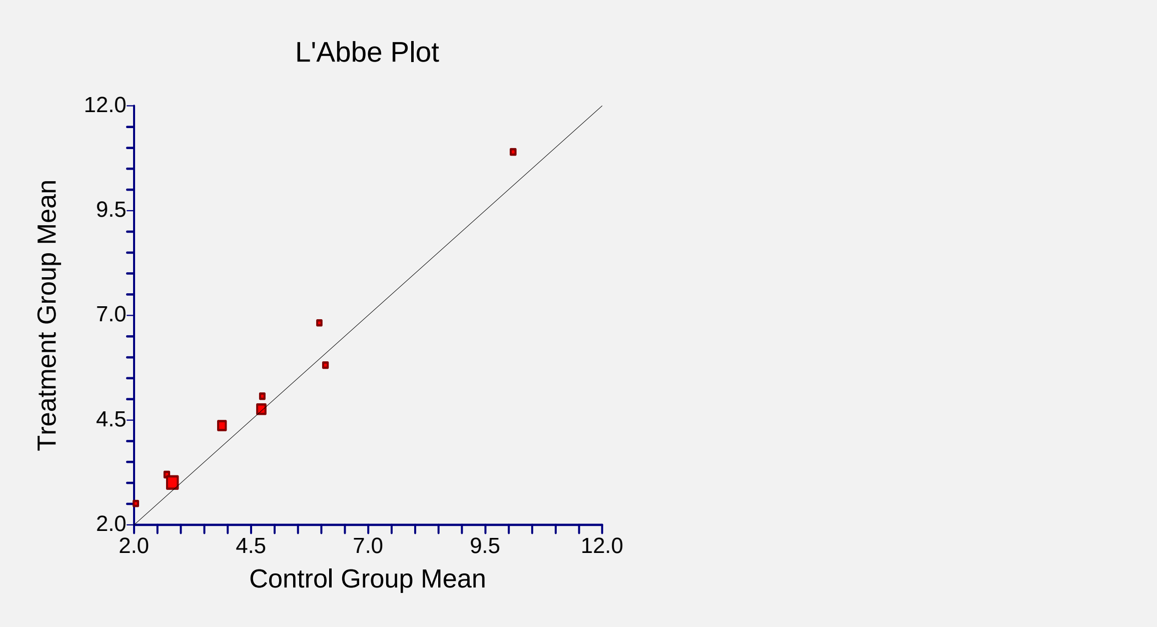
By studying this plot, you can determine the main conclusions that can be drawn from the set of studies. For example, you can determine how many studies were significant (the confidence limits do not intersect the vertical line at 0.0).

Radial Plot



The radial (or Galbraith) plot shows the z-statistic (outcome divided by standard error) on the vertical axis and a measure of weight on the horizontal axis. Studies that have the largest weight are closest to the Y axis. Studies within the limits are interpreted as homogeneous. Studies outside the limits may be outliers.

L’Abbe Plot



The L’Abbe plot displays the treatment mean on vertical axis versus the control mean on the horizontal axis. Homogenous studies will be arranged along the diagonal line. This plot is especially useful in determining if the relationship between the treatment group and the control group is the same for all values of the control group risk.

Chapter 456

Meta-Analysis of Proportions

Introduction

This module performs a meta-analysis of a set of two-group, binary-event studies. These studies have a treatment group (arm) and a control group. The results of each study may be summarized as counts in a 2-by-2 table. The program provides a complete set of numeric reports and plots to allow the investigation and presentation of the studies. The plots include the *forest plot, radial plot,* and *L'Abbe plot.* Both fixed-, and random-, effects models are available for analysis.

Meta-Analysis refers to methods for the systematic review of a set of individual studies with the aim to combine their results. Meta-analysis has become popular for a number of reasons:

1. The adoption of evidence-based medicine which requires that all reliable information is considered.
2. The desire to avoid narrative reviews which are often misleading.
3. The desire to interpret the large number of studies that may have been conducted about a specific treatment.
4. The desire to increase the statistical power of the results be combining many small-size studies.

The goals of meta-analysis may be summarized as follows. A meta-analysis seeks to systematically review all pertinent evidence, provide quantitative summaries, integrate results across studies, and provide an overall interpretation of these studies.

We have found many books and articles on meta-analysis. In this chapter, we briefly summarize the information in Sutton et al (2000) and Thompson (1998). Refer to those sources for more details about how to conduct a meta-analysis.

Treatment Effects

Suppose you have obtained the results for *k* studies, labeled $i = 1,\ldots,k$. Each study consists of a treatment group (T) and a control group (C). The results of each study are summarized by four counts:

a_i the number of subjects in the treatment group having the event of interest.

b_i the number of subjects in the control group having the event of interest.

c_i the number of subjects in the treatment group not having the event of interest.

d_i the number of subjects in the control group not having the event of interest.

Occasionally, one of these counts will be zero which causes calculation problems. To avoid this, the common procedure is to add a small value of 0.5 or 0.25 to all counts so that zero counts do not occur.

Risks

These counts may be used to calculate estimates of the event-risk in the treatment group as

$$\hat{p}_{Ti} = \frac{a_i}{a_i + c_i}$$

and in the control group as

$$\hat{p}_{Ci} = \frac{b_i}{b_i + d_i}$$

Based on these risks, three measures of treatment effect may be defined and used in the meta-analysis. These are the odds ratio, the risk ratio, and the risk difference.

Odds Ratio

The odds ratio is the most commonly used measure of treatment effect. It is defined as follows.

$$OR_i = \frac{\dfrac{p_{Ti}}{1 - p_{Ti}}}{\dfrac{p_{Ci}}{1 - p_{Ci}}}$$

For statistical analysis, the logarithm of the odds ratio is usually used because its distribution is more accurately approximated by the normal distribution for smaller sample sizes. The variance of the sample log odds ratio is estimated by

$$\hat{V}\left(\ln\left(OR_i\right)\right) = \frac{1_i}{a_i} + \frac{1_i}{b_i} + \frac{1_i}{c_i} + \frac{1_i}{d_i}$$

Risk Ratio or Relative Risk

The risk ratio is calculated as follows.

$$RR_i = \frac{p_{Ti}}{p_{Ci}}$$

Like the odds ratio, the logarithm of the risk ratio is usually used because its distribution is more accurately approximated by the normal distribution for smaller sample sizes. The variance of the sample log risk ratio is estimated by

$$\hat{V}\left(\ln\left(RR_i\right)\right) = \frac{1}{a_i} - \frac{1}{a_i + c_i} + \frac{1}{b_i} - \frac{1}{b_i + d_i}$$

Risk Difference

The risk difference is calculated as follows.

$$RD_i = p_{Ti} - p_{Ci}$$

The estimated variance of the sample risk difference is given by

$$\hat{V}\left(RD_i\right) = \frac{p_{Ti}\left(1 - p_{Ti}\right)}{a_i + c_i} + \frac{p_{Ci}\left(1 - p_{Ci}\right)}{b_i + d_i}$$

Defining the Study Parameters

Let θ_i represent the outcome measure created from the 2-by-2 table. That is, θ_i may be the odds ratio, risk ratio, or risk difference. Let $\hat{\theta}_i$ represent the estimate of θ_i from the study. Confidence intervals based on the normal distribution may be defined for θ_i in the usual manner.

$$\hat{\theta}_i \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}_i\right)}$$

In the case of the odds ratio and the risk ratio, the interval is created on the logarithmic scale and then transformed back to the original scale.

It will be useful in the sequel to make the following definition of the weights.

$$v_i = \hat{V}\left(\hat{\theta}_i\right)$$

$$w_i = 1/v_i$$

Hypothesis Tests

Several hypothesis tests have be developed to test the various hypotheses that may be of interest. These will be defined next.

Overall Null Hypothesis

Two statistical tests have been devised to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

Nondirectional Test

The nondirectional alternative hypothesis that at least one $\theta_i \neq 0$ may be tested by comparing the quantity

$$X_{ND} = \sum_{i=1}^{k} w_i \hat{\theta}_i^2$$

with a χ_k^2 distribution.

Directional Test

A test of the more interesting directional alternative hypothesis that $\theta_i = \theta \neq 0$ for all *i* may be tested by comparing the quantity

$$X_D = \frac{\left(\sum_{i=1}^{k} w_i \hat{\theta}_i\right)^2}{\sum_{i=1}^{k} w_i}$$

with a χ_1^2 distribution. Note that this tests the hypothesis that all effects are equal to the same nonzero quantity.

Effect-Equality (Heterogeneity) Test

When the overall null hypothesis is rejected, the next step is to test whether all effects are equal, that is, whether the effects are homogeneous. Specifically, the hypothesis is

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

versus the alternative that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction.

This hypothesis is tested using Cochran's Q test which is given by

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2$$

where

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

The test is conducted by comparing *Q* to a χ_{k-1}^2 distribution.

Fixed versus Random Effects Combined Confidence Interval

If the effects are assumed to be equal (homogeneous), either through testing or from other considerations, a *fixed effects model* may be used to construct a combined confidence interval. However, if the effects are heterogeneous, a *random effects model* should be used to construct the combined confidence interval.

Fixed Effects Model

The fixed effects model assumes homogeneity of study results. That is, it assumes that $\theta_i = \theta$ for all *i*. This assumption may not be realistic when combining studies with different patient pools, protocols, follow-up strategies, doses, durations, etc.

If the fixed effects model is adopted, the *inverse variance-weighted* method as described by Sutton (2000) page 58 is used to calculate the confidence interval for θ. The formulas used are

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

$$\hat{V}\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{k} w_i}$$

Random Effects Model

The random effects model assumes that the individual θ_i come from a random distribution with fixed mean $\bar{\theta}$ and variance σ^2. Sutton (2000) page 74 presents the formulas necessary to conduct a random effects analysis using the *weighted* method. The formulas used are

$$\hat{\hat{\theta}} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\hat{\theta}}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\hat{\theta}} = \frac{\sum_{i=1}^{k} \overline{w}_i \hat{\theta}_i}{\sum_{i=1}^{k} \overline{w}_i}$$

$$\hat{V}\left(\hat{\bar{\theta}}\right)=\frac{1}{\sum_{i=1}^{k}\overline{w}_i}$$

$$\overline{w}_i=\frac{1}{\frac{1}{w_i}+\hat{\tau}^2}$$

$$\hat{\tau}^2=\begin{cases}\frac{Q-k+1}{U} & \text{if } Q>k-1\\ 0 & \text{otherwise}\end{cases}$$

$$Q=\sum_{i=1}^{k}w_i\left(\hat{\theta}_i-\hat{\theta}\right)^2$$

$$U=(k-1)\left(\overline{w}-\frac{s_w^2}{k\overline{w}}\right)$$

$$s_w^2=\frac{1}{k-1}\left(\sum_{i=1}^{k}w_i^2-k\overline{w}^2\right)$$

$$\overline{w}=\frac{1}{k}\left(\sum_{i=1}^{k}w_i\right)$$

Graphical Displays

A number of plots have been devised to display the information in a meta-analysis. These include the forest plot, the radial plot, and the L'Abbe plot. More will be said about each of these plots in the Output section.

Data Structure

The data are entered into a dataset using one row per study. The four counts of the study's 2-by-2 table are entered into four columns. In addition to these, an additional variable is usually used to hold a short (3 or 4 character) label. Another variable may be needed to hold a grouping variable.

As an example, we will use data referred to in Sutton (2000) as the cholesterol-lowering intervention dataset. This data set reviews 34 randomized clinical trials that were conducted to study the effects of three cholesterol-lowering treatments: diet, drug, and surgery. The mortality of patients in a treatment arm and a control arm were recorded. These data are contained in the SUTTON 22 database. You should load this database to see how the data are arranged.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

The options on this screen control the variables that are used in the analysis.

Variables

Treatment Event (A) Variable

Specify the variable containing the count of the number of subjects in the treatment group in which the event of interest occurred. Each study may be represented by a 2-by-2 table of counts. This variable contains the treatment-event counts.

Control Event (B) Variable

Specify the variable containing the count of the number of subjects in the control group in which the event of interest occurred. Each study may be represented by a 2-by-2 table of counts. This variable contains the control-event counts.

Treatment Nonevent (C) Variable

Specify the variable containing the count of the number of subjects in the treatment group in which the event of interest did not occur. Each study may be represented by a 2-by-2 table of counts. This variable contains the treatment-nonevent counts.

Control Nonevent (D) Variable

Specify the variable containing the count of the number of subjects in the control group in which the event of interest did not occur. Each study may be represented by a 2-by-2 table of counts. This variable contains the control-nonevent counts.

Variables – Optional Variables

Label Variable

Specify an optional variable containing a label for each study (row) in the database. This label should be short (< 8 letters) so that it can fit on the plots.

Group Variable

Specify an optional variable containing a group identification value. Each unique value of this variable will receive its own plotting symbol on the forest plots. Some reports are sorted by these group values.

Combine Studies Method

Combine Studies Using

Specify the method used to combine treatment effects.

Use the Fixed Effects method when you do not want to account for the variation between studies.

Use the Random Effects method when you want to account for the variation between studies as well as the variation within the studies.

Zero Counts

Change Zero Counts To (Delta)

This is the value added to each cell to avoid having zero cell counts. Outcome measures like the odds ratio and risk ratio are not defined when certain counts are zero. By adding a small amount to each cell count, this option lets you analyze data with zero counts. You might consider running your analysis a couple of times with two or three difference delta values to determine if the delta value is making a big difference in the outcome (it should not).

If all cells in all rows are non-zero, enter 0. Otherwise, use 0.5 or 0.25. (Recent simulation studies have shown that 0.25 produces better results in some situations than the more traditional 0.5.)

Reports Tab

The options on this screen control the appearance of the reports.

Select Reports

Show Odds Ratio, Risk Ratio, or Risk Difference Reports/Plots

Indicate whether to display reports and plots about this outcome measure. You must check at least one of the three outcome measures.

Summary Report - Outcome Detail Reports

Indicate whether to display the corresponding report.

Select Plots

Forest Plot – L'Abbe Plot

Indicate whether to display the corresponding plot.

Report Options

Alpha Level

This setting controls the confidence coefficient used in the confidence limits. Note that 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05, which results in 95% confidence intervals.

Show Notes

Indicate whether to show the notes at the end of reports. Although these notes are helpful at first, they may tend to clutter the output. This option lets you omit them.

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.

Report Options – Decimal Places

Probability Values – Ratio Values

This setting controls the number of digits to the right of the decimal place that are displayed when showing this item.

Plot Options

Plot Options – Legend Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. Click the button on the right to specify the font size, color, and style of the legend text. The characters {G} are replaced with the name of the Group Variable.

Plot Options – Plot Symbol Options

Symbols Proportional to Sample Size

Check this box to cause the size of the plotting symbols on forest plots and L'Abbe plots to be proportional to relative study size. The larger the sample size, the larger the symbol. The range of the size of the symbol is controlled by the Size Min Pcnt and Size Max Pcnt options below.

Size Min Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the smallest sample size. The recommended value is 50. Typical values range from 20 to 99.

The formula for a symbol's size is

$$\text{Actual Symbol Size} = (\text{Normal Symbol Size})*\text{Radius}$$

where

$$\text{Radius} = [(\text{Min Pct}) + (\text{Max Pct - Min Pct})*(\text{Sample Size})/(\text{Max Sample Size})]/100$$

Size Max Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the largest sample size. The recommended value is 150. Typical values range from 101 to 200.

Forest Plot Tab

The options on this panel control the appearance of the forest plot.

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.

Log Scale

This option controls the scaling of horizontal axis. We suggest that you use a logarithmic scale for the odds ratio and risk ratio. The risk difference forest plot will automatically revert to a regular scale since the logarithm of negative numbers is not defined.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forest 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 Forest style file is used. These style files are created in the Scatter Plot procedure.

Line

This option lets you indicate whether to display the reference line and the characteristics of that line.

Ratio Value

This is the position of the reference line on the odds ratio and risk ratio forest plots.

Difference Value

This is the position of the reference line on the risk difference forest plots.

Titles

Plot Title

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

Radial Plot Tab

The options on this panel control the appearance of the radial plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Radial 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. These style files are created in the Scatter Plot procedure.

Symbol

Specify a symbol. Usually, no symbol is used.

Symbol Font Size

This option lets you specify the size of font used to display the row numbers or row labels.

Titles

Plot Title

This is the text of the title. The characters *{G}* are replaced by the output measure. Press the button on the right of the field to specify the font of the text.

L'Abbe Plot Tab

The options on this panel control the appearance of the L'Abbe plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

L'Abbe 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. These style files are created in the Scatter Plot procedure.

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 Tab

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fourteen groups may be used. When no Group Variable is specified, the options made for Symbol 1 are used to define the plot symbol. Following is an example of possible symbol settings for two groups:

Plotting Symbols

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol.

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 on the Spreadsheet

P1 – Risk Diff. Weights

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 – Meta-Analysis of Proportions

This section presents an example of how to analyze the data contained in the SUTTON 22 database. This dataset contains data for 34 randomized clinical trials that were conducted to study the effects of three cholesterol-lowering treatments: diet, drug, and surgery. The mortality of patients in a treatment arm and a control arm were recorded.

You may follow along here by making the appropriate entries or load the completed template **Example 1** from the Template tab of the Meta-Analysis of Proportions window.

1 **Open the SUTTON 22 dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SUTTON 22.s0**.
- Click **Open**.

2 **Open the Meta-Analysis of Proportions window.**

- On the menus, select **Analysis**, then **Meta-Analysis**, then **Meta-Analysis of Proportions**. The Meta-Analysis of Proportions procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Variables tab**.
- Set the **Treatment Event (A) Variable** to **TDeath**.
- Set the **Treatment Nonevent (C)** Variable to **TSurvive**.
- Set the **Control Event (B) Variable** to **CDeath**.
- Set the **Control Nonevent (D) Variable** to **CSurvive**.
- Set the **Label Variable** to **StudyId**.
- Set the **Group Variable** to **Treatment**.

4 Specify the reports.

- Select the **Reports tab**.
- Check the **Show Odds Ratio Reports/Plots** option box.
- Check the **Summary Report** option box.
- Check the **Heterogeneity Tests** option box.
- Check the **Outcome Detail Reports** option box.
- Check the **Forest Plot (By Group & Measure)** option box.
- Check **Radial Plot** option box.
- Check the **L'Abbe Plot** option box.

5 Specify the plotting symbols.

- Select the **Symbols tab**.
- Set the **Group 2 Symbol Type** to **Solid Circle**.
- Set the **Group 3 Symbol Type** to **Solid Hexagon**.

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).

Run Summary Section

```
Parameter                                 Value          Parameter         Value
Treatment Event-Count Variable            TDeath         Rows Processed    34
Treatment Nonevent-Count Variable         TSurvive       Number Groups     3
Control Event-Count Variable              CDeath         Delta Value       0.5
Control Nonevent-Count Variable           CSurvive
Row Label Variable                        StudyId
Group Variable                            Treatment
```

This report records the variables that were used and the number of rows that were processed.

Numeric Summary Section

```
[Treatment]                                                Odds      Risk       Risk
StudyId        Data                    P1         P2      Ratio     Ratio  Difference
[Diet]
S1             28/204  51/202        0.1373     0.2525    0.4750    0.5480    -0.1147
S7             41/206  55/206        0.1990     0.2670    0.6845    0.7477    -0.0676
S8             20/123  24/129        0.1626     0.1860    0.8529    0.8772    -0.0231
S9             111/1018  113/1015    0.1090     0.1113    0.9770    0.9795    -0.0023
S16            174/424  178/422      0.4104     0.4218    0.9542    0.9730    -0.0114
S17            28/199  31/194        0.1407     0.1598    0.8626    0.8821    -0.0190
S21            39/221  28/237        0.1765     0.1181    1.5910    1.4859     0.0582
S22            8/54  1/26            0.1481     0.0385    3.1075    2.7818     0.0990
S24            269/4541  248/4516    0.0592     0.0549    1.0835    1.0785     0.0043
Average (random effects model)                            0.9292    0.9440    -0.0082

[Drug]
S2             70/285  38/147        0.2456     0.2585    0.9305    0.9476    -0.0136
S3             37/156  40/119        0.2372     0.3361    0.6160    0.7077    -0.0986
S4             2/88  3/30            0.0227     0.1000    0.2271    0.2488    -0.0848
S5             0/30  3/33            0.0000     0.0909    0.1429    0.1567    -0.0868
S6             61/279  82/276        0.2186     0.2971    0.6636    0.7375    -0.0782
S10            81/427  27/143        0.1897     0.1888    0.9964    0.9971    -0.0006
S11            31/244  51/253        0.1270     0.2016    0.5801    0.6341    -0.0742
S12            17/50  12/50          0.3400     0.2400    1.6090    1.4000     0.0980
S13            23/47  20/48          0.4894     0.4167    1.3335    1.1702     0.0712
S15            1025/5552  723/2789   0.1846     0.2592    0.6470    0.7122    -0.0746
S18            42/350  48/367        0.1200     0.1308    0.9075    0.9187    -0.0107
S19            4/79  5/78            0.0506     0.0641    0.7965    0.8080    -0.0134
S20            37/1149  48/1129      0.0322     0.0425    0.7517    0.7597    -0.0103
S23            5/71  7/72            0.0704     0.0972    0.7223    0.7435    -0.0264
S26            0/94  1/94            0.0000     0.0106    0.3298    0.3333    -0.0105
S27            19/311  12/317        0.0611     0.0379    1.6293    1.5900     0.0232
S28            68/1906  71/1900      0.0357     0.0374    0.9534    0.9550    -0.0017
S29            44/2051  43/2030      0.0215     0.0212    1.0128    1.0125     0.0003
S30            33/6582  3/1663       0.0050     0.0018    2.4267    2.4194     0.0030
S31            236/5331  181/5296    0.0443     0.0342    1.3081    1.2945     0.0101
S32            0/48  1/49            0.0000     0.0204    0.3333    0.3401    -0.0198
S33            1/94  0/52            0.0106     0.0000    1.6845    1.6737     0.0064
S34            1/23  2/29            0.0435     0.0690    0.7333    0.7500    -0.0208
Average                                                   0.8863    0.9108    -0.0115

[Surgery]
S14            0/30  4/60            0.0000     0.0667    0.2058    0.2186    -0.0576
S25            46/421  62/417        0.1093     0.1487    0.7044    0.7369    -0.0393
Average                                                   0.6885    0.7238    -0.0439

[Combined]
Average                                                   0.8868    0.9100    -0.0112
```

This report summarizes the input data. You should scan it for any mistakes. Note that the ‘Average’ lines provide the estimated group averages. The values depend on your selection of whether the Random Effects model or Fixed Effects model was used. The ‘Combined’ line provides the combined results of all studies.

Data

These are the count values that were read from the database.

P1

This is the estimated event proportion in the treatment group. This is also known as the treatment-group risk.

P2

This is the estimated event proportion in the control group. This is also known as the treatment-group risk.

Odds Ratio

This is the estimated value of the odds ratio. Note that it depends not only on the data, but also on the delta value used.

Risk Ratio

This is the estimated value of the risk ratio. Note that it depends not only on the data, but also on the delta value used.

Risk Difference

This is the estimated value of the risk difference. Note that it depends not only on the data, but also on the delta value used.

Nondirectional Zero-Effect Test

```
              Outcome                                 Prob
Treatment     Measure        Chi-Square       DF      Level
Diet          Odds Ratio        16.9314        9     0.0498
Drug          Odds Ratio        95.6162       23     0.0000
Surgery       Odds Ratio         3.9568        2     0.1383
Combined      Odds Ratio       116.5043       34     0.0000
```

This reports the results of the nondirectional zero-effect chi-square test designed to test the null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative hypothesis is that at least one $\theta_i \neq 0$, that is, at least one study had a statistically significant result.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies.

Prob Level

This is the significance level of the test. If this value is less than the nominal value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Directional Zero-Effect Test

Treatment	Outcome Measure	Chi-Square	DF	Prob Level
Diet	Odds Ratio	0.1815	1	0.6701
Drug	Odds Ratio	33.7356	1	0.0000
Surgery	Odds Ratio	3.3032	1	0.0691
Combined	Odds Ratio	27.8056	1	0.0000

This reports the results of the directional zero-effect chi-square test designed to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative hypothesis is that $\theta_i = \theta \neq 0$ for all *i*, that is, that all effects are equal to the same, non-zero value.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal one.

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Effect-Equality (Heterogeneity) Test

Treatment	Outcome Measure	Cochran's Q	DF	Prob Level
Diet	Odds Ratio	16.7499	8	0.0328
Drug	Odds Ratio	61.8806	22	0.0000
Surgery	Odds Ratio	0.6536	1	0.4188
Combined	Odds Ratio	88.6987	33	0.0000

This reports the results of the effect-equality (homogeneity) test. This chi-square test was designed to test the null hypothesis that all treatment effects are equal. The null hypothesis is written

$$H_0 : \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative is that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction. This test may help you determine whether to use a Fixed Effects model (used for homogeneous effects) or a Random Effects model (heterogeneous effects).

Cochran's Q

This is the computed chi-square value for Cochran's Q statistic. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies minus one..

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Odds Ratio Detail Section

```
                                                   95.0%      95.0%    Percent
                                                   Lower      Upper    Random
[Treatment]                               Odds  Confidence Confidence  Effects
StudyId              P1         P2       Ratio      Limit      Limit   Weight
[Diet]
S1               0.1373     0.2525      0.4750     0.2863     0.7882   3.5636
S7               0.1990     0.2670      0.6845     0.4327     1.0828   3.9108
S8               0.1626     0.1860      0.8529     0.4469     1.6277   2.7201
S9               0.1090     0.1113      0.9770     0.7405     1.2889   5.4463
S16              0.4104     0.4218      0.9542     0.7261     1.2538   5.4819
S17              0.1407     0.1598      0.8626     0.4976     1.4952   3.2731
S21              0.1765     0.1181      1.5910     0.9450     2.6788   3.4641
S22              0.1481     0.0385      3.1075     0.5128    18.8317   0.5279
S24              0.0592     0.0549      1.0835     0.9073     1.2940   6.2826
Average                                 0.9292     0.7641     1.1300

[Drug]
S2               0.2456     0.2585      0.9305     0.5902     1.4668   3.9372
S3               0.2372     0.3361      0.6160     0.3637     1.0434   3.4236
.                     .          .           .          .          .   .
.                     .          .           .          .          .   .
.                     .          .           .          .          .   .
S32              0.0000     0.0204      0.3333     0.0132     8.3867   0.1737
S33              0.0106     0.0000      1.6845     0.0674    42.0926   0.1744
S34              0.0435     0.0690      0.7333     0.0898     5.9856   0.3966
Average                                 0.8863     0.7345     1.0696

[Surgery]
S14              0.0000     0.0667      0.2058     0.0107     3.9513   0.2060
S25              0.1093     0.1487      0.7044     0.4692     1.0575   4.3237
Average                                 0.6885     0.4603     1.0297

[Combined]
Average                                 0.8868     0.7739     1.0161
```

This report displays results for the odds ratio outcome measure. You can obtain a similar report for the risk ratio and the risk difference. The report gives you the

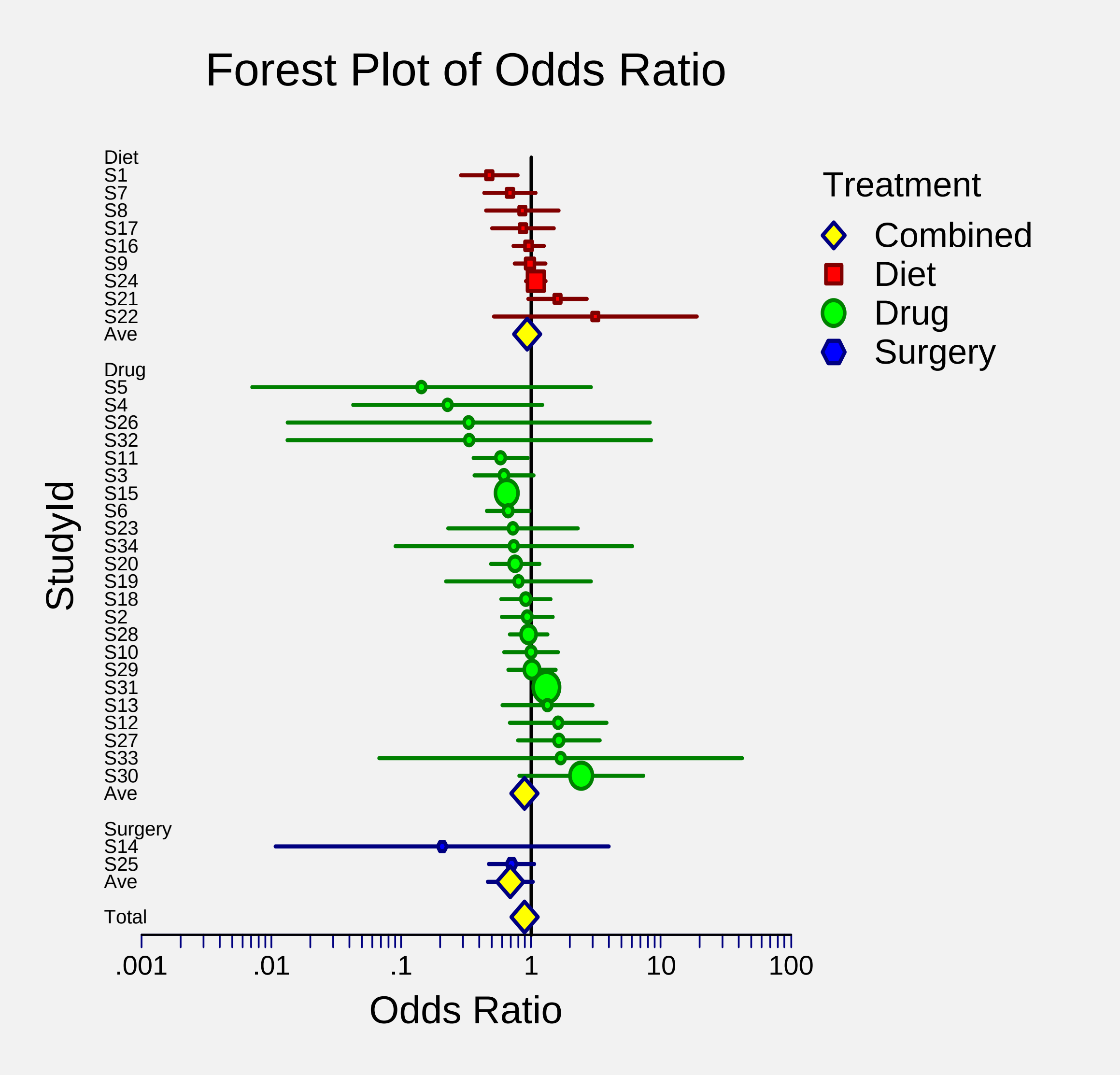
Confidence Limits

These are the lower and upper confidence limits (the formulas were given earlier in this chapter).

Weights

The last column gives the relative (percent) weight used in creating the weighted average. Using these values, you can decide how much influence each study has on the weighted average.

Forest Plot

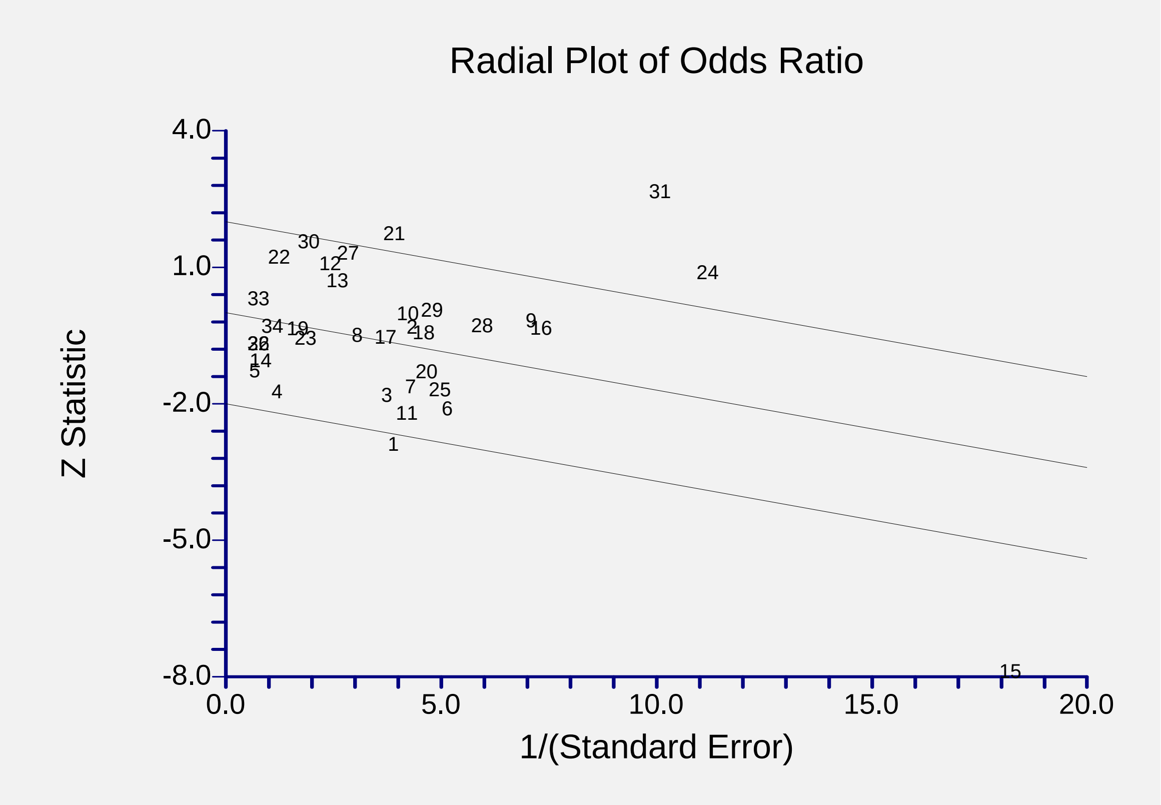


This plot presents the results for each study on one plot. The size of the plot symbol is proportional to the sample size of the study. The points on the plot are sorted by group and by the odds ratio. The lines represent the confidence intervals about the odds ratios. Note that the narrower the confidence limits, the better.

By studying this plot, you can determine the main conclusions that can be drawn from the set of studies. For example, you can determine how many studies were significant (the confidence limits do not intersect the vertical line at 1.0). You can see if there were different conclusions for the different groups.

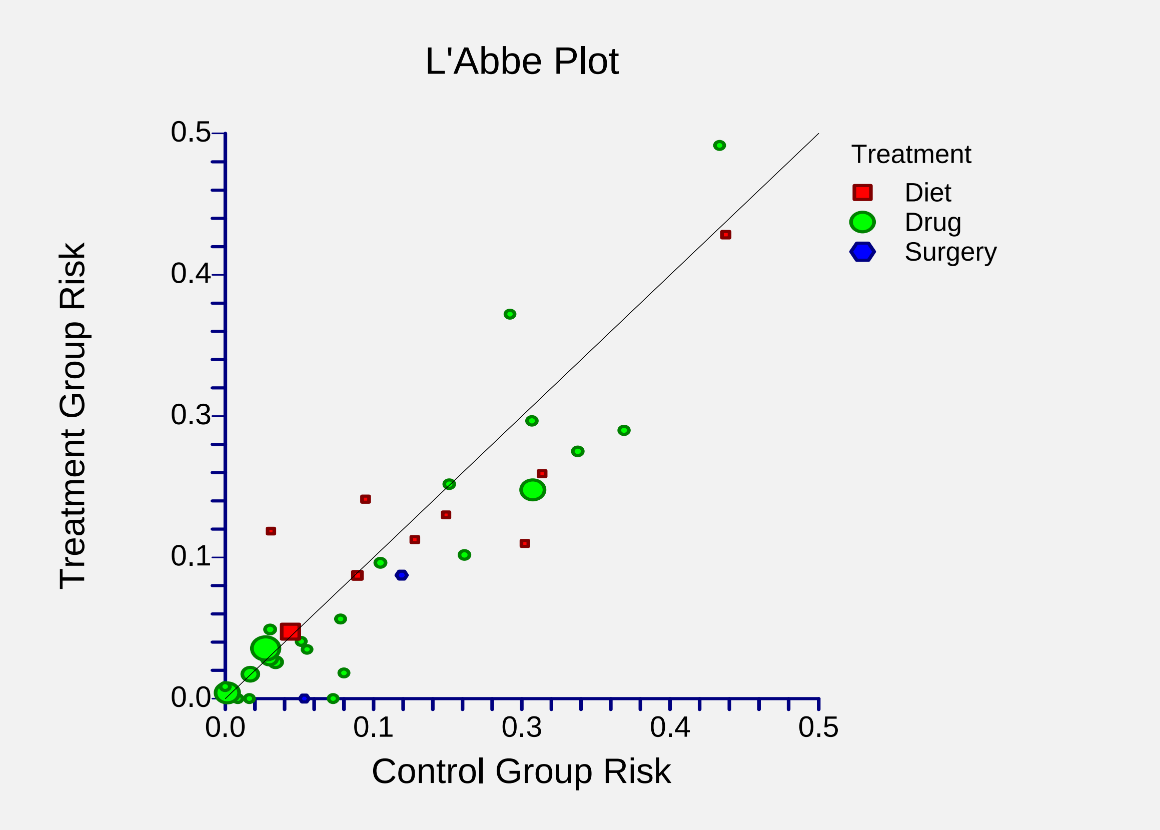
The results of the combining the studies are displayed at the end of each group.

Radial Plot



The radial (or Galbraith) plot shows the z-statistic (outcome divided by standard error) on the vertical axis and a measure of weight on the horizontal axis. Studies that have the largest weight are closest to the Y axis. Studies within the limits are interpreted as homogeneous. Studies outside the limits may be outliers.

L’Abbe Plot



The L’Abbe plot displays the treatment risk on vertical axis versus the control risk on the horizontal axis. Homogenous studies will be arranged along the diagonal line. This plot is especially useful in determining if the relationship between the treatment group and the control group is the same for all values of the control group risk.

Chapter 457

Meta-Analysis of Correlated Proportions

Introduction

This module performs a meta-analysis of a set of correlated, binary-event studies. These studies usually come from a design in which two dichotomous responses are made on each subject (or subject pair). The results of each study can be summarized as counts in a 2-by-2 table. For example, the binary response is recorded after treatment A and again after treatment B. The response is '1' if the event of interest occurs or '0' otherwise. This analysis also applies to *matched pairs* data in which each *case* subject is matched with a similar subject from a *control* group.

The program provides a complete set of numeric reports and plots to allow the investigation and presentation of the studies. The plots include the *forest plot*, *radial plot*, and *L'Abbe plot*. Both fixed-, and random-, effects models are available for analysis.

Meta-Analysis refers to methods for the systematic review of a set of individual studies with the aim to combine their results. Meta-analysis has become popular for a number of reasons:

1. The adoption of evidence-based medicine which requires that all reliable information is considered.
2. The desire to avoid narrative reviews which are often misleading.
3. The desire to interpret the large number of studies that may have been conducted about a specific treatment.
4. The desire to increase the statistical power of the results be combining many small-size studies.

The goals of meta-analysis may be summarized as follows. A meta-analysis seeks to systematically review all pertinent evidence, provide quantitative summaries, integrate results across studies, and provide an overall interpretation of these studies.

We have found many books and articles on meta-analysis. In this chapter, we briefly summarize the information in Sutton et al (2000) and Thompson (1998). Refer to those sources for more details about how to conduct a meta-analysis.

Treatment Effects

Suppose you have obtained the results for *k* studies, labeled *i = 1,…,k*. Each study consists of two dichotomous measurements Y_1 and Y_2 on each of *n* subjects (the 'subject' may be a pair of matched individuals). Measurement Y_1 represents the treatment response and Y_2 represents the control response. The results of each study are summarized by four counts:

a_i the number of $Y_1 = 1$ and $Y_2 = 1$.

b_i the number of $Y_1 = 1$ and $Y_2 = 0$.

c_i the number of $Y_1 = 0$ and $Y_2 = 1$.

d_i the number of $Y_1 = 0$ and $Y_2 = 0$.

Occasionally, one of these counts will be zero which causes calculation problems. To avoid this, the common procedure is to add a small value of 0.5 or 0.25 to all counts so that zero counts do not occur.

Odds Ratio

When a paired design is used, Sahai and Khurshid (1995) indicate that the odds ratio is estimated using the following simple formula of McNemar which is based on the Mantel-Haenszel estimator.

$$OR_i = \frac{b_i}{c_i}$$

For statistical analysis, the logarithm of the odds ratio is usually used because its distribution is more accurately approximated by the normal distribution for smaller sample sizes. Sahai and Khurshid (1995) page 119 give the variance of the sample log odds ratio is estimated by

$$\hat{V}\left(\ln\left(OR_i\right)\right) = \frac{1_i}{b_i} + \frac{1_i}{c_i}$$

Risk Ratio or Relative Risk

Following Sahai and Khurshid (1995) page 139, the risk ratio is estimated as follows.

$$RR_i = \frac{a_i + b_i}{a_i + c_i}$$

Like the odds ratio, the logarithm of the risk ratio is used because its distribution is more accurately approximated by the normal distribution for smaller sample sizes. The variance of the sample log risk ratio is estimated by

$$\hat{V}\left(\ln\left(RR_i\right)\right) = \frac{\left(b_i + c_i\right)}{\left(a_i + c_i\right)\left(a_i + b_i\right)}$$

Risk Difference

Following Sahai and Khurshid (1995) page 139, the risk difference is calculated as follows.

$$RD_i = \frac{b_i - c_i}{n_i}$$

The estimated variance of the sample risk difference is given by

$$\hat{V}(RD_i) = \frac{n_i(b_i + c_i) - (b_i - c_i)^2}{n_i^3}$$

Defining the Study Parameters

Let θ_i represent the outcome measure created from the 2-by-2 table. That is, θ_i may be the odds ratio, risk ratio, or risk difference. Let $\hat{\theta}_i$ represent the estimate of θ_i from the study. Confidence intervals based on the normal distribution may be defined for θ_i in the usual manner.

$$\hat{\theta}_i \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\theta}_i)}$$

In the case of the odds ratio and the risk ratio, the interval is created on the logarithmic scale and then transformed back to the original scale.

It will be useful in the sequel to make the following definition of the weights.

$$v_i = \hat{V}(\hat{\theta}_i)$$

$$w_i = 1 / v_i$$

Hypothesis Tests

Several hypothesis tests have be developed to test the various hypotheses that may be of interest. These will be defined next.

Overall Null Hypothesis

Two statistical tests have been devised to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0: \theta_i = 0 \quad i = 1, \cdots, k$$

Nondirectional Test

The nondirectional alternative hypothesis that at least one $\theta_i \neq 0$ may be tested by comparing the quantity

$$X_{ND} = \sum_{i=1}^{k} w_i \hat{\theta}_i^2$$

with a χ_k^2 distribution.

Directional Test

A test of the more interesting directional alternative hypothesis that $\theta_i = \theta \neq 0$ for all *i* may be tested by comparing the quantity

$$X_D = \frac{\left(\sum_{i=1}^{k} w_i \hat{\theta}_i\right)^2}{\sum_{i=1}^{k} w_i}$$

with a χ_1^2 distribution. Note that this tests the hypothesis that all effects are equal to the same nonzero quantity.

Effect-Equality (Heterogeneity) Test

When the overall null hypothesis is rejected, the next step is to test whether all effects are equal, that is, whether the effects are homogeneous. Specifically, the hypothesis is

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

versus the alternative that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction.

This hypothesis is tested using Cochran's Q test which is given by

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2$$

where

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

The test is conducted by comparing Q to a χ_{k-1}^2 distribution.

Fixed, Versus Random, Effects Combined Confidence Interval

If the effects are be assumed to be equal (homogeneous), either through testing or from other considerations, a *fixed effects model* may be used to construct a combined confidence interval. However, if the effects are heterogeneous, a *random effects model* should be used to construct the combined confidence interval.

Fixed Effects Model

The fixed effects model assumes homogeneity of study results. That is, it assumes that $\theta_i = \theta$ for all *i*. This assumption may not be realistic when combining studies with different patient pools, protocols, follow-up strategies, doses, durations, etc.

If the fixed effects model is adopted, the *inverse variance-weighted* method as described by Sutton (2000) page 58 is used to calculate the confidence interval for θ. The formulas used are

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\theta}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

$$\hat{V}\left(\hat{\theta}\right) = \frac{1}{\sum_{i=1}^{k} w_i}$$

Random Effects Model

The random effects model assumes that the individual θ_i come from a random distribution with fixed mean $\bar{\theta}$ and variance σ^2. Sutton (2000) page 74 presents the formulas necessary to conduct a random effects analysis using the *weighted* method. The formulas used are

$$\hat{\bar{\theta}} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\bar{\theta}}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\bar{\theta}} = \frac{\sum_{i=1}^{k} \overline{w}_i \hat{\theta}_i}{\sum_{i=1}^{k} \overline{w}_i}$$

$$\hat{V}\left(\hat{\bar{\theta}}\right)=\frac{1}{\sum_{i=1}^{k}\overline{w}_i}$$

$$\overline{w}_i=\frac{1}{\frac{1}{w_i}+\hat{\tau}^2}$$

$$\hat{\tau}^2=\begin{cases}\frac{Q-k+1}{U} & \text{if } Q>k-1\\ 0 & \text{otherwise}\end{cases}$$

$$Q=\sum_{i=1}^{k}w_i\left(\hat{\theta}_i-\hat{\theta}\right)^2$$

$$U=(k-1)\left(\overline{w}-\frac{s_w^2}{k\overline{w}}\right)$$

$$s_w^2=\frac{1}{k-1}\left(\sum_{i=1}^{k}w_i^2-k\overline{w}^2\right)$$

$$\overline{w}=\frac{1}{k}\left(\sum_{i=1}^{k}w_i\right)$$

Graphical Displays

A number of plots have been devised to display the information in a meta-analysis. These include the forest plot, the radial plot, and the L'Abbe plot. More will be said about each of these plots in the Output section.

Data Structure

The data are entered into a dataset using one row per study. The four counts of the study's 2-by-2 table are entered into four columns. In addition to these, an additional variable is usually used to hold a short (3 or 4 character) label. Another variable may be needed to hold a grouping variable.

As an example, we will use the METACPROP dataset which presents the results of 24 matched case-control studies that were conducted to study the effectiveness of a certain treatment. The goal of each study was to compare the proportion of cases that responding with a 'Yes' to the corresponding proportion of control responses with a 'Yes'. The studies were grouped into two diets, but these were not their main focus. These data are contained in the METACPROP database. You should load this database to see how the data are arranged.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

The options on this screen control the variables that are used in the analysis.

Data and Variables

N11 Count (A) Variable

Specify the variable containing the count of the number of subjects that responded with a ‘1’ (Yes) to both variables. In a matched case control study, this variable contains the number of case-control pairs that both showed the event of interest.

N10 Count (B) Variable

Specify the variable containing the count of the number of subjects that responded with a ‘1’ (Yes) to first variable and a ‘0’ (No) to the second. In a matched case control study, this variable contains the number of case-control pairs that had a positive case and a negative control.

N01 Count (C) Variable

Specify the variable containing the count of the number of subjects that responded with a ‘0’ (No) to first variable and a ‘1’ (Yes) to the second. In a matched case control study, this variable contains the number of case-control pairs that had a negative case and a positive control.

N00 Count (D) Variable

Specify the variable containing the count of the number of subjects that responded with a ‘0’ (No) to both variables. In a matched case control study, this variable contains the number of case-control pairs that were negative for both the case and the control.

Data and Variables – Optional Variables

Label Variable

Specify an optional variable containing a label for each study (row) in the database. This label should be short (< 8 letters) so that it can fit on the plots.

Group Variable

Specify an optional variable containing a group identification value. Each unique value of this variable will receive its own plotting symbol on the forest plots. Some reports are sorted by these group values.

Combine Studies Method

Combine Studies Using

Specify the method used to combine treatment effects.

Use the Fixed Effects method when you do not want to account for the variation between studies.

Use the Random Effects method when you want to account for the variation between studies as well as the variation within the studies.

Zero Counts

Change Zero Counts To (Delta)

This is the value added to each cell to avoid having zero cell counts. Outcome measures like the odds ratio and risk ratio are not defined when certain counts are zero. By adding a small amount to each cell count, this option lets you analyze data with zero counts. You might consider running your analysis a couple of times with two or three difference delta values to determine if the delta value is making a big difference in the outcome (it should not).

If all cells in all rows are non-zero, enter 0. Otherwise, use 0.5 or 0.25. (Recent simulation studies have shown that 0.25 produces better results in some situations than the more traditional 0.5.)

Reports Tab

The options on this screen control the appearance of the reports.

Select Reports

Odds Ratio Reports/Plots - Risk Difference Reports/Plots

Indicate whether to display reports and plots about this outcome measure. You must check at least one of the three outcome measures.

Summary Report - Outcome Detail Reports

Indicate whether to display the corresponding report.

Select Plots

Forest Plot – L'Abbe Plot

Indicate whether to display the corresponding plot.

Report Options

Alpha Level

This setting controls the confidence coefficient used in the confidence limits. Note that 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05, which results in 95% confidence intervals.

Show Notes

Indicate whether to show the notes at the end of reports. Although these notes are helpful at first, they may tend to clutter the output. This option lets you omit them.

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.

Report Options – Decimal Places

Probability Values – Ratio Values

This setting controls the number of digits to the right of the decimal place that are displayed when showing this item.

Plot Options

Plot Options – Legend Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. Click the button on the right to specify the font size, color, and style of the legend text. The characters {G} are replaced with the name of the Group Variable.

Plot Options – Plot Symbol Options

Symbols Proportional to Sample Size

Check this box to cause the size of the plotting symbols on forest plots and L'Abbe plots to be proportional to relative study size. The larger the sample size, the larger the symbol. The range of the size of the symbol is controlled by the Size Min Pcnt and Size Max Pcnt options below.

Size Min Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the smallest sample size. The recommended value is 50. Typical values range from 20 to 99.

The formula for a symbol's size is

$$\text{Actual Symbol Size} = (\text{Normal Symbol Size})*\text{Radius}$$

where

$$\text{Radius} = [(\text{Min Pct}) + (\text{Max Pct} - \text{Min Pct})*(\text{Sample Size})/(\text{Max Sample Size})]/100$$

Size Max Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the largest sample size. The recommended value is 150. Typical values range from 101 to 200.

Forest Plot Tab

The options on this panel control the appearance of the forest plot.

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.

Log Scale

This option controls the scaling of horizontal axis. We suggest that you use a logarithmic scale for the odds ratio and risk ratio. The risk difference forest plot will automatically revert to a regular scale since the logarithm of negative numbers is not defined.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forest 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 Forest style file is used. These style files are created in the Scatter Plot procedure.

Line

This option lets you indicate whether to display the reference line and the characteristics of that line.

Ratio Value

This is the position of the reference line on the odds ratio and risk ratio forest plots.

Difference Value

This is the position of the reference line on the risk difference forest plots.

Titles

Plot Title

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

Radial Plot Tab

The options on this panel control the appearance of the radial plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Radial 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. These style files are created in the Scatter Plot procedure.

Symbol

Specify a symbol. Usually, no symbol is used.

Symbol Font Size

This option lets you specify the size of font used to display the row numbers or row labels.

Titles

Plot Title

This is the text of the title. The characters *{G}* are replaced by the output measure. Press the button on the right of the field to specify the font of the text.

L'Abbe Plot Tab

The options on this panel control the appearance of the L'Abbe plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

L'Abbe 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. These style files are created in the Scatter Plot procedure.

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 Tab

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fourteen groups may be used. When no Group Variable is specified, the options made for Symbol 1 are used to define the plot symbol.

Plotting Symbols

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol.

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 on the Spreadsheet

P1 – Risk Diff. Weights

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 – Meta-Analysis of Correlated Proportions

This section presents an example of how to analyze the data contained in the METACPROP database. This dataset contains data for 24 matched case-control studies. The response of each case subject was compared to the response of a matched control subject.

You may follow along here by making the appropriate entries or load the completed template **Example 1** from the Template tab of the Meta-Analysis of Correlated Proportions window.

1 Open the METACPROP dataset.

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

2 Open the Meta-Analysis of Correlated Proportions window.

- On the menus, select **Analysis**, then **Meta-Analysis**, then **Meta-Analysis of Correlated Proportions**. The procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Variables tab**.
- Set the **N11 Count (A) Variable** to **CaseYes**.
- Set the **N10 Count (B) Variable** to **CaseNo**.
- Set the **N01 Count (C) Variable** to **ControlYes**.
- Set the **N00 Count (D) Variable** to **ControlNo**.
- Set the **Label Variable** to **Study**.
- Set the **Group Variable** to **Diet**.
- Set **Combine Studies Using** to **Random Effects Method**.
- Set the **Change Zero Counts To (Delta)** to **0.0**.

4 Specify the reports.

- Select the **Reports tab**.
- Check the **Odds Ratio Reports/Plots** option box.
- Check the **Summary Report** option box.
- Check the **Heterogeneity Tests** option box.
- Check the **Outcome Detail Reports** option box.
- Check the **Forest Plot (By Group & Measure)** option box.
- Check **Radial Plot** option box.
- Check the **L'Abbe Plot** option box.

5 Specify the L'Abbe plot.

- Select the **L'Abbe Plot tab**.
- Set the **Vertical Axis Minimum** to **0.2**.
- Set the **Vertical Axis Maximum** to **0.8**.
- Set the **Horizontal Axis Minimum** to **0.2**.
- Set the **Horizontal Axis Maximum** to **0.8**.
- Press the **Vertical Axis – Tick Label Setting** button and set the decimal places to **2**.
- Press the **Horizontal Axis – Tick Label Settings** button and set the decimal places to **2**.

6 Specify the plotting symbols.

- Select the **Symbols tab**.
- Set the **Group 2 Symbol Type** to **Solid Circle**.

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).

Run Summary Section

Parameter	Value	Parameter	Value
N11 Count (A) Variable	CaseYes	Rows Processed	24
N10 Count (B) Variable	CaseNo	Number Groups	2
N01 Count (C) Variable	ControlYes	Delta Value	0
N00 Count (D) Variable	ControlNo		
Row Label Variable	Study		
Group Variable	Diet		

This report records the variables that were used and the number of rows that were processed.

Numeric Summary Section

```
[Treatment]                                                  Odds       Risk        Risk
StudyId        Data                       P1        P2       Ratio      Ratio  Difference
[A]
S1             25/43  6/23            0.6515    0.4697      3.0000     1.3871      0.1818
S2             44/79  15/49           0.6172    0.4609      2.3333     1.3390      0.1563
S4             26/51  10/29           0.6375    0.4500      2.5000     1.4167      0.1875
S7             26/73  10/26           0.7374    0.3636      4.7000     2.0278      0.3737
S10            23/48  8/21            0.6957    0.4493      3.1250     1.5484      0.2464
S13            28/66  6/23            0.7416    0.3820      6.3333     1.9412      0.3596
S16            25/42  10/29           0.5915    0.4930      1.7000     1.2000      0.0986
S19            29/46  10/26           0.6389    0.5417      1.7000     1.1795      0.0972
S20            44/76  18/47           0.6179    0.5041      1.7778     1.2258      0.1138
S22            25/43  8/21            0.6719    0.5156      2.2500     1.3030      0.1563
S24            75/123  15/97          0.5591    0.4091      3.2000     1.3667      0.1500
Average                                                     2.6640     1.4040      0.1906

[B]
S3             53/72  21/43           0.6261    0.6435      0.9048     0.9730     -0.0174
S5             73/108  49/97          0.5268    0.5951      0.7143     0.8852     -0.0683
S6             58/97  37/103          0.4850    0.4750      1.0541     1.0211      0.0100
S8             42/74  18/47           0.6116    0.4959      1.7778     1.2333      0.1157
S9             56/98  14/39           0.7153    0.5109      3.0000     1.4000      0.2044
S11            71/112  21/63          0.6400    0.5257      1.9524     1.2174      0.1143
S12            60/108  28/89          0.5482    0.4467      1.7143     1.2273      0.1015
S14            46/81  15/49           0.6231    0.4692      2.3333     1.3279      0.1538
S15            58/77  21/43           0.6417    0.6583      0.9048     0.9747     -0.0167
S17            74/126  13/61          0.6738    0.4652      4.0000     1.4483      0.2086
S18            62/101  31/97          0.5101    0.4697      1.2581     1.0860      0.0404
S21            58/77  14/39           0.6638    0.6207      1.3571     1.0694      0.0431
S23            117/158  11/53         0.7488    0.6066      3.7273     1.2344      0.1422
Average                                                     1.6166     1.1481      0.0804

[Combined]
Average                                                     1.9972     1.2448      0.1259

Note: This report shows the input data and the three outcomes for each study in the analysis. The 'Average'
values are actually weighted averages with weights based on the effects model that was selected.
```

This report summarizes the input data. You should scan it for any mistakes. Note that the ‘Average’ lines provide the estimated group averages. The values depend on your selection of whether the Random Effects model or Fixed Effects model was used. The ‘Combined’ line provides the combined results of all studies.

Data

These are the count values that were read from the database.

P1

This is the estimated event proportion for variable 1 (the cases).

P2

This is the estimated event proportion for variable 2 (the controls).

Odds Ratio

This is the estimated value of the odds ratio. Note that it depends not only on the data, but also on the delta value used.

Risk Ratio

This is the estimated value of the risk ratio. Note that it depends not only on the data, but also on the delta value used.

Risk Difference

This is the estimated value of the risk difference. Note that it depends not only on the data, but also on the delta value used.

Nondirectional Zero-Effect Test

Diet	Outcome Measure	Chi-Square	DF	Prob Level
A	Odds Ratio	90.7010	11	0.0000
B	Odds Ratio	74.6044	13	0.0000
Combined	Odds Ratio	165.3054	24	0.0000

This reports the results of the nondirectional zero-effect chi-square test designed to test the null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative hypothesis is that at least one $\theta_i \neq 0$, that is, at least one study had a statistically significant result.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies.

Prob Level

This is the significance level of the test. If this value is less than the nominal value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Directional Zero-Effect Test

Diet	Outcome Measure	Chi-Square	DF	Prob Level
A	Odds Ratio	78.7597	1	0.0000
B	Odds Ratio	29.4196	1	0.0000
Combined	Odds Ratio	90.9788	1	0.0000

This reports the results of the directional zero-effect chi-square test designed to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = 0 \quad i = 1, \cdots, k$$

The alternative hypothesis is that $\theta_i = \theta \neq 0$ for all *i*, that is, that all effects are equal to the same, non-zero value.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal one.

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Effect-Equality (Heterogeneity) Test

```
                     Outcome        Cochran's                 Prob
Diet                 Measure                Q       DF       Level
A                    Odds Ratio       11.9413       10      0.2890
B                    Odds Ratio       45.1848       12      0.0000
Combined             Odds Ratio       74.3266       23      0.0000
```

This reports the results of the effect-equality (homogeneity) test. This chi-square test was designed to test the null hypothesis that all treatment effects are equal. The null hypothesis is written

$$H_0\colon \theta_i = \theta \quad i = 1, \cdots, k$$

The alternative is that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction. This test may help you determine whether to use a Fixed Effects model (used for homogeneous effects) or a Random Effects model (heterogeneous effects).

Cochran's Q

This is the computed chi-square value for Cochran's Q statistic. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies minus one..

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Odds Ratio Detail Section

				95.0%	95.0%	Percent
				Lower	Upper	Random
[Diet]			Odds	Confidence	Confidence	Effects
Study	P1	P2	Ratio	Limit	Limit	Weight
[A]						
S1	0.6515	0.4697	3.0000	1.1909	7.5576	3.0697
S2	0.6172	0.4609	2.3333	1.2744	4.2723	4.3247
S4	0.6375	0.4500	2.5000	1.2007	5.2051	3.7800
S7	0.7374	0.3636	4.7000	2.3750	9.3009	3.9902
S10	0.6957	0.4493	3.1250	1.4096	6.9280	3.5315
S13	0.7416	0.3820	6.3333	2.6773	14.9818	3.2895
S16	0.5915	0.4930	1.7000	0.7784	3.7126	3.5898
S19	0.6389	0.5417	1.7000	0.7784	3.7126	3.5898
S20	0.6179	0.5041	1.7778	0.9979	3.1671	4.4454
S22	0.6719	0.5156	2.2500	0.9783	5.1746	3.3928
S24	0.5591	0.4091	3.2000	1.7921	5.7140	4.4352
Average			2.6640	2.1011	3.3776	
[B]						
S3	0.6261	0.6435	0.9048	0.4864	1.6828	4.2560
S5	0.5268	0.5951	0.7143	0.4629	1.1022	5.0815
S6	0.4850	0.4750	1.0541	0.6722	1.6528	5.0115
S8	0.6116	0.4959	1.7778	0.9979	3.1671	4.4454
S9	0.7153	0.5109	3.0000	1.6385	5.4930	4.3247
S11	0.6400	0.5257	1.9524	1.1538	3.3035	4.6743
S12	0.5482	0.4467	1.7143	1.0756	2.7321	4.9400
S14	0.6231	0.4692	2.3333	1.2744	4.2723	4.3247
S15	0.6417	0.6583	0.9048	0.4864	1.6828	4.2560
S17	0.6738	0.4652	4.0000	2.1783	7.3452	4.3120
S18	0.5101	0.4697	1.2581	0.7850	2.0161	4.9156
S21	0.6638	0.6207	1.3571	0.6805	2.7067	3.9575
S23	0.7488	0.6066	3.7273	1.9158	7.2514	4.0623
Average			1.6166	1.2010	2.1759	
[Combined]						
Average			1.9972	1.5913	2.5065	

This report displays results for the odds ratio outcome measure. You can obtain a similar report for the risk ratio and the risk difference. The report gives you the

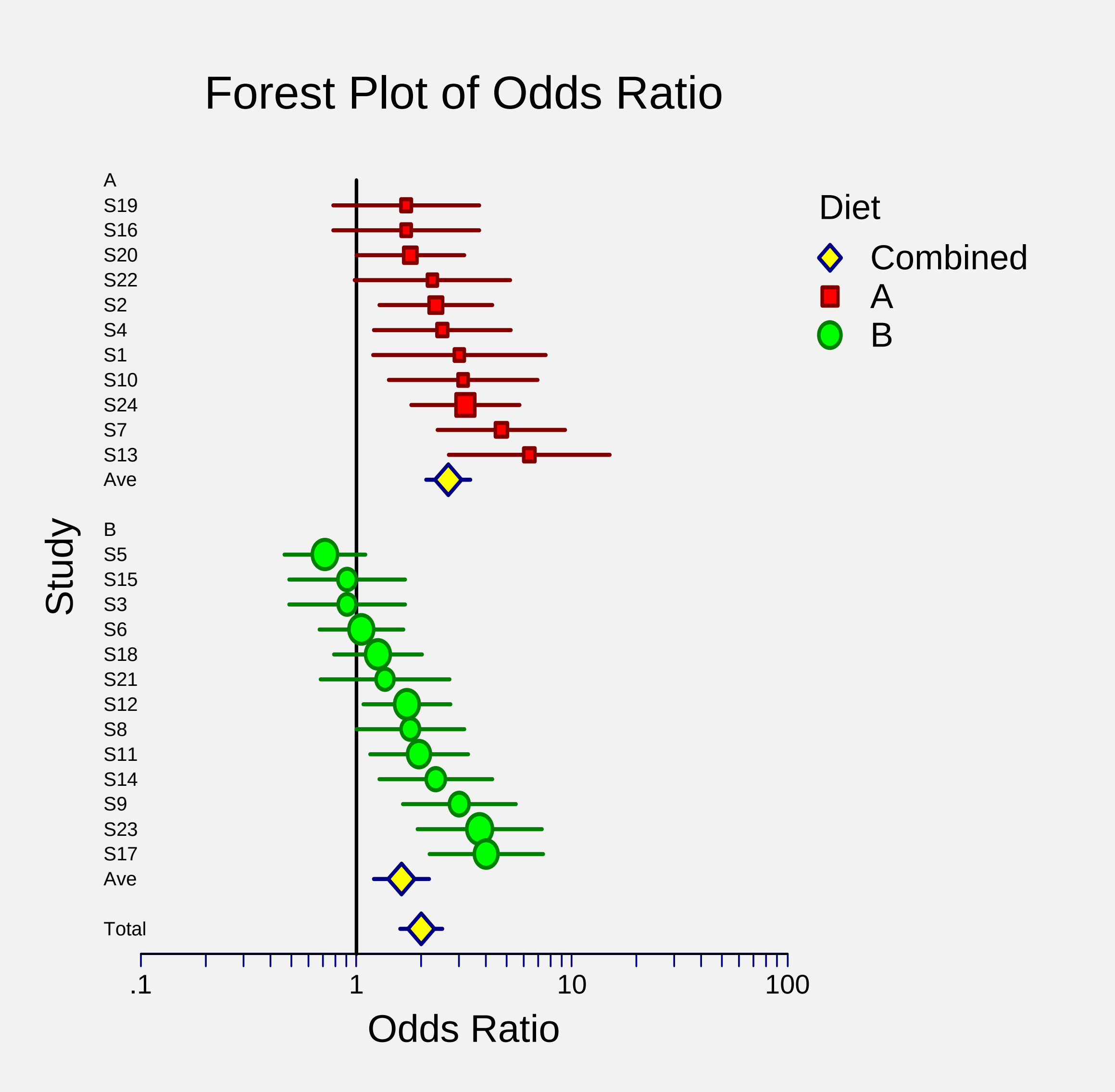
Confidence Limits

These are the lower and upper confidence limits (the formulas were given earlier in this chapter).

Weights

The last column gives the relative (percent) weight used in creating the weighted average. Using these values, you can decide how much influence each study has on the weighted average.

Forest Plot

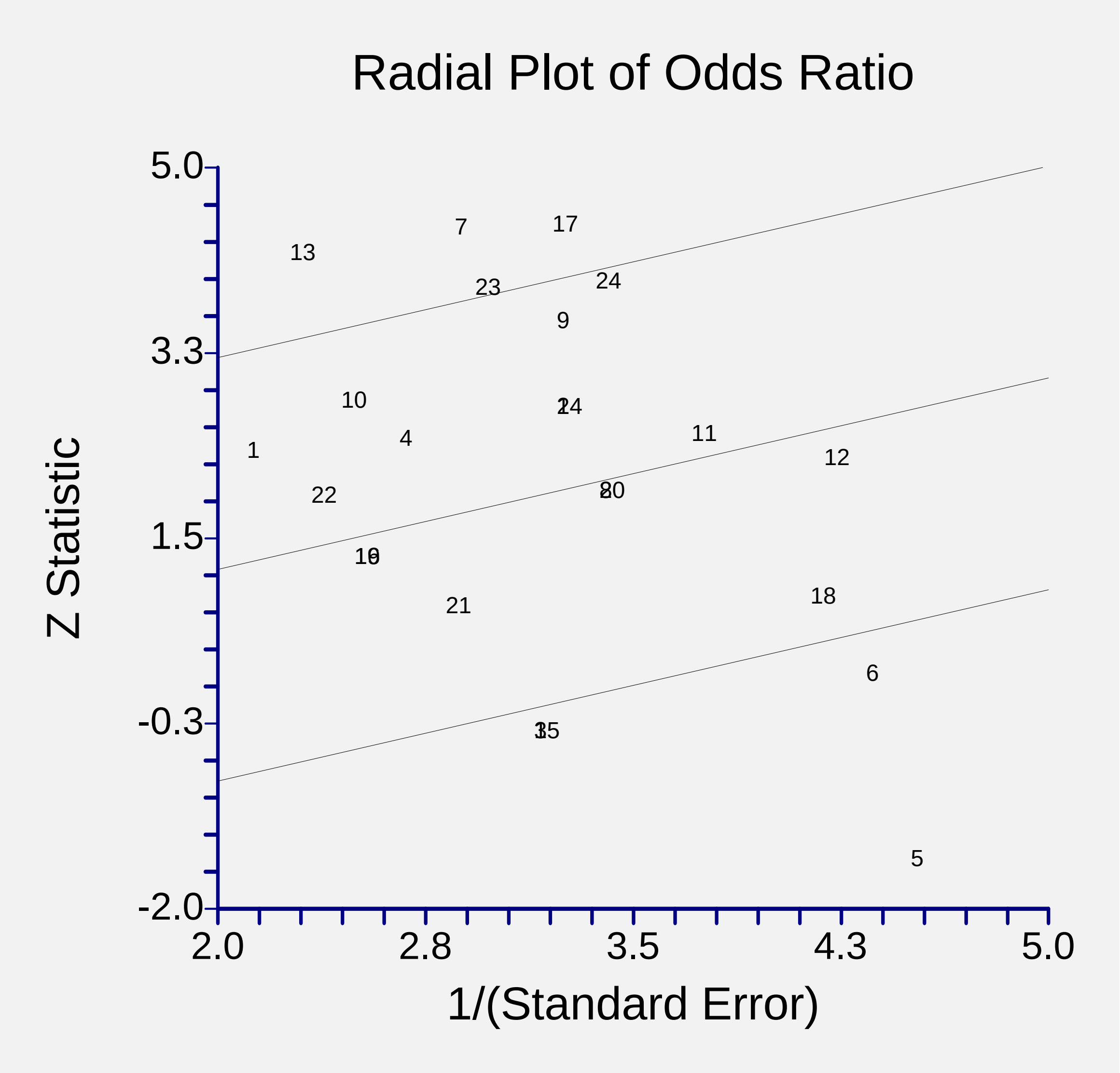


This plot presents the results for each study on one plot. The size of the plot symbol is proportional to the sample size of the study. The points on the plot are sorted by group and by the odds ratio. The lines represent the confidence intervals about the odds ratios. Note that the narrower the confidence limits, the better.

By studying this plot, you can determine the main conclusions that can be drawn from the set of studies. For example, you can determine how many studies were significant (the confidence limits do not intersect the vertical line at 1.0). You can see if there were different conclusions for the different groups.

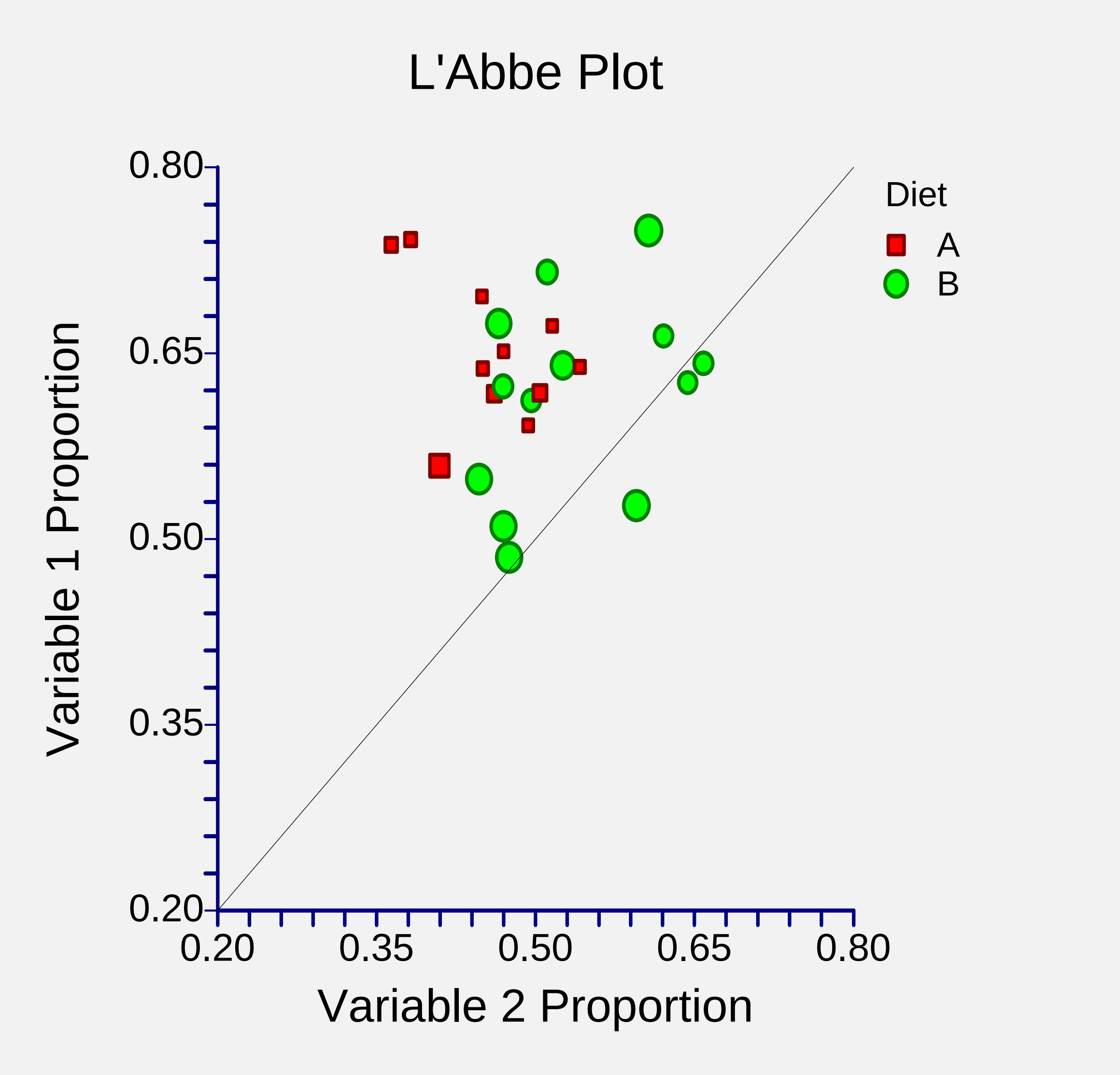
The results of the combining the studies are displayed at the end of each group.

Radial Plot



The radial (or Galbraith) plot shows the z-statistic (outcome divided by standard error) on the vertical axis and a measure of weight on the horizontal axis. Studies that have the largest weight are closest to the Y axis. Studies within the limits are interpreted as homogeneous. Studies outside the limits may be outliers.

L’Abbe Plot



The L’Abbe plot displays the variable 1 (case) proportion on vertical axis versus the variable 2 (control) proportion on the horizontal axis. Homogenous studies will be arranged along the diagonal line. This plot is especially useful in determining if the relationship between the two variables is the same for all values of variable 2.

Chapter 458

Meta-Analysis of Hazard Ratios

Introduction

This module performs a meta-analysis on a set of two-group, time to event (survival), studies in which some data may be censored. These studies have a treatment group and a control group. Each study's result may be summarized by the log hazard ratio and its standard error. The program provides a complete set of numeric reports and plots to allow the investigation and presentation of the studies. The plots include the *forest plot* and *radial plot.* Both fixed-, and random-, effects models are available for analysis.

Meta-Analysis refers to methods for the systematic review of a set of individual studies with the aim to combine their results. Meta-analysis has become popular for a number of reasons:

1. The adoption of evidence-based medicine which requires that all reliable information is considered.
2. The desire to avoid narrative reviews which are often misleading.
3. The desire to interpret the large number of studies that may have been conducted about a specific treatment.
4. The desire to increase the statistical power of the results be combining many small-size studies.

The goals of meta-analysis may be summarized as follows. A meta-analysis seeks to systematically review all pertinent evidence, provide quantitative summaries, integrate results across studies, and provide an overall interpretation of these studies.

We have found many books and articles on meta-analysis. In this chapter, we briefly summarize the information in Sutton *et al.* (2000) and Thompson (1998). Refer to those sources for more details about how to conduct a meta-analysis.

As for the particular topic of combining hazard ratio studies in a meta-analysis, the book by Parmar and Machin (1995) and the paper by Parmar *et al.*(1998) are essential reading. The paper provides instructions on how to obtain estimates of the hazard ratio and its standard error from trials that do not report these items explicitly (a situation that is common).

Treatment Effect – Hazard Ratio

The most recommended single summary statistic for quantifying the treatment effect in studies using survival data is the (log) hazard rate. This statistic is chosen because it can be calculated from time-to-event data with censoring and because it measures the size of the difference between two Kaplan-Meier curves.

The Cox-Mantel estimate of the *hazard ratio* is formed by dividing the hazard rate under treatment by the hazard rate under control. Thus, it measures the change in risk of treatment versus control over the follow-up period. Since the distribution of the log hazard ratio is nearly normal, the log transformation is applied. The formula for the hazard rate is

$$\begin{aligned} HR_{CM} &= \frac{H_T}{H_C} \\ &= \frac{O_T / E_T}{O_C / E_C} \end{aligned}$$

where O_i is the observed number of events (deaths) in group i , E_i is the expected number of events (deaths) in group *i*, and H_i is the overall hazard rate for the ith group. The calculation of the E_i is explained in Parmar and Machin (1995).

A confidence interval for *HR* is found by first transforming to the log scale which is better approximated by the normal distribution, calculating the limits, and then transforming back to the original scale. The calculation is made using

$$\ln\left(HR_{CM}\right) \pm z_{1-\alpha/2}\left(SE_{\ln HR_{CM}}\right)$$

where

$$SE_{\ln HR_{CM}} = \sqrt{\frac{1}{E_T} + \frac{1}{E_C}}$$

An alternative estimate of *HR* that is sometimes used is the Mantel-Haenszel estimator which is calculated using

$$HR_{MH} = \exp\left(\frac{O_T - E_T}{V}\right)$$

where *V* is the hypergeometric variance. For further details, see Parmar and Machin (1995). A confidence interval for *HR* is found by first transforming to the log scale which is better approximated by the normal distribution, calculating the limits, and then transforming back to the original scale. The calculation is made using

$$\ln\left(HR_{MH}\right) \pm z_{1-\alpha/2}\left(SE_{\ln HR_{MH}}\right)$$

where

$$SE_{\ln HR_{MH}} = \sqrt{\frac{1}{V}}$$

If the log hazard ratio and its standard error are not reported in a particular study it will have to be estimated from the logrank test statistic, p-value, or from the Kaplan-Meier curves. Details of how to do this are presented in Parmar *et al.* (1998).

Suppose you have obtained the results for *k* studies, labeled *i = 1,…,k*. Each study consists of a treatment group (T) and a control group (C). The results of each study are summarized by two statistics:

$\ln(HR_i)$ the log hazard ratio.

$SE_{\ln(HR_i)}$ the standard error of the log hazard ratio.

It will be useful in the sequel to make the following definition of the weights.

$$v_i = (SE_{\ln HR})^2$$

$$w_i = 1/v_i$$

Hypothesis Tests

In the discussion below, we let θ_i represent $\ln HR_i$. Several hypothesis tests have be developed to test the various hypotheses that may be of interest. These will be defined next.

Overall Null Hypothesis

Two statistical tests have been devised to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = \theta \quad i = 1,\cdots,k$$

Nondirectional Test

The nondirectional alternative hypothesis that at least one $\theta_i \neq 0$ may be tested by comparing the quantity

$$X_{ND} = \sum_{i=1}^{k} w_i \theta_i^2$$

with a χ_k^2 distribution.

Directional Test

A test of the more interesting directional alternative hypothesis that $\theta_i = \theta \neq 0$ for all *i* may be tested by comparing the quantity

$$X_D = \frac{\left(\sum_{i=1}^{k} w_i \hat{\theta}_i\right)^2}{\sum_{i=1}^{k} w_i}$$

with a χ_1^2 distribution. Note that this tests the hypothesis that all effects are equal to the same nonzero quantity.

Effect-Equality (Heterogeneity) Test

When the overall null hypothesis is rejected, the next step is to test whether all effects are equal, that is, whether the effects are homogeneous. Specifically, the hypothesis is

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

versus the alternative that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction.

This hypothesis is tested using Cochran's Q test which is given by

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2$$

where

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

The test is conducted by comparing Q to a χ^2_{k-1} distribution.

Fixed versus Random Effects Combined Confidence Interval

If the effects are assumed to be equal (homogeneous), either through testing or from other considerations, a *fixed effects model* may be used to construct a combined confidence interval. However, if the effects are heterogeneous, a *random effects model* should be used to construct the combined confidence interval.

Fixed Effects Model

The fixed effects model assumes homogeneity of study results. That is, it assumes that $\theta_i = \theta$ for all *i*. This assumption may not be realistic when combining studies with different patient pools, protocols, follow-up strategies, doses, durations, etc.

If the fixed effects model is adopted, the *inverse variance-weighted* method as described by Sutton (2000) page 58 is used to calculate the confidence interval for θ. The formulas used are

$$\hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}\left(\hat{\theta}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

$$\hat{V}\left(\hat{\theta}\right)=\frac{1}{\sum_{i=1}^{k} w_i}$$

Random Effects Model

The random effects model assumes that the individual θ_i come from a random distribution with fixed mean $\bar{\theta}$ and variance σ^2. Sutton (2000) page 74 presents the formulas necessary to conduct a random effects analysis using the *weighted* method. The formulas used are

$$\hat{\hat{\theta}} \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\hat{\theta}}\right)}$$

where $z_{1-\alpha/2}$ is the appropriate percentage point from the standardized normal distribution and

$$\hat{\hat{\theta}} = \frac{\sum_{i=1}^{k} \bar{w}_i \hat{\theta}_i}{\sum_{i=1}^{k} \bar{w}_i}$$

$$\hat{V}\left(\hat{\hat{\theta}}\right) = \frac{1}{\sum_{i=1}^{k} \bar{w}_i}$$

$$\bar{w}_i = \frac{1}{\frac{1}{w_i} + \hat{\tau}^2}$$

$$\hat{\tau}^2 = \begin{cases} \dfrac{Q-k+1}{U} & \text{if } Q > k-1 \\ 0 & \text{otherwise} \end{cases}$$

$$Q = \sum_{i=1}^{k} w_i \left(\hat{\theta}_i - \hat{\theta}\right)^2$$

$$U = (k-1)\left(\bar{w} - \frac{s_w^2}{k\bar{w}}\right)$$

$$s_w^2 = \frac{1}{k-1}\left(\sum_{i=1}^{k} w_i^2 - k\bar{w}^2\right)$$

$$\bar{w} = \frac{1}{k}\left(\sum_{i=1}^{k} w_i\right)$$

Graphical Displays

A number of plots have been devised to display the information in a meta-analysis. These include the forest plot, the radial plot, and the L'Abbe plot. More will be said about each of these plots in the Output section.

Data Structure

The data are entered into a dataset using one row per study. Two variables are required to hold the log hazard ratio and its standard error. In addition to these, an additional variable is usually used to hold a short (3 or 4 character) label. Another variable may be used to hold a grouping variable.

As an example, we will use a dataset giving the results for survival studies. The results of these studies are recorded in the METAHR database. You should load this database to see how the data are arranged.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

The options on this screen control the variables that are used in the analysis.

Variables

Log(Hazard Ratio) Variable

Specify the variable containing the log hazard ratio of each study. Each row of data represents a separate study. Note that the base of the logarithm (e or 10) is arbitrary. However, it must be consistent throughout the dataset.

S.E. Log(Hazard Ratio) Variable

Specify the variable containing the standard error of the log hazard ratio of each study. Each row of data represents a separate study. Note that the base of the logarithm (e or 10) is arbitrary. However, it must be consistent throughout the dataset.

Variables – Optional Variables

Label Variable

Specify an optional variable containing a label for each study (row) in the database. This label should be short (< 8 letters) so that it can fit on the plots.

Group Variable

Specify an optional variable containing a group identification value. Each unique value of this variable will receive its own plotting symbol on the forest plots. Some reports are sorted by these group values.

Combine Studies Method

Combine Studies Using

Specify the method used to combine treatment effects.

Use the **Fixed Effects** method when you do not want to account for the variation between studies.

Use the **Random Effects** method when you want to account for the variation between studies as well as the variation within the studies.

Reports Tab

The options on this screen control the appearance of the reports.

Select Reports

Summary Report - Outcome Detail Reports

Indicate whether to display the corresponding report.

Select Plots

Forest Plot – Radial Plot

Indicate whether to display the corresponding plot.

Report Options

Alpha Level

This setting controls the confidence coefficient used in the confidence limits. Note that 100 x (1 - alpha)% confidence limits will be calculated. This must be a value between 0 and 0.5. The most common choice is 0.05, which results in 95% confidence intervals.

Show Notes

Indicate whether to show the notes at the end of reports. Although these notes are helpful at first, they may tend to clutter the output. This option lets you omit them.

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.

Report Options – Decimal Places

Probability Values – Z Values

This setting controls the number of digits to the right of the decimal place that are displayed when showing this item.

Plot Options

Plot Options – Legend Options

Show Legend

Specifies whether to display the legend.

Legend Text

Specifies the title of the legend. Click the button on the right to specify the font size, color, and style of the legend text. The characters {G} are replaced with the name of the Group Variable.

Plot Options – Plot Symbol Options

Symbols Proportional to Sample Size

Check this box to cause the size of the plotting symbols on forest plots and L'Abbe plots to be proportional to relative study size. The larger the sample size, the larger the symbol. The range of the size of the symbol is controlled by the Size Min Pcnt and Size Max Pcnt options below.

Size Min Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the smallest sample size. The recommended value is 50. Typical values range from 20 to 99.

The formula for a symbol's size is

$$\text{Actual Symbol Size} = (\text{Normal Symbol Size})*\text{Radius}$$

where

$$\text{Radius} = [(\text{Min Pct}) + (\text{Max Pct} - \text{Min Pct})*(\text{Sample Size})/(\text{Max Sample Size})]/100$$

Size Max Pcnt

When the Symbols Proportional to Sample Size option is checked, this is percentage adjustment that occurs to the largest sample size. The recommended value is 150. Typical values range from 101 to 200.

Forest Plot Tab

The options on this panel control the appearance of the forest plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forest 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 Forest style file is used. These style files are created in the Scatter Plot procedure.

Ref. Line

This option lets you indicate whether to display the reference line and the characteristics of that line.

Difference Reference Value

This is the position of the reference line on the forest plot.

Titles

Plot Title

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

Radial Plot Tab

The options on this panel control the appearance of the radial plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Radial 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. These style files are created in the Scatter Plot procedure.

Symbol

Specify a symbol. Usually, no symbol is used.

Symbol Font Size

This option lets you specify the size of font used to display the row numbers or row labels.

Titles

Plot Title

This is the text of the title. The characters *{G}* are replaced by the output measure. Press the button on the right of the field to specify the font of the text.

Symbols Tab

These options set the type, color, size, and style of the plotting symbols. Symbols for up to fourteen groups may be used. When no Group Variable is specifed, the options made for Symbol 1 are used to define the plot symbol.

Plotting Symbols

Click on the symbol or on the button to its right to display a window that allows you to change the characteristics of the plotting symbol.

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 are automatically replaced, so be careful.

Data Storage Options – Select Items to Store on the Spreadsheet

Log(HR) - Weights

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 – Meta-Analysis of Hazard Ratios

This section presents an example of how to analyze the data contained in the METAHR database. This dataset contains data for sixteen randomized clinical trials with survival endpoints.

You may follow along here by making the appropriate entries or load the completed template **Example 1** from the Template tab of the Meta-Analysis of Hazard Ratios window.

1 **Open the METAHR dataset.**

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

2 **Open the Meta-Analysis of Hazard Ratios window.**

- On the menus, select **Analysis**, then **Meta-Analysis**, then **Meta-Analysis of Hazard Ratios**. The Meta-Analysis of Hazard Ratios procedure window will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Select the variables.

- Select the **Variables tab**.
- Set the **Log(Hazard Ratio) Variable** to **LogHR**.
- Set the **S.E. Log(Hazard Ratio) Variable** to **SELogHR**.
- Set the **Label Variable** to **Study**.

4 Specify the reports.

- Select the **Reports tab**.
- Check the **Summary Report** option box.
- Check the **Heterogeneity Tests** option box.
- Check the **Outcome Detail Reports** option box.
- Check the **Forest Plot (By Measure)** option box.
- Check **Radial Plot** option box.

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
Log HR Variable           LogHR          SE(Log HR) Variable     SELogHR
Group Variable            None           Number Groups           1
Row Label Variable        Study          Rows Processed          16
```

This report records the variables that were used and the number of rows that were processed.

Numeric Summary Section

```
Study       Log HR    SE(Log HR)
S1         -0.1350        0.0799
S2         -0.2570        0.0734
S3         -0.4610        0.0492
S4          0.2030        0.0401
S5         -0.7980        0.1203
S6         -0.3240        0.0933
S7         -0.5510        0.0577
S8         -0.6820        0.1084
S9         -0.3340        0.1385
S10        -0.3840        0.0472
S11         0.0564        0.0671
S12        -0.9910        0.0528
S13        -0.7230        0.0319
S14        -0.4240        0.0289
S15         0.0178        0.0817
S16        -0.1870        0.0203

[Combined]
Average    -0.3712
```

This report shows the input data. You should scan it for any mistakes. Note that the ‘Average’ line provides the estimated group average.

Nondirectional Zero-Effect Test

	Outcome			Prob
Rows	Measure	Chi-Square	DF	Level
Combined	Log(Hazard Ratio)	1554.1876	16	0.0000

This reports the results of the nondirectional zero-effect chi-square test designed to test the null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

The alternative hypothesis is that at least one $\theta_i \neq 0$, that is, at least one study had a statistically significant result.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies.

Prob Level

This is the significance level of the test. If this value is less than the nominal value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Directional Zero-Effect Test

	Outcome			Prob
Rows	Measure	Chi-Square	DF	Level
Combined	Log(Hazard Ratio)	902.7014	1	0.0000

This reports the results of the directional zero-effect chi-square test designed to test the overall null hypothesis that all treatment effects are zero. The null hypothesis is written

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

The alternative hypothesis is that $\theta_i = \theta \neq 0$ for all *i*, that is, that all effects are equal to the same, non-zero value.

Chi-Square

This is the computed chi-square value for this test. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal one.

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Effect-Equality (Heterogeneity) Test

```
               Outcome             Cochran's                   Prob
Treatment      Measure                     Q          DF      Level
Combined       Log(Hazard Ratio)    651.4861          15     0.0000
```

This reports the results of the effect-equality (homogeneity) test. This chi-square test was designed to test the null hypothesis that all treatment effects are equal. The null hypothesis is written

$$H_0 : \theta_i = \theta \quad i = 1, \cdots, k$$

The alternative is that at least one effect is different, that is, that the effects are heterogeneous. This may also be interpreted as a test of the study-by-treatment interaction. This test may help you determine whether to use a Fixed Effects model (used for homogeneous effects) or a Random Effects model (heterogeneous effects).

Cochran’s Q

This is the computed chi-square value for Cochran’s Q statistic. The formula was presented earlier.

DF

This is the degrees of freedom. For this test, the degrees of freedom is equal to the number of studies minus one..

Prob Level

This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Log(Hazard Ratio) Detail Section

Study	Log Hazard Difference	Standard Error	95.0% Lower Confidence Limit	95.0% Upper Confidence Limit	Percent Random Effects Weight
S1	-0.1350	0.0799	-0.2917	0.0217	6.1950
S2	-0.2570	0.0734	-0.4009	-0.1131	6.2557
S3	-0.4610	0.0492	-0.5574	-0.3646	6.4428
S4	0.2030	0.0401	0.1244	0.2816	6.4960
S5	-0.7980	0.1203	-1.0338	-0.5622	5.7451
S6	-0.3240	0.0933	-0.5069	-0.1411	6.0591
S7	-0.5510	0.0577	-0.6641	-0.4379	6.3843
S8	-0.6820	0.1084	-0.8945	-0.4695	5.8891
S9	-0.3340	0.1385	-0.6055	-0.0625	5.5118
S10	-0.3840	0.0472	-0.4765	-0.2915	6.4553
S11	0.0564	0.0671	-0.0751	0.1879	6.3104
S12	-0.9910	0.0528	-1.0945	-0.8875	6.4190
S13	-0.7230	0.0319	-0.7855	-0.6605	6.5352
S14	-0.4240	0.0289	-0.4806	-0.3674	6.5474
S15	0.0178	0.0817	-0.1423	0.1779	6.1779
S16	-0.1870	0.0203	-0.2268	-0.1472	6.5759
[Combined]					
Average	-0.3712	0.0800	-0.5280	-0.2145	

This report displays results for the log hazard ratio.

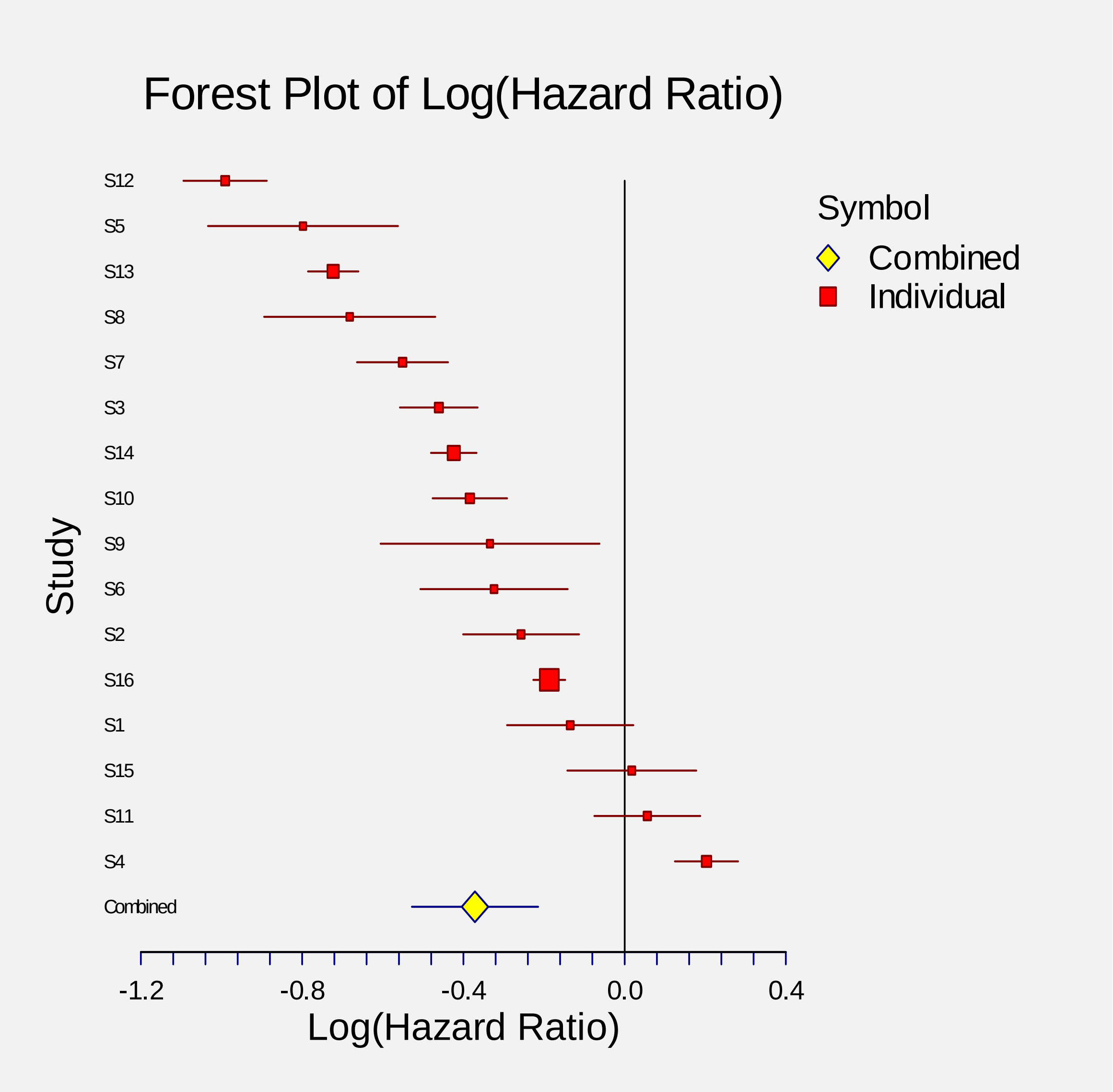
Confidence Limits

These are the lower and upper confidence limits (the formulas were given earlier in this chapter).

Weights

The last column gives the relative (percent) weight used in creating the weighted average. Using these values, you can decide how much influence each study has on the weighted average.

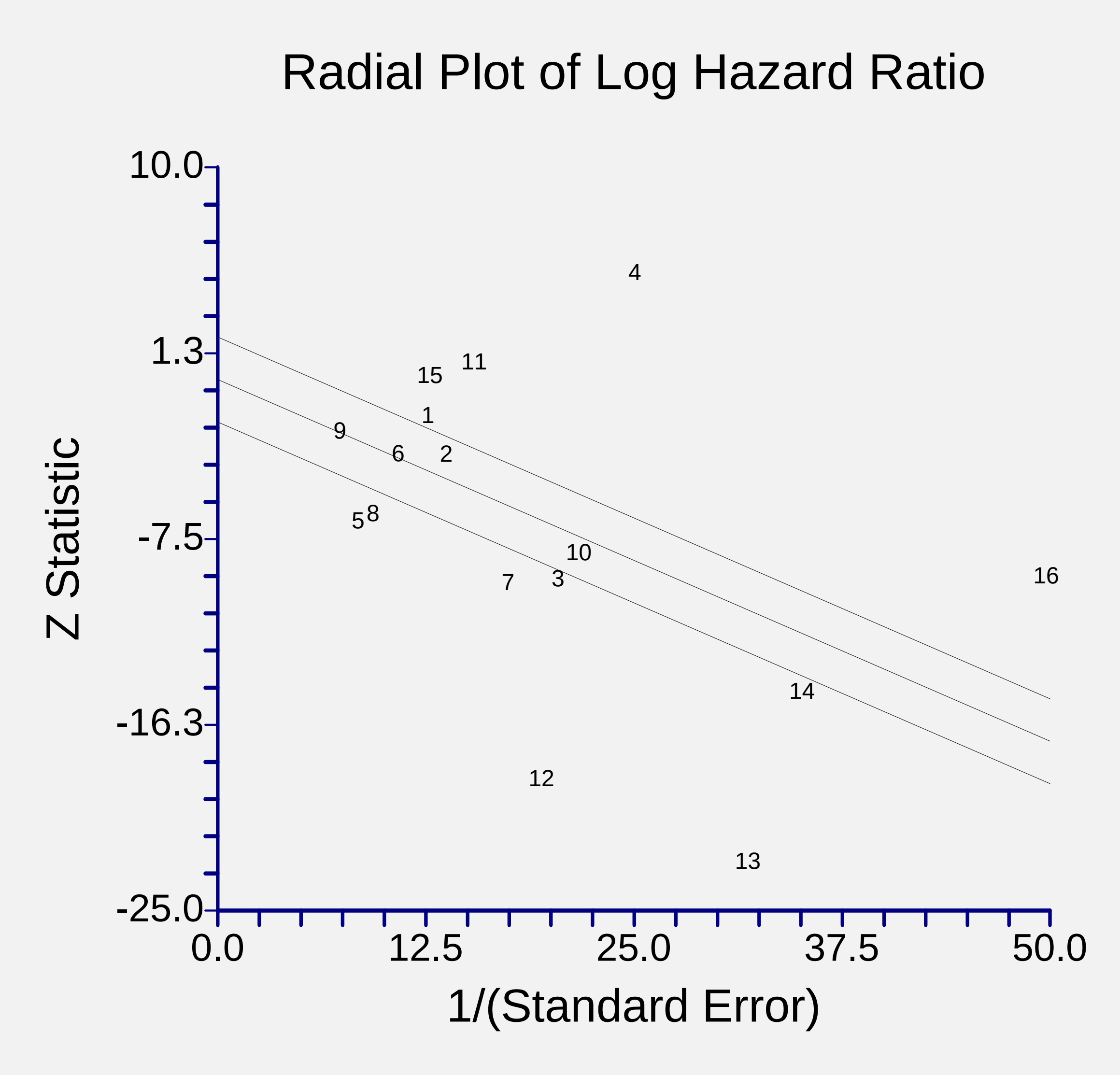
Forest Plot



This plot presents the results for each study on one plot. The size of the plot symbol is proportional to the sample size of the study. The points on the plot are sorted by the mean difference. The lines represent the confidence intervals about the log hazard ratios. Note that the narrower the confidence limits, the better.

By studying this plot, you can determine the main conclusions that can be drawn from the set of studies. For example, you can determine how many studies were significant (the confidence limits do not intersect the vertical line at 0.0).

Radial Plot



The radial (or Galbraith) plot shows the z-statistic (outcome divided by standard error) on the vertical axis and a measure of weight on the horizontal axis. Studies that have the largest weight are closest to the Y axis. Studies within the limits are interpreted as homogeneous. Studies outside the limits may be outliers.

Chapter 465

Exponential Smoothing – Horizontal

Introduction

Simple exponential smoothing forecasts horizontal series: those without trends or seasonal patterns. It is appropriate for short-term forecasts of series using a weighted average of the most recent observations.

The forecasting algorithm makes use of the following formulas:

$$F_t = \alpha X_t + (1-\alpha)F_{t-1}$$

Here α is the smoothing constant which is between zero and one.

The forecast at time *T* for the value at time *T*+*k* is F_T .

Another form of the above equation which shows how this procedure received its name is

$$F_t = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \alpha(1-\alpha)^3 X_{t-3} + \cdots$$

From this equation we see that the method constructs a weighted average of the observations. The weight of each observation decreases exponentially as we move back in time. Hence, since the weights decrease exponentially and averaging is a form of smoothing, the technique was name exponential smoothing.

Smoothing Constants

Notice that the *smoothing constant*, α , determines how fast the weights of the series decays. The value may be chosen either subjectively or objectively. Values near one put almost all weight on the most recent observations. Values of the smoothing constant near zero allow the distant past observations to have a large influence.

When selecting the smoothing constant *subjectively*, you use your own experience with this, and similar, series. Also, specifying the smoothing constant yourself lets you tune the forecast to your own beliefs about the future of the series. If you believe that the mechanism generating the series has recently gone through some fundamental changes, use a smoothing constant value of 0.9 which will cause distant observations to be ignored. If, however, you think the series is fairly stable and only going through random fluctuations, use a value of 0.1.

To select the value of the smoothing constant *objectively*, you search for a value that is best in some sense. Our program searches for that value that minimizes the size of the combined forecast errors of the currently available series. Three methods of summarizing the amount of error in the forecasts are available: the mean square error (MSE), the mean absolute error (MAE), and the mean absolute percent error (MAPE). The forecast error is the difference between the forecast of the current period made at the last period and the value of the series at the current period. This is written as

$$e_t = X_t - F_{t-1}$$

Using this formulation, we can define the three error-size criterion as follows:

$$MSE = \frac{1}{n}\sum e_t^2$$

$$MAE = \frac{1}{n}\sum \left| e_t \right|$$

$$MAPE = \frac{100}{n}\sum \left| \frac{e_t}{X_t} \right|$$

To find the value of the smoothing constants objectively, we select one of these criterion and search for that value of α that minimize this function. The program conducts a search for the appropriate values using an efficient grid-searching algorithm.

Initial Values

Exponential smoothing requires special initialization since the forecast for period one requires the forecast at period zero, which we do not, by definition, have. Several methods have been proposed for generating starting values. We have adopted the backcasting method which is currently considered to be one of the best. Backcasting is simply reversing the series so that we forecast into the past instead of into the future. This produces the required starting value. Once we have done this, we can then switch the series back and apply the algorithm in the regular manor.

Relationship to ARIMA Method

It can be shown that both exponential smoothing is equivalent to the ARIMA(0,1,1) model (see Kendall and Ord (1990) page 130). This is why backcasting is recommended for initial values.

Assumptions and Limitations

This algorithm is useful for short-term forecasting of nonseasonal time series with no apparent upward or downward. The series is assumed to have a changing (or evolving) mean that is not fixed over all time. We assume that future values of this average are unpredictable, so that the current level (current average or mean) of the series is the best forecast of future values.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by this implementation of exponential smoothing. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable(s) on which to run the analysis.

Time Series Variables

Time Series Variable(s)

Specify the variable(s) on which to run the analysis. A separate analysis will be conducted for each variable listed.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable. The forecasts are converted back to there original metric before display.

Forecasting Options

Number of Forecasts

This option specifies the number of forecasts to be generated.

Smoothing Constant Search Options

Search Method

This option specifies whether a search is conducted for the best value of the smoothing constant and what the criterion for the search will be.

- **Specified Value**

  No search is conducted. Use the value of the smoothing constant that is set in Alpha box.

- **Search on MSE**

  A search is conducted to find the value of the smoothing constant that minimizes MSE.

- **Search on MAE**

  A search is conducted to find the value of the smoothing constant that minimizes MAE.

- **Search on MAPE**

  A search is conducted to find the value of the smoothing constant that minimizes MAPE.

Alpha Smoothing Constant

When the Search Method is set to Specified Value, this option specifies the value of the smoothing constant to be used. The limits of this value are zero and one. Usually, a value between 0.1 and 0.3 are used. As the value gets closer to one, more and more weight is given to recent observations.

Reports Tab

The following options control which reports are displayed.

Select Reports

Summary Report

This option specifies whether the indicated report is displayed.

Forecast Report

This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Select Plots

Forecast Plot - Residual 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 only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Page Title

Specify a title to be shown at the top of the reports.

Forecast Plot Tab

A plot of the data and forecast over time may be displayed.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forecast 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

This option controls the attributes of the plotting 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.

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.

Residual Plot Tab

This section controls the residual plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Residual 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

This option controls the attributes of the plotting 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.

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.

Storage Tab

The forecasts and residuals may be stored on the current database for further analysis. These options let you designate which statistics (if any) should be stored by designating which variables should receive the statistics.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Forecasts

The forecasts are stored in this variable.

Residuals

The residuals are stored in this variable.

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 – Horizontal Exponential Smoothing

This section presents an example of how to generate a forecast of a horizontal series. The data in the INTEL database gives price and volume data for Intel stock during August, 1995. We will forecast values for daily volumes. These values are contained in the variable Intel_Volume.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Exponential Smoothing – Horizontal window.

1 Open the Intel dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Intel.S0**.
- Click **Open**.

2 Open the Exponential Smoothing – Horizontal window.

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Exponential Smoothing** – **Horizontal**. The Exponential Smoothing – Horizontal 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 Exponential Smoothing – Horizontal window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **Intel_Volume** from the list of variables and then click **Ok**.

4 Specify the reports.

- On the Exponential Smoothing – Horizontal window, select the **Reports tab**.
- Select **Data and Forecasts** in the **Forecast Report** list box.

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).

Forecast Summary Section

```
Forecast Summary Section

Variable                    Intel_Volume
Number of Rows              20
Mean                        10974.54
Pseudo R-Squared            0.000000
Mean Square Error           1.632774E+07
Mean |Error|                2876.168
Mean |Percent Error|        25.98573

Alpha Search                Mean Square Error
Alpha                       0.3769885
Forecast                    13100.84
```

This report summarizes the forecast equation.

Variable

The name of the variable for which the forecasts are generated.

Mean

The mean of the variable across all time periods.

Pseudo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Mean Square Error

The average squared residual (MSE) is a measure of how closely the forecasts track the actual data. The statistic is popular because it shows up in analysis of variance tables. However, because of the squaring, it tends to exaggerate the influence of outliers (points that do not follow the regular pattern).

Mean |Error|

The average absolute residual (MAE) is a measure of how closely the forecasts track the actual data without the squaring.

Mean |Percent Error|

The average percent absolute residual (MAPE) is a measure of how closely the forecasts track the actual data put on a percentage basis.

Alpha Search

If a search was made to find the best value of the smoothing constant, this row gives the criterion used during the search.

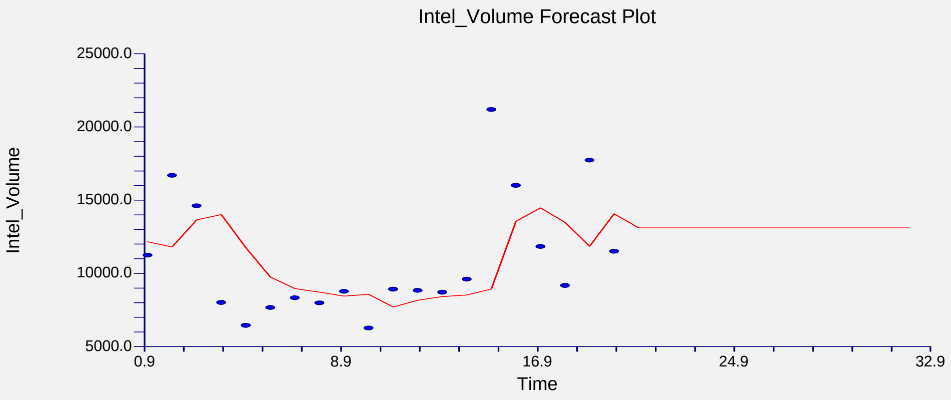
Alpha

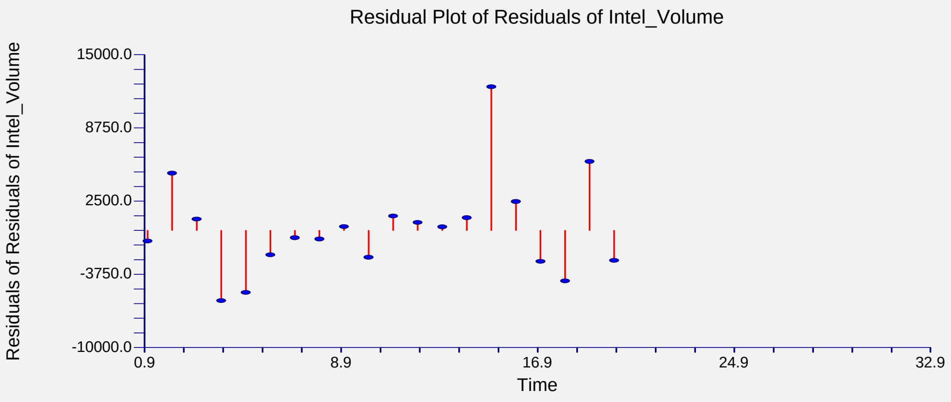
The value of the smoothing constant that was used to generate the forecasts.

Forecast

The value of the forecast. This is the value that used to forecast future values from this point on. Remember that this method does not adjust for trend or seasonality, so only the current average is used for forecasting.

Forecast and Residuals Plots





Forecast Plot

The forecast plot lets you analyze how closely the forecasts track the data. The plot also shows the forecasts at the end of the data series.

Residual Plot

This plot lets you analyze the residuals themselves. You are looking for patterns, outliers, or any other information that may help you improve the forecasting model. The first thing to compare is the scale of the Residual Plot versus the scale of the Forecast Plot. If your forecasting is working well, the vertical scale of the Residual Plot will be much less than the scale of the Forecast Plot.

Forecasts Section

```
Forecasts Section

Row    Forecast       Actual
No.    Intel_Volume   Intel_Volume   Residuals
1      12153.88       11242.2        -911.6825
2      11810.19       16689.9        4879.711
3      13649.78       14613.3        963.5162
4      14013.02       8009           -6004.018
5      11749.57       6441.8         -5307.772
6      9748.604       7664.5         -2084.103
7      8962.92        8330.3         -632.6202
8      8724.43        7983           -741.4297
9      8444.919       8767.1         322.1808
10     8566.378       6266.4         -2299.978
11     7699.313       8915.3         1215.987
12     8157.726       8833           675.2742
13     8412.297       8709.7         297.4036
14     8524.414       9603           1078.586
15     8931.028       21185.2        12254.17
16     13550.71       16006.5        2455.79
17     14476.51       11832.4        -2644.115
18     13479.71       9168.1         -4311.614
19     11854.29       17729.3        5875.015
20     14069.1        11500.7        -2568.398
21     13100.84
22     13100.84
23     13100.84
24     13100.84
25     13100.84
26     13100.84
27     13100.84
28     13100.84
29     13100.84
30     13100.84
31     13100.84
32     13100.84
```

This section shows the values of the forecasts, the actual values, and the residuals.

Chapter 466

Exponential Smoothing – Trend

Introduction

This module forecasts series with upward or downward trends. Three techniques are available: least squares trend, double smoothing, and Holt's linear trend algorithm.

Least Squares Trend

Least squares trend computes a straight-line trend equation through the data using standard least squares techniques in which the dependent variable is the time series and the independent variable is the row (sequence) number. The forecasting equation is

$$F_t = a + bt$$

where F_t is the forecast at time period t, *a* is the y-intercept, and *b* is the slope of the trend. The slope indicates how much is added (or subtracted if *b* is negative) from each time period to the next.

This method is useful for series that show a stable, long-term trend. It places the largest weights in estimation on the two ends of the series, while the rows near the middle with an insignificant impact on the estimates.

Double Exponential Smoothing

Double exponential smoothing computes a trend equation through the data using a special weighting function that places the greatest emphasis on the most recent time periods. The forecasting equation changes from period to period.

The forecasting algorithm makes use of the following formulas:

$$F_t = a_t + b_t$$

$$a_t = X_t + (1-\alpha)^2 e_t$$

$$b_t = b_{t-1} + \alpha^2 e_t$$

$$e_t = F_t - X_t$$

The smoothing constant, α, dictates the amount of smoothing that takes place. It ranges from zero to one.

The forecast at time period *T* for the value at time period *T*+*k* is $a_T + b_T k$. Double smoothing is discussed in detail in Thomopoulos (1980).

This method is included more for its historical significance, since Holt's algorithm is usually preferred to it.

Holt's Linear Trend

Holt's Linear Trend computes an evolving trend equation through the data using a special weighting function that places the greatest emphasis on the most recent time periods. Instead of the global trend equation of the least squares trend algorithm, this technique uses a local trend equation. The trend equation is modified from period to period. The forecasting equation changes from period to period.

The forecasting algorithm makes use of the following formulas:

$$a_t = \alpha X_t + (1-\alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

Here α and β are smoothing constants which are each between zero and one. Again, a_t gives the y-intercept (or level) at time *t*, while b_t is the slope at time *t*.

The forecast at time *T* for the value at time *T*+*k* is $a_T + b_T k$.

Smoothing Constants

Notice that in both double smoothing and Holt's linear trend, the *smoothing constant(s)* determines how fast the weights of the series decays. The values may be chosen either subjectively or objectively. Values of a smoothing constant near one put almost all weight on the most recent observations. Values of a smoothing constant near zero allow the distant past observations to have a large influence.

When selecting the smoothing constant *subjectively*, you use your own experience with this, and similar, series. Also, specifying the smoothing constant yourself lets you tune the forecast to your own beliefs about the future of the series. If you believe that the mechanism generating the series has recently gone through some fundamental changes, use a smoothing constant value of 0.9 which will cause distant observations to be ignored. If, however, you think the series is fairly stable and only going through random fluctuations, use a value of 0.1.

To select the value of the smoothing constant(s) *objectively*, you search for values that are best in some sense. Our program searches for that values that minimizes the size of the combined forecast errors of the currently available series. Three methods of summarizing the amount of error in the forecasts are available: the mean square error (MSE), the mean absolute error (MAE), and the mean absolute percent error (MAPE). The forecast error is the difference between the forecast of the current period made at the last period and the value of the series at the current period. This is written as

$$e_t = X_t - F_{t-1}$$

Using this formulation, we can define the three error-size criterion as follows:

$$MSE = \frac{1}{n}\sum e_t^2$$

$$MAE = \frac{1}{n}\sum \left| e_t \right|$$

$$MAPE = \frac{100}{n}\sum \left| \frac{e_t}{X_t} \right|$$

To find the value of the smoothing constants objectively, we select one of these criterion and search for those values of α and β that minimize this function. The program conducts a search for the appropriate values using an efficient grid-searching algorithm.

Initial Values

Both double smoothing and Holt's linear trend require initialization since the forecast for period one requires the forecast at period zero, which we do not, by definition, have. Several methods have been proposed for generating starting values. We have adopted the backcasting method which is currently considered to be one of the best methods. Backcasting is simply reversing the series so that we forecast into the past instead of into the future. This produces the required starting values for the slope and intercept. Once we have done this, we can then switch the series back and apply the algorithm in the regular manor.

Relationship to ARIMA Method

It can be shown that both double exponential smoothing and Holt's linear trend technique are equivalent to the ARIMA(0,2,2) model (see Kendall and Ord (1990) page 133). This is why backcasting is recommended for initial values.

Assumptions and Limitations

These algorithms are useful for forecasting non-seasonal time series with (local or global) trend.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by these algorithms. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable(s) on which to run the analysis.

Time Series Variables

Time Series Variable(s)

Specify the variable(s) on which to run the analysis. A separate analysis will be conducted for each variable listed.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable. The forecasts are converted back to there original metric before display.

Forecasting Options

Forecast Method

Select LS (least squares) linear trend, double exponential smoothing, or Holt's linear trend.

Number of Forecasts

This option specifies the number of forecasts to be generated.

Smoothing Constant Search Options

Search Method

This option specifies whether a search is conducted for the best values of the smoothing constants and what the criterion for the search will be.

- **Specified Value**

  No search is conducted. The values of the smoothing constants that are given in the next options are used.

- **Search on MSE**

  A search is conducted to find the values of the smoothing constants that minimize MSE.

- **Search on MAE**

  A search is conducted to find the values of the smoothing constants that minimize MAE.

- **Search on MAPE**

  A search is conducted to find the values of the smoothing constants that minimize MAPE.

Smoothing Constant Search Options – Pre-Specified Smoothing Constants

Alpha Smoothing Constant

When the Search Method is set to Specified Value, this option specifies the value of alpha used in double exponential smoothing and Holt's linear trend. The limits of this value are zero and one. Usually, a value between 0.1 and 0.3 are used. As the value gets closer to one, more and more weight is given to recent observations.

Beta Smoothing Constant

When the Search Method is set to Specified Value, this option specifies the value of beta used in Holt's linear trend. The limits of this value are zero and one. Usually, a value between 0.1 and 0.3 are used. As the value gets closer to one, more and more weight is given to recent observations.

Reports Tab

The following options control which reports are displayed.

Select Reports

Summary Report

This option specifies whether the indicated report is displayed.

Forecast Report

This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Select Plots

Forecast Plot - Residual 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 only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Page Title

Specify a title to be shown at the top of the reports.

Forecast Plot Tab

A plot of the data and forecast over time may be displayed.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forecast 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

This option controls the attributes of the plotting 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.

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.

Residual Plot Tab

This section controls the residual plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Residual 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

This option controls the attributes of the plotting 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.

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.

Storage Tab

The forecasts and residuals may be stored on the current database for further analysis. These options let you designate which statistics (if any) should be stored by designating which variables should receive the statistics.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Forecasts

The forecasts are stored in this variable.

Residuals

The residuals are stored in this variable.

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 – Trend Exponential Smoothing

This section presents an example of how to generate a forecast of a series using Holt's linear trend. The data in the INTEL database gives price and volume data for Intel stock during August, 1995. We will forecast values for daily volumes. These values are contained in the variable Intel_Volume.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Exponential Smoothing – Trend window.

1 Open the Intel dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Intel.S0**.
- Click **Open**.

2 Open the Exponential Smoothing – Trend window.

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Exponential Smoothing** – **Trend**. The Exponential Smoothing – Trend 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 Exponential Smoothing – Trend window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **Intel_Volume** from the list of variables and then click **Ok**.
- Select **Holt's Linear Trend** in the **Forecast Method** list box.

4 Specify the reports.

- On the Exponential Smoothing – Trend window, select the **Reports tab**.
- Select **Data and Forecasts** in the **Forecast Report** list box.

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).

Forecast Summary Section

```
Forecast Summary Section

Variable                    Intel_Volume
Number of Rows              20
Mean                        10974.54
Pseudo R-Squared            0.000000
Mean Square Error           1.80109E+07
Mean |Error|                3229.513
Mean |Percent Error|        28.64767

Forecast Method             Holt's Linear Trend
Search Iterations           59
Search Criterion            Mean Square Error
Alpha                       0.4157034
Beta                        0.1182474
Intercept (A)               9277.523
Slope (B)                   210.8949
```

This report summarizes the forecast equation.

Variable

The name of the variable for which the forecasts are generated.

Mean

The mean of the variable across all time periods.

Pseudo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Mean Square Error

The average squared residual (MSE) is a measure of how closely the forecasts track the actual data. The statistic is popular because it shows up in analysis of variance tables. However, because of the squaring, it tends to exaggerate the influence of outliers (points that do not follow the regular pattern).

Mean |Error|

The average absolute residual (MAE) is a measure of how closely the forecasts track the actual data without the squaring.

Mean |Percent Error|

The average percent absolute residual (MAPE) is a measure of how closely the forecasts track the actual data put on a percentage basis.

Forecast Method

This line shows which of the three possible trend forecasting algorithms was selected.

Search Iterations

This line shows how many iterations were needed to find the best value(s) for the smoothing constant(s).

Search Criterion

If a search was made to find the best values of the smoothing constants, this row gives the criterion used during the search.

Alpha

The value of the smoothing constant alpha that was used to generate the forecasts.

Beta

The value of the smoothing constant beta that was used to generate the forecasts.

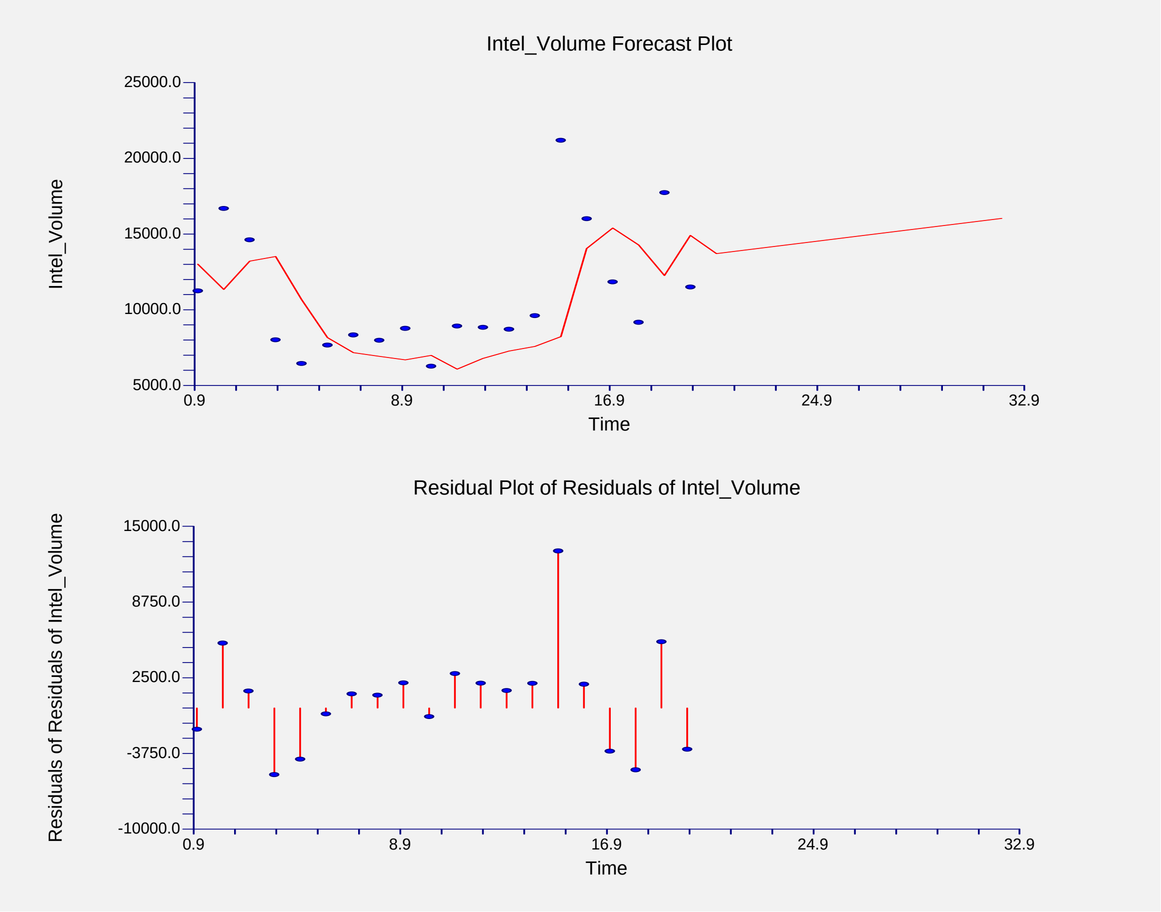
Intercept (A)

The value of the y-intercept for time period one! Hence, to forecast for time period 21 (the next period after the current period) we would use 9277.523 + 21(210.8949) = 13706.32.

Slope (B)

The value of the slope.

Forecast and Residuals Plots



Forecast Plot

The forecast plot lets you analyze how closely the forecasts track the data. The plot also shows the forecasts at the end of the data series.

Residual Plot

This plot lets you analyze the residuals themselves. You are looking for patterns, outliers, or any other information that may help you improve the forecasting model. The first thing to compare is the scale of the Residual Plot versus the scale of the Forecast Plot. If your forecasting is working well, the vertical scale of the Residual Plot will be much less than the scale of the Forecast Plot.

Forecasts Section

```
Row     Forecast        Actual
No.     Intel_Volume    Intel_Volume    Residuals
1       13003.77        11242.2         -1761.574
2       11334.51        16689.9         5355.391
3       13210.71        14613.3         1402.594
4       13512.66        8009            -5503.658
5       10673.12        6441.8          -4231.32
6       8154.504        7664.5          -490.0037
7       7167.079        8330.3          1163.221
8       6924.084        7983            1058.916
9       6689.781        8767.1          2077.319
10      6980.944        6266.4          -714.5444
11      6076.396        8915.3          2838.904
12      6788.578        8833            2044.422
13      7270.985        8709.7          1438.714
14      7572.32         9603            2030.68
15      8219.556        21185.2         12965.64
16      14049.83        16006.5         1956.668
17      15399.82        11832.4         -3567.42
18      14278.07        9168.1          -5109.966
19      12263.89        17729.3         5465.414
20      14914.58        11500.7         -3413.885
21      13706.32
22      13917.21
23      14128.11
24      14339
25      14549.9
26      14760.79
27      14971.68
```

This section shows the values of the forecasts, the actual values, and the residuals.

Chapter 467

Exponential Smoothing – Trend & Seasonal

Introduction

This module forecasts seasonal series with upward or downward trends using the Holt-Winters exponential smoothing algorithm. Two seasonal adjustment techniques are available: additive and multiplicative.

Additive Seasonality

Given observations $X_1, X_2, \cdots, X_t$ of a time series, the Holt-Winters additive seasonality algorithm computes an evolving trend equation with a seasonal adjustment that is additive. *Additive* means that the amount of the adjustment is constant for all levels (average value) of the series.

The forecasting algorithm makes use of the following formulas:

$$a_t = \alpha\left(X_t - F_{t-s}\right) + (1-\alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

$$F_t = \gamma\left(X_t - a_t\right) + (1-\gamma)F_{t-s}$$

Here α, β, and γ are smoothing constants which are between zero and one. Again, a_t gives the y-intercept (or level) at time *t*, while b_t is the slope at time *t.* The letter *s* represents the number of periods per year, so the quarterly data is represented by s = 4 and monthly data is represented by s = 12.

The forecast at time *T* for the value at time *T*+*k* is $a_T + b_T k + F_{[(T+k-1)/s]+1}$. Here *[(T+k-1)/s]* is means the remainder after dividing *T+k-1* by *s*. That is, this function gives the season (month or quarter) that the observation came from.

Multiplicative Seasonality

Given observations $X_1, X_2, \cdots, X_t$ of a time series, the Holt-Winters multiplicative seasonality algorithm computes an evolving trend equation with a seasonal adjustment that is multiplicative. *Multiplicative* means that the amount of the adjustment is varies with the level (average value) of the series. Note that the nature of most economic time series make the multiplicative model more popular than the additive model.

The forecasting algorithm makes use of the following formulas:

$$a_t = \alpha\left(X_t / F_{t-s}\right) + (1-\alpha)\left(a_{t-1} + b_{t-1}\right)$$

$$b_t = \beta\left(a_t - a_{t-1}\right) + (1-\beta)b_{t-1}$$

$$F_t = \gamma\left(X_t / a_t\right) + (1-\gamma)F_{t-s}$$

Here α, β, and γ are smoothing constants which are between zero and one. Again, a_t gives the y-intercept (or level) at time *t*, while b_t is the slope at time *t*. The letter *s* represents the number of periods per year, so the quarterly data is represented by s = 4 and monthly data is represented by s = 12.

The forecast at time *T* for the value at time *T*+*k* is $(a_T + b_T k)F_{[(T+k-1)/s]+1}$. Here *[(T+k-1)/s]* is means the remainder after dividing *T*+*k-1* by *s*. That is, this function gives the season (month or quarter) that the observation came from.

Smoothing Constants

Notice that the *smoothing constants* determines how fast the weights of the series decays. The values may be chosen either subjectively or objectively. Values of a smoothing constant near one put almost all weight on the most recent observations. Values of a smoothing constant near zero allow the distant past observations to have a large influence.

Note that α is associated with the level of the series, β is associated with the trend, and γ is associated with the seasonality factors.

When selecting the smoothing constant *subjectively*, you use your own experience with this, and similar, series. Also, specifying the smoothing constant yourself lets you tune the forecast to your own beliefs about the future of the series. If you believe that the mechanism generating the series has recently gone through some fundamental changes, use a smoothing constant value of 0.9 which will cause distant observations to be ignored. If, however, you think the series is fairly stable and only going through random fluctuations, use a value of 0.1.

To select the value of the smoothing constants *objectively*, you search for values that are best in some sense. Our program searches for that values that minimize the size of the combined forecast errors of the currently available series. Three methods of summarizing the amount of error in the forecasts are available: the mean square error (MSE), the mean absolute error (MAE), and the mean absolute percent error (MAPE). The forecast error is the difference between the forecast of the current period made at the last period and the value of the series at the current period. This is written as

$$e_t = X_t - F_{t-1}$$

Using this formulation, we can define the three error-size criterion as follows:

$$MSE = \frac{1}{n}\sum e_t^2$$

$$MAE = \frac{1}{n}\sum \left|e_t\right|$$

$$MAPE = \frac{100}{n}\sum \left|\frac{e_t}{X_t}\right|$$

To find the value of the smoothing constants objectively, we select one of these criterion and search for those values of α and β that minimize this function. The program conducts a search for the appropriate values using an efficient grid-searching algorithm.

Initial Values

Winters method requires initialization since the forecast for period one requires the forecast at period zero, which we do not, by definition, have. It also requires the seasonal adjustment factors. Several methods have been proposed for generating starting values. We have adopted the backcasting method which is currently considered to be one of the best methods. Backcasting is simply reversing the series so that we forecast into the past instead of into the future. This produces the required starting values for the slope, intercept, and seasonal factors. Once we have done this, we can switch the series back and apply the algorithm in the regular manner.

Relationship to ARIMA Method

The multiplicative seasonal adjustment model does not have an ARIMA counterpart, while the additive model does.

Assumptions and Limitations

These algorithms are useful for forecasting seasonal time series with (local or global) trend.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by these algorithms. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable(s) on which to run the analysis.

Time Series Variables

Time Series Variable(s)

Specify the variable(s) on which to run the analysis. A separate analysis will be conducted for each variable listed.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable. The forecasts are converted back to there original metric before display.

Forecasting Options

Number of Forecasts

This option specifies the number of forecasts to be generated.

Seasonal Model Options

Seasonal Adjustment

Select either the Additive or Multiplicative adjustment scheme.

Seasonality Options

Seasons

Specify the number of seasons per year in the series. Use ‘4’ for quarterly data or ‘12’ for monthly data.

First Season

Specify the first season of the series. This value is used to format the reports and plots. For example, if you have monthly data beginning with March, you would enter a ‘3’ here.

First Year

Specify the first year of the series. This value is used to format the reports and plots.

Smoothing Constant Search Options

Search Method

This option specifies whether a search is conducted for the best values of the smoothing constants and what the criterion for the search will be.

- **Specified Value**

  No search is conducted. The values of the smoothing constants given in the next options are used.

- **Search on MSE**

  A search is conducted to find the values of the smoothing constants that minimize MSE.

- **Search on MAE**

  A search is conducted to find the values of the smoothing constants that minimize MAE.

- **Search on MAPE**

  A search is conducted to find the values of the smoothing constants that minimize MAPE.

Smoothing Constant Search Options – Pre-Specified Smoothing Constants

Alpha Smoothing Constant

When the Search Method is set to Specified Value, this option specifies the value of alpha. Alpha is the smoothing constant for the level of the series. The limits of this value are zero and one. Usually, a value between 0.1 and 0.3 are used. As the value gets closer to one, more and more weight is given to recent observations.

Beta Smoothing Constant

When the Search Method is set to Specified Value, this option specifies the value of beta. Beta is the smoothing constant for the trend. The limits of this value are zero and one. Usually, a value between 0.1 and 0.3 are used. As the value gets closer to one, more and more weight is given to recent observations.

Gamma Smoothing Constant

When the Search Method is set to Specified Value, this option specifies the value of gamma. Gamma is the smoothing constant for the seasonal factors. The limits of this value are zero and one. Usually, a value between 0.1 and 0.3 are used. As the value gets closer to one, more and more weight is given to recent observations.

Reports Tab

The following options control which reports are displayed.

Select Reports

Summary Report

This option specifies whether the indicated report is displayed.

Forecast Report

This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Select Plots

Forecast Plot - Residual 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 only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Page Title

Specify a title to be shown at the top of the reports.

Forecast Plot Tab

This section controls the forecast plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forecast 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

This option controls the attributes of the plotting 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.

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.

Residual Plot Tab

This section controls the residual plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Residual 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

This option controls the attributes of the plotting 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.

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.

Storage Tab

The forecasts and residuals may be stored on the current database for further analysis. These options let you designate which statistics (if any) should be stored by designating which variables should receive the statistics.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Forecasts

The forecasts are stored in this variable.

Residuals

The residuals are stored in this variable.

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 – Trend & Seasonal Exponential Smoothing

This section presents an example of how to generate forecasts of a series using Winters multiplicative seasonal model. The data in the SALES database will be used. We will forecast the values of the Sales variable for the next twelve months.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Exponential Smoothing – Trend / Seasonal window.

1 Open the Sales dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sales.S0**.
- Click **Open**.

2 Open the Exponential Smoothing – Trend / Seasonal window.

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Exponential Smoothing - Trend/Seasonal**. The Exponential Smoothing – Trend / Seasonal 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 Exponential Smoothing – Trend / Seasonal window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **Sales** from the list of variables and then click **Ok**.
- Enter **1970** in the **First Year** box.

4 Specify the reports.

- On the Exponential Smoothing – Trend / Seasonal window, select the **Reports tab**.
- Select **Data and Forecasts** in the **Forecast Report** list box.

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).

Forecast Summary Section

```
Forecast Summary Section

Variable                     Sales
Number of Rows               144
Mean                         174.2847
Pseudo R-Squared             0.974893
Mean Square Error            20.36166
Mean |Error|                 3.582509
Mean |Percent Error|         2.043473

Forecast Method              Winter's with multiplicative seasonal adjustment.
Search Iterations            146
Search Criterion             Mean Square Error
Alpha                        0.3120825
Beta                         2.528956E-02
Gamma                        0.4916648
Intercept (A)                117.1492
Slope (B)                    0.7311453
Season 1 Factor              0.9097279
Season 2 Factor              0.8559249
Season 3 Factor              0.964533
Season 4 Factor              0.9912127
Season 5 Factor              1.03415
Season 6 Factor              1.022021
Season 7 Factor              0.9964323
Season 8 Factor              1.001555
Season 9 Factor              0.9602281
Season 10 Factor             1.029593
Season 11 Factor             1.011403
Season 12 Factor             1.223219
```

This report summarizes the forecast equation.

Variable

The name of the variable for which the forecasts are generated.

Number of Rows

The number of rows that were in the series. This is provided to allow you to double-check that the correct series was used.

Mean

The mean of the variable across all time periods.

Pseudo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where SSE is the sum of square residuals and SST is the total sum of squares after correcting for the mean.

Mean Square Error

The average squared residual (MSE) is a measure of how closely the forecasts track the actual data. The statistic is popular because it shows up in analysis of variance tables. However, because of the squaring, it tends to exaggerate the influence of outliers (points that do not follow the regular pattern).

Mean |Error|

The average absolute residual (MAE) is a measure of how closely the forecasts track the actual data without the squaring.

Mean |Percent Error|

The average percent absolute residual (MAPE) is a measure of how closely the forecasts track the actual data put on a percentage basis.

Forecast Method

This line shows which of the two possible seasonal adjustment algorithms was selected.

Search Iterations

This line shows how many iterations were needed to find the best values for the smoothing constants.

Search Criterion

If a search was made to find the best values of the smoothing constants, this row gives the criterion used during the search.

Alpha

The value of the smoothing constant alpha that was used to generate the forecasts.

Beta

The value of the smoothing constant beta that was used to generate the forecasts.

Gamma

The value of the smoothing constant gamma that was used to generate the forecasts.

Intercept (A)

The value of the y-intercept for time period one!

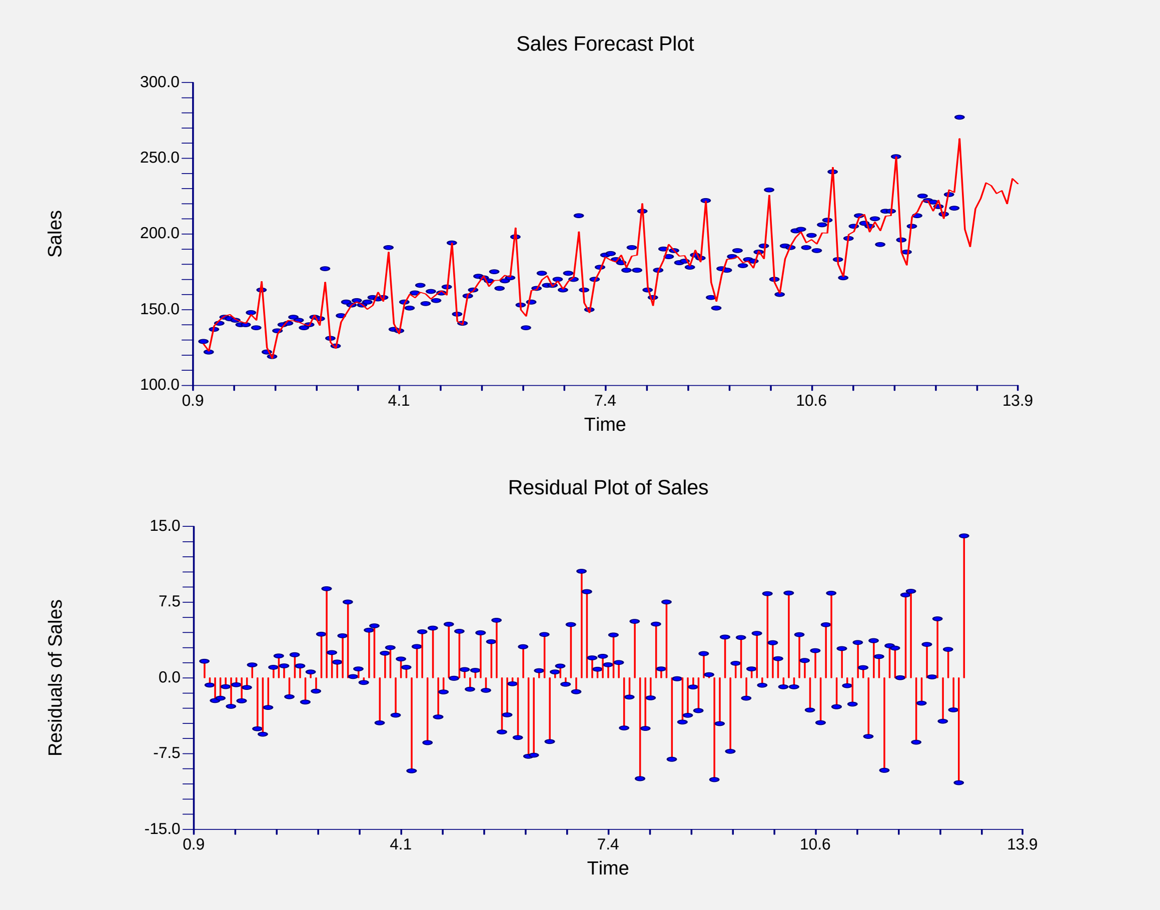
Slope (B)

The value of the slope.

Season (1-12) Factor

The values of the multiplicative seasonal factors.

Forecast and Residuals Plots



Forecast Plot

The forecast plot lets you analyze how closely the forecasts track the data. The plot also shows the forecasts at the end of the data series.

Residual Plot

This plot lets you analyze the residuals themselves. You are looking for patterns, outliers, or any other information that may help you improve the forecasting model. The first thing to compare is the scale of the Residual Plot versus the scale of the Forecast Plot. If your forecasting algorithm is working well, the vertical scale of the Residual Plot will be much less than the scale of the Forecast Plot.

Forecasts Section

```
Forecasts Section

Row                     Forecast        Actual
No.     Date            Sales           Sales           Residuals
1       1970 1          127.3592        129             1.640789
2       1970 2          122.7182        122             -0.7182454
3       1970 3          139.2607        137             -2.260717
4       1970 4          143.0193        141             -2.019351
5       1970 5          145.8727        145             -0.8727194
6       1970 6          146.8304        144             -2.830379
7       1970 7          143.6819        143             -0.6819006
8       1970 8          142.2842        140             -2.284242
9       1970 9          140.9541        140             -0.9540558
10      1970 10         146.7194        148             1.280589
11      1970 11         143.0544        138             -5.054361
12      1970 12         168.5834        163             -5.583423
.       .               .               .               .
.       .               .               .               .
.       .               .               .               .
140     1981 8          222.2978        218             -4.297828
141     1981 9          210.1906        213             2.809448
142     1981 10         229.1777        226             -3.177652
143     1981 11         227.3908        217             -10.39081
144     1981 12         262.9469        277             14.05311
145     1982 1          203.0197
146     1982 2          191.6385
147     1982 3          216.6607
148     1982 4          223.3784
149     1982 5          233.8107
150     1982 6          231.8159
151     1982 7          226.7403
152     1982 8          228.6382
153     1982 9          219.9061
154     1982 10         236.5445
155     1982 11         233.1048
156     1982 12         282.8179
```

This section shows the values of the forecasts, the dates, the actual values, and the residuals.

467-14 Exponential Smoothing – Trend & Seasonal

Chapter 468

Spectral Analysis

Introduction

This program calculates and displays the periodogram and spectrum of a time series. This is sometimes known as harmonic analysis or the frequency approach to time series analysis.

Suppose we believe that a time series, X_t, contains a periodic (cyclic) component. A natural model of the periodic component would be

$$X_t = R\cos(ft + d) + e_t$$

where

- R is the amplitude of variation. Normally, the cosine varies between -1 and 1. Hence, if *R* is 6, then the term would vary between -6 and 6. The impact of the amplitude is in the size (height or magnitude) of the wave. The length of the wave is not influenced by the amplitude.
- f is the frequency of periodic variation, measured in number of radians per unit time. This is the 'frequency' scale of the plots. If we divide 2π by *f*, we get the corresponding *wavelength*. This is the 'wavelength' scale of the plots. The impact of the frequency is to change the length of a cycle. As *f* increases, the length of the cycle decreases. A model with *f* = 2 would have a cycle length equal to one-half the cycle length of a model with *f* = 1.
- d is the phase. Changing the phase causes a shift in the beginning of the cycle.
- e_t is the random error (noise) of the series about the period component.
- t is the time period number. Usually, *t=1, 2, 3, ..., N*.

Since *cos(ft+d) = cos(ft) cos(d) - sin(ft) sin(d),* this model may be written in the alternative form

$$X_t = a\cos(ft) + b\sin(ft) + e_t$$

where *a = R cos(d)* and *b = -R sin(d).*

This model is a multiple regression model with two independent variables. In this case, the independent variables are *X1 = cos(ft)* and *X2 = sin(ft).* The regression coefficients are *B1 = a* and *B2 = b*. In practice, the variation in a time series may be modeled as the sum of several different individual waves occurring at different frequencies.

The generalization of this model to the sum of *k* frequencies may be written symbolically as

$$X_t = \sum_{j=1}^{k} R_j \cos\left(f_j t + d_j\right) + e_t$$

or, using the alternative form, as

$$X_t = \sum_{j=1}^{k} a_j \cos\left(f_j t\right) + \sum_{j=1}^{k} b_j \sin\left(f_j t\right) + e_t$$

Note that if the f_j were known constants, and we let $W_{tr} = \cos\left(f_r t\right)$ and $Z_{ts} = \sin\left(f_s t\right)$, then this could be rewritten in the usual multiple regression form:

$$X_t = \sum_{j=1}^{k} a_j W_{tj} + \sum_{j=1}^{k} b_j Z_{tj} + e_t$$

where the *a's* and the *b's* are the regression coefficients to be estimated. This is an example of a harmonic regression.

Fourier analysis is the study of approximating functions using the sum of sine and cosine terms. This sum is called the Fourier series representation of the function. Spectral analysis is identical to Fourier analysis except that instead of approximating a function, the sum of sine and cosine terms approximates a time series that includes a random component. Note that the coefficients (the *a's* and *b's*) may be estimated using multiple regression.

One question that arises is how to select the frequencies. The highest frequency that can be fit to the data is π. The lowest is one cycle for the whole length of series, which amounts to a frequency of $2\pi / N$ (*N* is the length of the series). Hence, one popular choice of frequencies is to select the *N/2* frequencies given by

$$f_k = 2\pi k / N, \quad (k = 1, 2, \cdots, N / 2)$$

The kth frequency is often referred to as the kth harmonic.

This set of frequencies is particularly popular when working by hand because it results in certain simplifications due to well-known trigonometric identities. However, there is nothing in nature that says that a series will follow these rather than some other set. That is why the program lets you specify a range of frequencies.

In the analysis of variance, we study the partitioning of the total variation (sum of squares) given by

$$SST = \sum_{t=1}^{N} \left(X_t - \overline{X}\right)^2$$

into the sum of squares for factor A, factor B, etc. Similarly, in spectral analysis we are interested in partitioning the total sum of squares into amounts associated with each frequency. It turns out that the sum of squares for a particular frequency, SS_k, is given by

$$SS_k = \frac{N}{2}\left(a_k^2 + b_k^2\right)$$

If we regard SS_k as the portion of the total sum of squares accounted for by frequencies in the range

$$f_k \pm \frac{\pi}{N},$$

we can draw a histogram so that the area of each bar is proportional SS_k. The height of the histogram would be

$$I\left(f_k\right)=\frac{N}{4\pi}\left(a_k^2+b_k^2\right)$$

The plot of *I(f)* versus *f* is called the *periodogram*.

This definition of the periodogram equates the total sum of squares to the area under the periodogram. *I(f)* may be calculated directly from the data as

$$I\left(f_k\right)=\frac{\left[\sum X_t\cos\left(2\pi kt/N\right)\right]^2+\left[\sum X_t\sin\left(2\pi kt/N\right)\right]^2}{N\pi}$$

The periodogram is sometimes calculated using the fast Fourier transform (FFT). This method is not used in this program for three reasons. First, the increase in speed of the FFT is not significant until *N* is greater than one thousand. For series of the length we normally anticipate for our users, the FFT would provide little speed improvement.

Second, when using the FFT, the length of the series (*N*) must be a power of 2 (2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, etc.). If *N* is not a power of 2, then enough zeros must be added to bring the length of the series to the next power of 2. Suppose the length of a particular series was 260. You would need to add 252 zeros to bring the length to 512. This could dramatically distort your results. (FFT users use various “windows” or “filters” to remove the effect of these zeros. Since we do not pad with zeros, we do not need these filters.)

Third, we can calculate the periodogram for any set of frequencies, not just the set given above. This is very useful when you want to investigate a particular range of frequencies.

The sample periodogram has been shown to have some poor statistical properties. Recently, techniques for spectral analysis have improved on the periodogram by smoothing it. The smoothed periodogram is an estimate of the *power spectral density* or simply the *spectral density* of the series. The smoothing used in this program is simply an *m-term* moving average of the periodogram. The value of *m* is specified as the Smoothing Length option. Practitioners suggest that a value of *m* near *N/40* is reasonable. A large value of *m* may make the graph too smooth while a value too small may include spurious peaks.

Spectral analysis offers an interesting addition to other methods of time series analysis. For those who wish to find more out about it, we strongly recommend the book by C. Chatfield (1984). It offers a thorough, readable treatment of a difficult, but useful, subject.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by this algorithm. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable on which to run the analysis.

Time Series Variable

Time Series Variable

Specify the variable on which to run the analysis.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable.

Data Adjustment Options

Remove Mean

Checking this option indicates that the series average should be subtracted from the data. This is almost always done.

Remove Trend

Checking this option indicates that the least squares trend line should be subtracted from the data. This is sometimes done, although differencing is usually used to remove trends instead.

Regular Differencing

This option lets you designate whether the original series, the first differences, or the second differences are analyzed. The first difference series, *W*, is calculated using the formula:

$$W_t = X_t - X_{t-1}$$

which may be written using the backshift operator, B, as:

$$W_t = (1 - B) X_t$$

The second difference series, *Z*, is the first difference of the *W* series. The formula is:

$$Z_t = W_t - W_{t-1}$$

which may be written using the backshift operator, B, as:

$$Z_t = (1 - B)^2 X_t$$

Seasonal Differencing

This option lets you designate whether the original series, the first seasonal differences, or the second seasonal differences are analyzed. Assuming the number of seasons is *s*, the first seasonal difference series, *W*, is calculated using the formula:

$$W_t = X_t - X_{t-s}$$

which may be written using the backshift operator, B, as:

$$W_t = \left(1 - B^s\right) X_t$$

The second seasonal difference series, *Z*, is the first seasonal difference of the *W* series. The formula is:

$$Z_t = W_t - W_{t-s}$$

which may be written using the backshift operator, B, as:

$$Z_t = \left(1 - B^s\right)^2 X_t$$

Seasonality Options

Seasons

Specify the number of seasons, *s*, in the series. Use '4' for quarterly data or '12' for monthly data. Note that this option is used only when seasonal differencing is used.

Reports Tab

The following options control which reports are displayed.

Select Reports

Fourier Report

This option specifies whether the indicated report is displayed.

Select Plots

Data Plot - Spectrum

Each of these options specifies whether the indicated plot is displayed.

Periodogram / Spectrum Calculation Options

Number of Frequencies

Specify the number of frequencies that are calculated and displayed. This controls the resolution of the periodogram and spectrum. The frequencies are equi-spaced between the minimum and maximum wavelengths.

Smoothing Length

The spectral density function is a moving average of the periodogram. This option specifies the value of *m*, the number of periodogram terms averaged.

Minimum Wavelength

The minimum wavelength value to be used in calculating and displaying the periodogram and spectral density.

Maximum Wavelength

The maximum wavelength value to be used in calculating and displaying the periodogram and spectral density. The maximum value possible is *N*, the sample size.

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.

Data Plot Tab

A plot of the data over time may be displayed. This panel controls the appearance of this plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Data 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

This option controls the attributes of the plotting 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.

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.

Periodogram / Spectrum Plot Tab

This section controls the appearance of the periodogram and spectrum plots.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis 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

This option controls the attributes of the plotting 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.

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.

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 – Spectral Analysis

This section presents an example of how to do a spectral analysis of a time series. The Spots variable in the SUNSPOT database will be used.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Spectral Analysis window.

1 **Open the Sunspot dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sunspot.S0**.
- Click **Open**.

2 **Open the Spectral Analysis window.**

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Spectral Analysis**. The Spectral Analysis procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

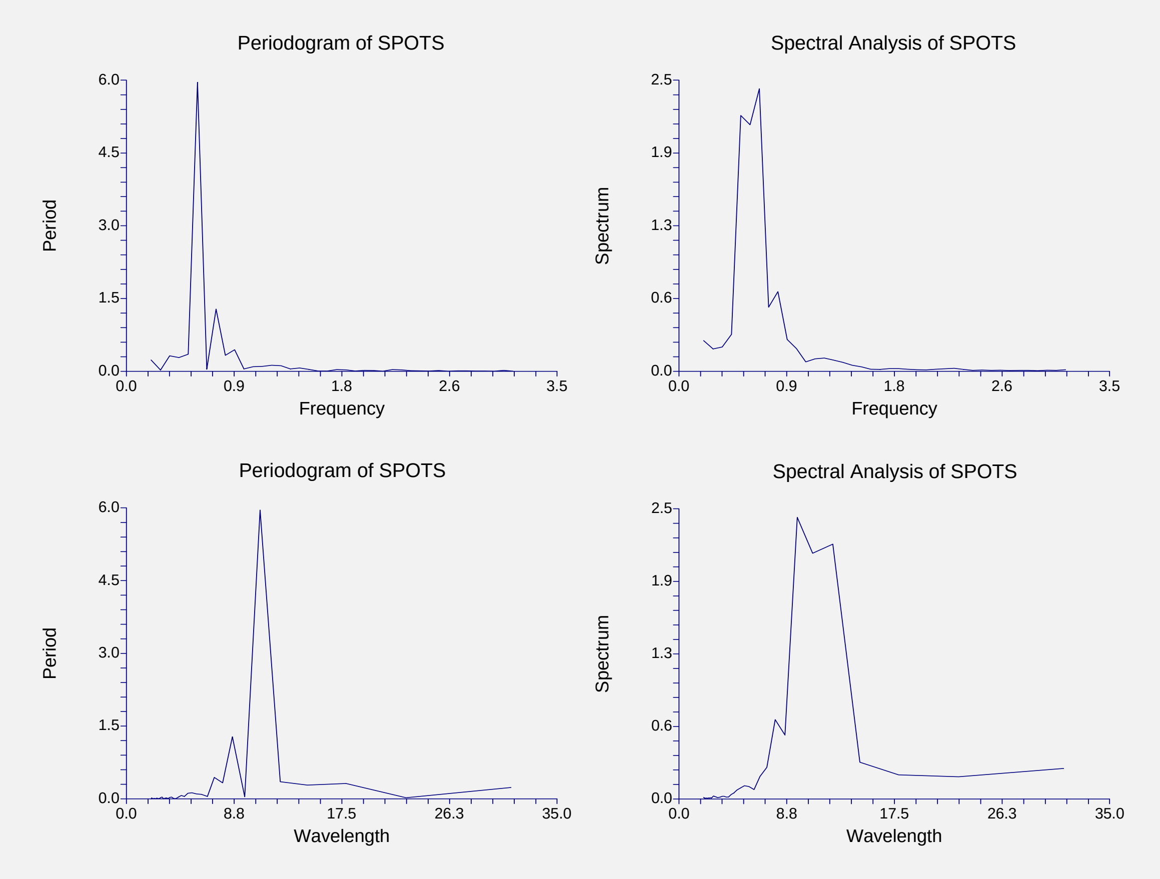
3 **Specify the variables.**

- On the Spectral Analysis window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **SPOTS** from the list of variables and then click **Ok**.

4 **Run the procedure.**

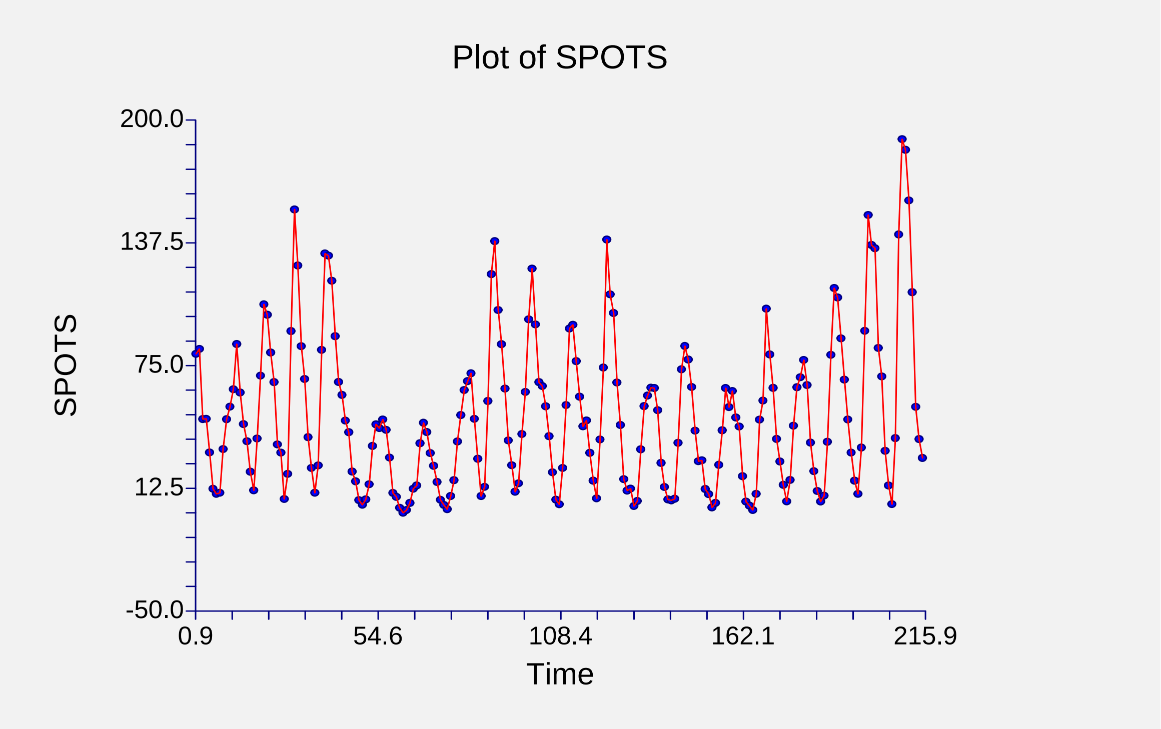
- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

Fourier Plot Section



This section displays the periodogram and the spectrum the top two plots are in the frequency scale. The bottom two plots are in the wavelength scale. Remember that the wavelength is in terms of the number of observations.

Data Plot Section



This section displays a plot of the data values.

Fourier Analysis Section

```
Fourier Analysis of SPOTS (0,0,12,1,0)
Frequency     Wavelength    Period           Cosine(a's)   Sine(b's)    Sprectrum
0.2010619     31.25         3200767          -75.89938     -425.8151    3590384
0.2764601     22.72727      324968.6         -13.48533     137.1568     2618775
0.3518584     17.85714      4330590          92.67065      -494.4972    2837492
0.4272566     14.70588      3856917          33.71187      -473.5963    4333876
0.5026549     12.5          4814120          298.3864      -438.5685    2.997943E+07
.             .             .                .             .            .
.             .             .                .             .            .
.             .             .                .             .            .
```

This section shows the values of the various components of the spectral analysis. The numbers in parentheses, (d,D,s,M,T), are defined as follows:

d is the regular differencing order.

D is the seasonal differencing order.

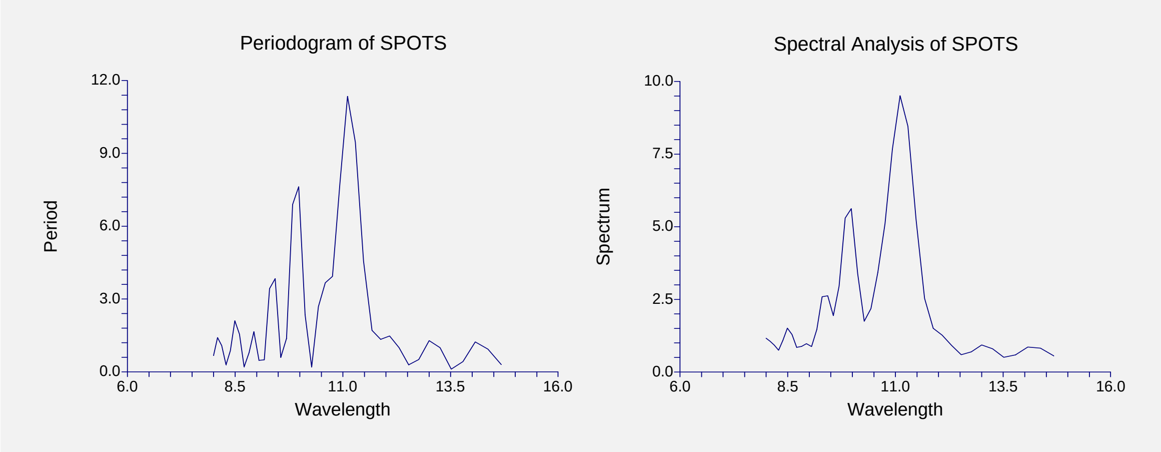
s is the number of seasons (ignored if D is 0).

M is 1 if the mean is subtracted, 0 otherwise.

T is 1 if the trend is subtracted, 0 otherwise.

Fourier Plot Section

To complete this example, we rerun the analysis with the minimum wavelength set to 8 and the maximum wavelength set to 15. This appears to be portion of the periodogram and spectrum that show the most promise. Doing this produces the following wavelength plots.



Now we can see the famous sunspot cycle of just over eleven years.

Chapter 469

Decomposition Forecasting

Introduction

Classical time series decomposition separates a time series into five components: mean, long-range trend, seasonality, cycle, and randomness. The decomposition model is

Value = (Mean) x (Trend) x (Seasonality) x (Cycle) x (Random).

Note that this model is multiplicative rather than additive. Although additive models are more popular in other areas of statistics, forecasters have found that the multiplicative model fits a wider range of forecasting situations.

Decomposition is popular among forecasters because it is easy to understand (and explain to others). While complex ARIMA models are often popular among statisticians, they are not as well accepted among forecasting practitioners. For seasonal (monthly, weekly, or quarterly) data, decomposition methods are often as accurate as the ARIMA methods and they provide additional information about the trend and cycle which may not be available in ARIMA methods.

Decomposition has one disadvantage: the cycle component must be input by the forecaster since it is not estimated by the algorithm. You can get around this by ignoring the cycle, or by assuming a constant value. Some forecasters consider this a strength because it allows the forecaster to enter information about the current business cycle into the forecast.

Decomposition Method

The basic decomposition method consists of estimating the five components of the model

$$X_t = UT_tC_tS_tR_t$$

where

X_t denotes the series or, optionally, log of series.

U denotes the mean of the series.

T_t denotes the linear trend.

C_t denotes cycle.

S_t denotes season.

R_t denotes random error.

t denotes the time period.

We will now take you through the steps used by the program to perform a decomposition of a time series. Most of this information is from Makridakis (1978), chapter 15.

Step 1 – Remove the Mean

The first step is to remove the mean by dividing each individual value by the series mean. This creates a new series that has values near one. This step is represented symbolically as

$$Y_t = X_t / U$$

If the absolute value of the mean of the series is less than 0.0000001, no division takes place.

Step 2 – Calculate a Moving Average

The next step calculates an L-step moving average centered at the time period, *t*, where *L* is the length of the seasonality (e.g., *L* would be 12 for a monthly series or 4 for quarterly series). Since the moving average gives the mean of a year's data, the seasonality factor is removed. Usually, the averaging removes the randomness component as well. Symbolically, this step is represented as

$$M_t = \sum Y_t$$

where for odd *L*, the summation runs from *t-[L/2]* to *t+[L/2]*. The symbols *[x]* mean take the integer part of *x*. Hence [6.43]=6 and [11/2]=5. Notice that this summation range centers the moving average at *t*.

For even *L*, the values usually found in practice (2, 4 and 12), it is a little more difficult to center the moving average on the time period *t*. For example, the average of the first 12 terms of a series would be centered at 6.5 rather than 6. To center the average right on 7, we must compute the moving average centered at 6.5 and at 7.5 and then average these. The resulting double moving average is centered at the desired value of 7.

Another complexity that must be dealt with is what to do at the ends of the series. Because the average is centered, the first and last *L/2* averages cannot be computed (because of the lack of data). Many different end-effect techniques have been proposed.

Our end-effect strategy can best be explained by considering an example. Suppose we have a monthly series that runs from January of 1980 to December of 1988. To compute the moving average centered at January, 1980, we will need estimated data back through July, 1979. The estimate of July 1979 is obtained by subtracting the difference of July, 1980 and July, 1981 from July, 1980.

At the other end of the series we will need estimated values through June, 1989. To compute the estimated value for June, 1989, we add the difference of June, 1987 and June, 1988 to June, 1988.

This method of estimating end-effects preserves local trends in the series. However, it is especially sensitive to outliers. You should remember that strange patterns in the last *L/2* time periods may be from the end-effect calculation and not from a pattern in the series itself.

Step 3 – Calculate the Trend

The next step is to calculate and remove the trend component of the series. This calculation is made on the moving averages, M_t, rather than on the Y_t series. A least squares fit is made of the of the model

$$M_t = a + bt + e_t$$

where

a is the intercept.

b is the slope.

e_t is the residual or lack-of-linear-fit.

The linear portion of the above model is used to define the trend. That is, we use

$$T_t = a + bt$$

Note that because of the problems of end-effects, the first and last *L/2* terms are omitted in the trend calculation.

Step 4 – Calculate the Cycle

The cycle term is found by dividing the moving average by the computed trend. Symbolically, this is

$$C_t = \frac{M_t}{T_t}$$

Step 5 – Calculate the Seasonality

The seasonality is computed by dividing the *Y* series by the moving averages. Symbolically, this is

$$K_t = \frac{Y_t}{M_t}$$

Note that the *K* series is composed of both the seasonality and the randomness. To calculate the seasonal component for each season, we simple average all like seasons. That is, the average of all Januarys is computed, the average of all Februarys gives the seasonal value for February, and so on. Mathematically, this is stated as

$$S_g = \sum K_t$$

where the summation is over all *t* in which the season is *g*.

Step 6 – Calculate the Randomness

The final step is to calculate the randomness component. This is accomplished by dividing the *K* series by S_i where the values of $S_1, S_2, \cdots, S_g$ are repeated as needed. This is represented mathematically as follows

$$R_t = \frac{K_t}{S_t}$$

Creating Forecasts

Once the series decomposition is complete, forecasts may be generated fairly easily. The trend component is calculated using

$$T_t = a + bt\ ,$$

the seasonal factor is read from

$$S_g = \sum K_t\ ,$$

the cycle factor is input by hand, and the random factor is assumed to be one. If the series was transformed using the log transformation, the forecasts are transformed back using the appropriate inverse function.

Assumptions and Limitations

These algorithms are useful for forecasting seasonal time series with (local or global) trend.

Data Structure

The data are entered in a single variable.

This section describes the options available in this procedure.

Missing Values

Missing values are not tolerated by this algorithm. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable(s) on which to run the analysis.

Time Series Variable

Time Series Variable

Specify the variable(s) on which to run the analysis. A separate analysis will be conducted for each variable listed.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable. The forecasts are converted back to there original metric before display.

Cycle-Input Variable

Cycle-Input Variable

Specifies the name of a variable containing estimated values of the cycle component. If this option is left blank, a value of one is used for all future cycle components. This ignores the cycle in the forecasts.

The program ignores the first *n* rows (where *n* is the number of rows in the original series) and begins reading the cycle ratios in the rows that match the forecast rows. For example, suppose you have a series with 48 rows over data and you want to forecast the next 12 rows. The first 48 rows of this variable are ignored. The 49th through 60th rows are used to provide the cyclical component.

Note that since the forecasting model is multiplicative, if you enter a '1' for the cycle, the forecast will not be changed. If you feel the cycle influence will be a 5% increase, you would enter '1.05' for this time period.

Use Cycle

This option specifies whether a cycle component is found or ignored.

Forecasting Options

Number of Forecasts

This option specifies the number of forecasts to be generated.

Seasonality Options

Seasons

Specify the number of seasons in the series. Use '4' for quarterly data or '12' for monthly data.

First Season

Specify the first season of the series. This value is used to format the reports and plots. For example, if you have monthly data beginning with March, you would enter a '3' here.

First Year

Specify the first year of the series. This value is used to format the reports and plots.

Reports Tab

The following options control which reports are displayed.

Select Reports

Summary Report

This option specifies whether the indicated report is displayed.

Forecast Report

This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Select Plots

Forecast Plot- Decomposition Ratio Plots

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 only.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Page Title

Specify a title to be shown at the top of the reports.

Forecast Plot Tab

This section controls the forecast plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forecast 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

This option controls the attributes of the plotting 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.

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.

Decomp Ratio Plots Tab

This section controls the decomposition ratio plots.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Ratio 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

This option controls the attributes of the plotting 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.

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.

Storage Tab

Several statistics, including the forecasts 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.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Forecasts

The forecasts are stored in this variable.

Residuals

The residuals are stored in this variable.

Trend Ratios

The trend ratios, the T_t series, are stored in this variable.

Cycle Ratios

The cycle ratios, the C_t series, are stored in this variable.

Season Ratios

The season ratios, the S_t series, are stored in this variable.

Error Ratios

The error, or random, ratios, the R_t series, are stored in this variable.

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 – Decompositions Forecasting

This section presents an example of how to generate forecasts of a series using the time series decomposition forecasting method. The data in the SALES database will be used. We will forecast the values of the Sales variable for the next twelve months.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Decomposition Forecasting window.

1 Open the Sales dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Sales.S0**.
- Click **Open**.

2 Open the Decomposition Forecasting window.

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Decomposition Forecasting**. The Decomposition Forecasting 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 Decomposition Forecasting window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **Sales** from the list of variables and then click **Ok**.
- Enter **1970** in the **First Year** 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).

Forecast Summary Section

```
Forecast Summary Section

Forecast                  (Mean) x (Trend) x (Cycle) x (Season)
Variable                  Sales
Number of Rows            144
Mean                      174.2847
Pseudo R-Squared          0.9872795
Forecast Std. Error       3.211923
Trend Equation            Trend = (0.762387) + (0.003224) * (Time Period Number)
Number of Seasons         12
First Year                1970
First Season              1

Seasonal Component Ratios
No.  Ratio           No.  Ratio           No.  Ratio           No.  Ratio
1    0.901933        2    0.855207        3    0.971591        4    0.998289
5    1.030011        6    1.028496        7    0.997262        8    1.005337
9    0.975528        10   1.024248        11   1.007887        12   1.205982
```

This report summarizes the forecast equation.

Variable

The name of the variable for which the forecasts are generated.

Number of Rows

The number of rows that were in the series. This is provided to allow you to double-check that the correct series was used.

Mean

The mean of the variable across all time periods.

Pseudo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Forecast Std. Error

The estimated standard deviation of the forecast errors (the difference between the actual and predicted). This value is calculated by squaring and summing all of the forecast errors, dividing by the number of observations, and taking the square root.

Trend Equation

The equation used to predict the trend. The equation is

$$\text{Trend} = a + bt$$

where

a is the intercept.

b is the slope.

t is the time period.

Note that the trend value obtained from this equation will be a ratio type value that will be multiplied by the mean to obtain the actual forecast.

Number of Seasons

The number of rows per year. For example, monthly data would have a value of 12.

First Year

The value of the first year of the series.

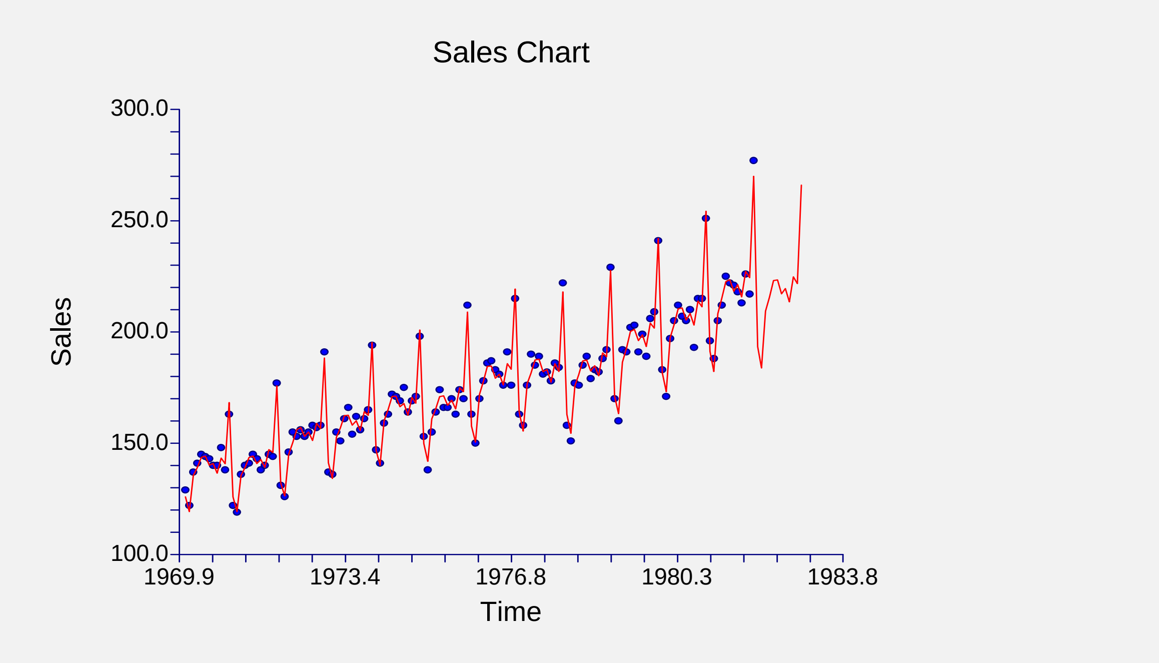
First Season

The value of the season of the first observation.

Season Component Ratios

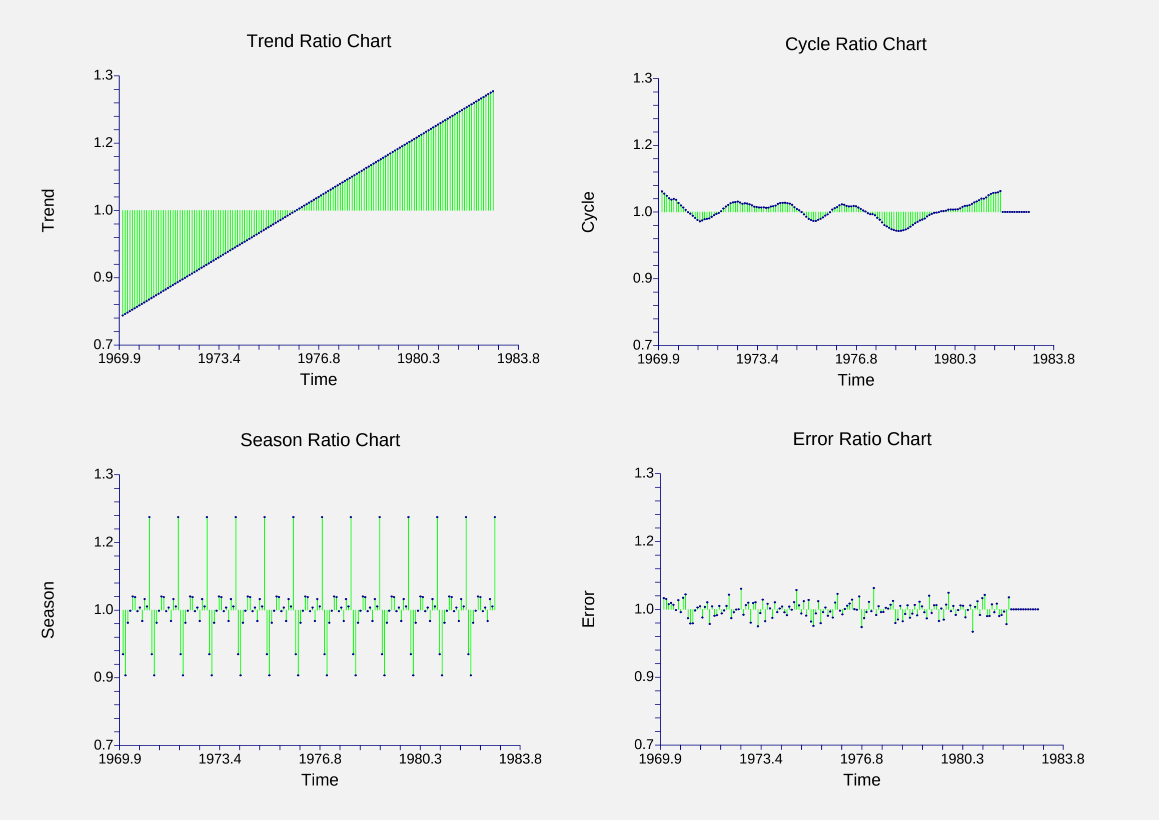
The ratios used to adjust for each season (month or quarter). For example, the last ratio in this example is 1.205982. This indicates that the December correction factor is a 20.5982% increase in the forecast.

Data and Forecast Plot



The data plot lets you analyze how closely the forecasts track the data. The plot also shows the forecasts at the end of the data series.

Ratio Plots



Ratio Plots

These plots let you see the various components of the forecast. Each of these plots is centered at one since this is the value that will leave the forecast unchanged. By studying these plots, you can see which factors influence the forecasts the most.

Forecasts Section

```
Forecasts Section

Row    Year          Forecast      Actual                    Trend      Cycle       Season     Error
No.    Season        Sales         Sales      Residual       Factor     Factor      Factor     Factor
1      1970  1       125.8949      129        3.105124       0.7656     1.0461      0.9019     1.0247
2      1970  2       119.3013      122        2.698687       0.7688     1.0411      0.8552     1.0226
3      1970  3       135.456       137        1.544002       0.7721     1.0361      0.9716     1.0114
.      .             .             .          .              .          .           .          .
.      .             .             .          .              .          .           .          .
144    1981  12      269.8886      277        7.111356       1.2266     1.0468      1.2060     1.0263
145    1982  1       193.3271                                1.2299     1.0000      0.9019     1.0000
146    1982  2       183.7919                                1.2331     1.0000      0.8552     1.0000
147    1982  3       209.3499                                1.2363     1.0000      0.9716     1.0000
.      .             .             .          .              .          .           .          .
.      .             .             .          .              .          .           .          .
```

This section shows the values of the forecasts, the dates, the actual values, the residuals, and the forecast ratios.

Note that the forecasted cycle ratios are all equal to one. This is because we did not supply cycle values to be used. If we had, they would have shown up here.

Chapter 470

The Box-Jenkins Method

Introduction

Box - Jenkins Analysis refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models. The method is appropriate for time series of medium to long length (at least 50 observations).

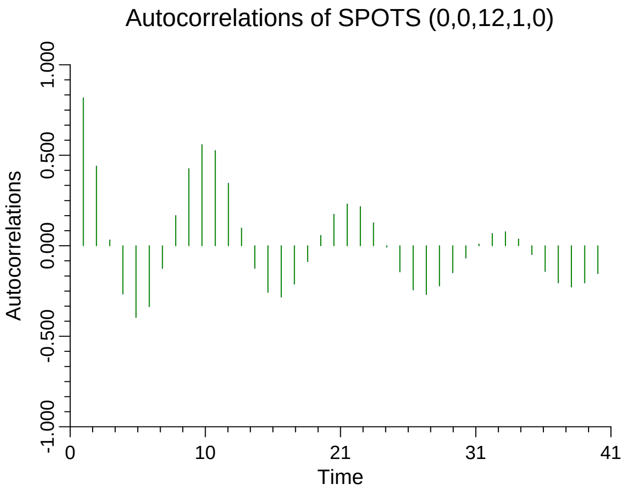
In this chapter we will present an overview of the Box-Jenkins method, concentrating on the how-to parts rather than on the theory. Most of what is presented here is summarized from the landmark book on time series analysis written by George Box and Gwilym Jenkins (1976).

A time series is a set of values observed sequentially through time. The series may be denoted by $X_1, X_2, \cdots, X_t$, where *t* refers to the time period and *X* refers to the value. If the X's are exactly determined by a mathematical formula, the series is said to be *deterministic*. If future values can be described only by their probability distribution, the series is said to be a *statistical* or *stochastic* process.

A special class of stochastic processes is a *stationary stochastic process*. A statistical process is stationary if the probability distribution is the same for all starting values of *t*. This implies that the mean and variance are constant for all values of *t*. A series that exhibits a simple trend is not stationary because the values of the series depend on *t*. A stationary stochastic process is completely defined by its mean, variance, and autocorrelation function. One of the steps in the Box - Jenkins method is to transform a non-stationary series into a stationary one.

Autocorrelation Function

The stationary assumption allows us to make simple statements about the correlation between two successive values, X_t and X_{t+k}. This correlation is called the *autocorrelation of lag k* of the series. The autocorrelation function displays the autocorrelation on the vertical axis for successive values of k on the horizontal axis. The following figure shows the autocorrelation function of the sunspot data.



Since a stationary series is completely specified by its mean, variance, and autocorrelation function, one of the major (and most subjective) tasks in Box-Jenkins analysis is to identify an appropriate model from the sample autocorrelation function. Although the sample autocorrelations contains random fluctuations, for moderate sample sizes they are fairly accurate in signaling the order of the ARIMA model.

The ARMA Model

The ARMA (autoregressive, moving average) model is defined as follows:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

where the $\phi' s$ (phis) are the autoregressive parameters to be estimated, the $\theta' s$ (thetas) are the moving average parameters to be estimated, the *X's* are the original series, and the *a's* are a series of unknown random errors (or residuals) which are assumed to follow the normal probability distribution.

Box-Jenkins use the backshift operator to make writing these models easier. The backshift operator, *B*, has the effect of changing time period *t* to time period *t-1*. Thus $BX_t = X_{t-1}$ and $B^2 X_t = X_{t-2}$. Using this backshift notation, the above model may be rewritten as:

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) X_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right) a_t$$

This may be abbreviated even further by writing:

$$\phi_p(B) X_t = \theta_q(B) a_t$$

where

$$\phi_p(B) = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$$

and

$$\theta_q(B) = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$$

These formulas show that the operators $\phi_p(B)$ and $\theta_q(B)$ are polynomials in *B* of orders *p* and *q* respectively. One of the benefits of writing models in this fashion is that we can see why several models may be equivalent.

For example, consider the model

$$X_t = 0.8X_{t-1} - 0.15X_{t-2} + a_t - 0.3a_{t-1}$$

This could be rewritten in the form of (8.3) as:

$$\left(1 - 0.8B + 0.15B^2\right)X_t = (1 - 0.3B)a_t$$

Notice that the polynomial on the left may be factored, so that we can rewrite the model as

$$(1 - 0.5B)(1 - 0.3B)X_t = (1 - 0.3B)a_t$$

Finally, canceling the *(1 - 0.3B)* from both sides leaves the simpler, but equivalent, model

$$(1 - 0.5B)X_t = a_t$$

or

$$X_t = 0.5X_{t-1} + a_t$$

Note that this is a much simpler model!

This type of model rearrangement is used by experienced Box-Jenkins forecasters to obtain the simplest models possible. The Theoretical ARIMA program displays the roots of the two polynomials, $\phi_p(B)$ and $\theta_q(B)$, so you can see possible model simplifications.

Nonstationary Models

Many time series encountered in practice exhibit nonstationary behavior. Usually, the nonstationarity is due to a trend, a change in the local mean, or seasonal variation. Since the Box-Jenkins methodology is for stationary models only, we have to make some adjustments before we can model these nonstationary series.

We use one of two methods for reducing a nonstationary series with trend to a stationary series (without trend):

1. Use the first differences of the series, $W_t = X_t - X_{t-1}$. Note that this can be rewritten as $W_t = (1 - B)X_t$. A more general form of this equation is:

   $$\phi_p(B)(1 - B)^d X_t = \theta_q(B)a_t$$

   where *d* is the order of differencing. This is known as the *ARIMA(p,d,q)* model.
2. Fit a least squares trend and fit the Box-Jenkins model to the residuals.

If the model exhibits an occasional change of mean, first differences will result in a stationary model.

For seasonal series, Box-Jenkins provided a modification to this equation that will be the subject of the next section.

Seasonal Time Series

To deal with series containing seasonal fluctuations, Box-Jenkins recommend the following general model:

$$\phi_p(B)\Phi_P(B)(1-B)^d\left(1-B^s\right)^D X_t = \theta_q(B)\,\Theta_Q\left(B^s\right)a_t$$

where d is the order of differencing, s is the number of seasons per year, and D is the order of seasonal differencing. The operator polynomials are

$$\phi_p(B)=\left(1-\phi_1 B-\cdots-\phi_p B^p\right)$$

$$\theta_q(B)=\left(1-\theta_1 B-\cdots-\theta_q B^q\right)$$

$$\Phi_P\left(B^s\right)=\left(1-\Phi_1 B^s-\cdots-\Phi_p B^{sp}\right)$$

$$\Theta_Q\left(B^s\right)=\left(1-\Theta_1 B^s-\cdots-\Theta_Q B^{sQ}\right)$$

Note that $\left(1-B^s\right)X_t = X_t - X_{t-s}$.

Box-Jenkins explain that the maximum value of *d, D, p, q, P*, and *Q* is two. Hence, these operator polynomials are usually simple expressions.

Partial Autocorrelation Function

We previously discussed the autocorrelation function, which gives the correlations between different lags of a series. The Partial Autocorrelation Function is a second function that expresses information useful in determining the order of an ARIMA model.

This function is constructed by calculating the partial correlation between X_t and X_{t-1}, X_t and X_{t-2}, and so on, statistically adjusting out the influence of intermediate lags. For example, the partial autocorrelation of lag four is the partial correlation between X_t and X_{t-4} after statistically removing the influence of X_{t-1}, X_{t-2}, and X_{t-3} from both X_t and X_{t-4}.

The autoregressive order, *p*, is estimated as the lag of the last large partial autocorrelation. For example, suppose the partial autocorrelations were

Lag	Partial Autocorrelation
1	0.55
2	0.21
3	0.11
4	0.72
5	0.06
6	0.09
7	0.13

We would conclude that a reasonable value for *p* is four, since the partial autocorrelations are relatively small after the fourth lag.

Box-Jenkins Methodology – An Overview

The Box-Jenkins method refers to the iterative application of the following three steps:

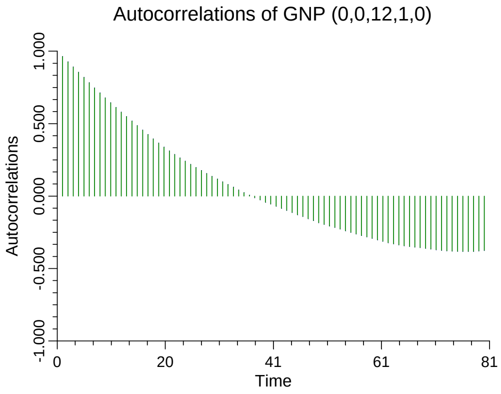
1. ***Identification***. Using plots of the data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA models is selected. This amounts to estimating appropriate values for *p*, *d*, and *q*.
2. ***Estimation***. The phis and thetas of the selected model are estimated using maximum likelihood techniques, backcasting, etc., as outlined in Box-Jenkins (1976).
3. ***Diagnostic Checking***. The fitted model is checked for inadequacies by considering the autocorrelations of the residual series (the series of residual, or error, values).

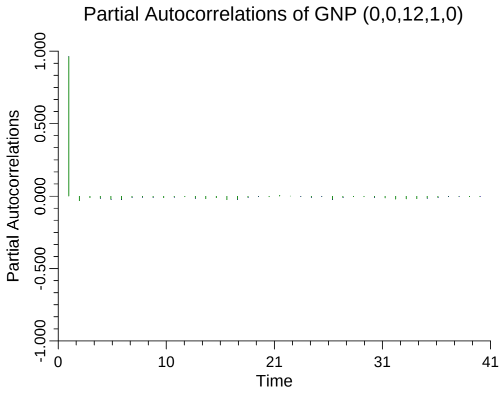
These steps are applied iteratively until step three does not produce any improvement in the model. We will now go over these steps in detail.

Model Identification

Assuming for the moment that there is no seasonal variation, the objective of the model identification step is to select values of *d* and then *p* and *q* in the *ARIMA(p,d,q)* model. When the series exhibits a trend, we may either fit and remove a deterministic trend or difference the series. Box-Jenkins seem to prefer differencing, while several other authors prefer the deterministic trend removal.

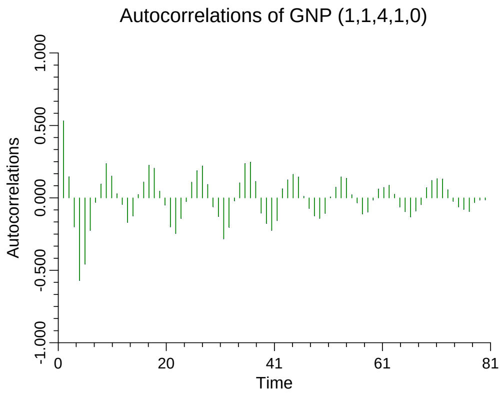
The first step, in either case, is to look at the plots of the autocorrelations and partial autocorrelations. A series with a trend will have an autocorrelation patterns similar to the following:

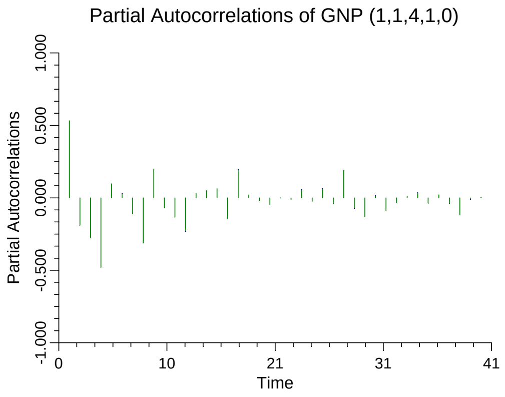




We notice that the large autocorrelations persist even after several lags. This indicates that either a trend should be removed or that the series should be differenced. The next step would be to difference the series.

When the series is differenced, the autocorrelation plots might appear as follows:





Differencing usually reduces the number of large autocorrelations considerably. If the differenced series still does not appear stationary, we would have to difference it again.

It is often useful to determine the magnitude of a large autocorrelation and partial autocorrelation coefficient. An autocorrelation must be at least $2/\sqrt{N}$, in absolute value to be statistically significant. The following list gives some common values of significant autocorrelations for various sample sizes. Note that even though an autocorrelation is statistically significant, it may not be large enough to worry about.

N	Large Autocorrelation
25	0.40
50	0.28
75	0.23
100	0.23
200	0.14
500	0.09
1000	0.06

By considering the patterns of the autocorrelations and the partial autocorrelations, we can guess a reasonable model for the data. The following chart shows the autocorrelation patterns that are produced by various types of ARMA models.

Model	Autocorrelations	Partial Autocorrelations
ARIMA(p,d,0)	Infinite. Tails off.	Finite. Cuts off after *p* lags.
ARIMA(0,d,q)	Finite. Cuts off after *q* lags.	Infinite. Tails off.
ARIMA(p,d,q)	Infinite. Tails off.	Infinite. Tails off.

The identification phase determines the values of *d* (differencing), *p* (autoregressive order), and *q* (moving average order). By studying the two autocorrelation plots, you estimate these values.

Differencing

The level of differencing is estimated by considering the autocorrelation plots. When the autocorrelations die out quickly, the appropriate value of *d* has been found.

Value of *p*

The value of *p* is determined from the partial autocorrelations of the appropriately differenced series. If the partial autocorrelations cut off after a few lags, the last lag with a large value would be the estimated value of *p*. If the partial autocorrelations do not cut off, you either have a moving average model *(p=0)* or an ARIMA model with positive *p* and *q*.

Value of *q*

The value of *q* is found from the autocorrelations of the appropriately differenced series. If the autocorrelations cut off after a few lags, the last lag with a large value would be the estimated value of *q*. If the autocorrelations do not cut off, you either have an autoregressive model *(q=0)* or an ARIMA model with a positive *p* and *q*.

Mixed Model

When neither the autocorrelations or the partial autocorrelations cut off, a *mixed model* is suggested. In an *ARIMA(p,d,q)* model, the autocorrelation function will be a mixture of exponential decay and damped sine waves after the first *q-p* lags. The partial autocorrelation function have the same pattern after *p-q* lags. By studying the first few correlations of each plot, you may be able to obtain reasonable guesses for *p* and *q*.

Our experience has been that directly identifying the values of *p* and *q* in mixed models is very difficult. Instead, we use a trial and error approach in which successively more complex models are fit until the residuals show no further structure (large autocorrelations). Usually, we try fitting an *ARIMA(1,d,0),* an *ARIMA(2,d,1),* and an *ARMA(4,3).* We would select the simplest model that had a reasonably good fit. (We have found that the *ARIMA(2,d,1)* often works well and we usually begin with it.)

Identification of a seasonal series is much more difficult. Box-Jenkins describe methods for model identification, but the user must be very skilled and experienced to successfully identify the model order. We have found that trial and error must usually be used. Usually, you want to keep the number of parameters to a minimum, so the values of *p, P, q, Q, d,* and *D* that you select should be less than or equal to two.

As you can see, the identification step is subjective. One of the frequent objections about the Box-Jenkins method is that two trained forecasters will arrive at different forecasting models, even though they are using the same software. However, as we showed earlier, often models that appear to be very different on the surface are actually quite similar.

Model Estimation and Diagnostic Checking

Maximum Likelihood Estimation

Once you have guestimated values of *p, d,* and *q*, you are ready to estimate the phis and thetas. This program follows the maximum likelihood estimation process outlined in Box-Jenkins (1976). The maximum likelihood equation is solved by nonlinear function maximization.

Backcasting is used to obtain estimates of the initial residuals. The estimation process is calculation intensive and iterative, so it often takes a few seconds to obtain a solution.

Diagnostic Checking

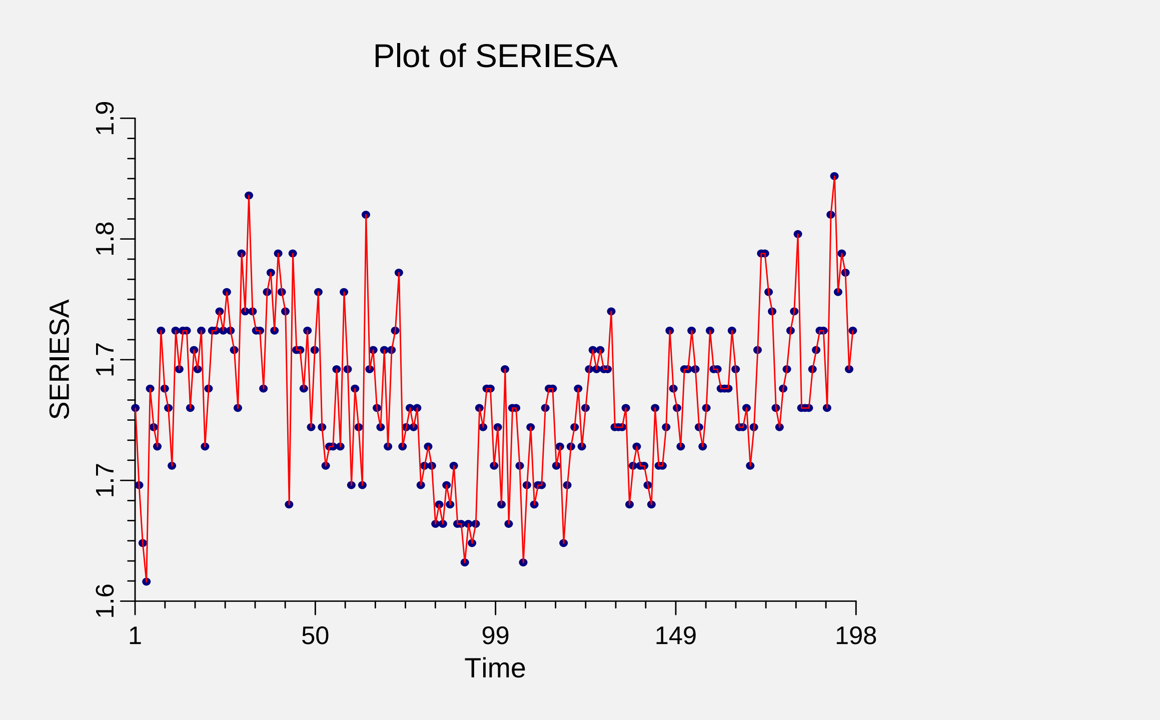
Once a model has been fit, the final step is the diagnostic checking of the model. The checking is carried out by studying the autocorrelation plots of the residuals to see if further structure (large correlation values) can be found. If all the autocorrelations and partial autocorrelations are small, the model is considered adequate and forecasts are generated. If some of the autocorrelations are large, the values of *p* and/or *q* are adjusted and the model is re-estimated.

This process of checking the residuals and adjusting the values of *p* and *q* continues until the resulting residuals contain no additional structure. Once a suitable model is selected, the program may be used to generate forecasts and associated probability limits.

Example 1 – Chemical Process Concentrations

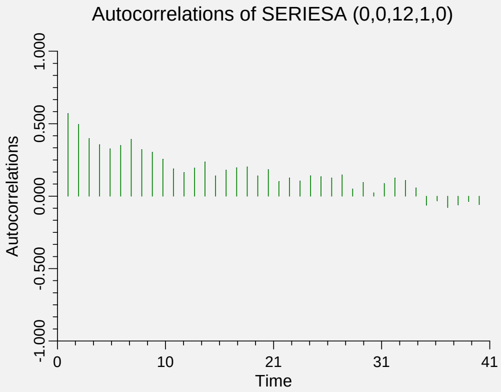
To complete this chapter, we will construct forecasts for two example problems. The first example we consider is called Series A by Box-Jenkins, and is from their book. This is a set of 197 concentration values from a chemical process taken at two-hour intervals. The data are stored in the SERIESA database. If you want to follow along, you should open this data base now. The following figure shows a plot of the data.

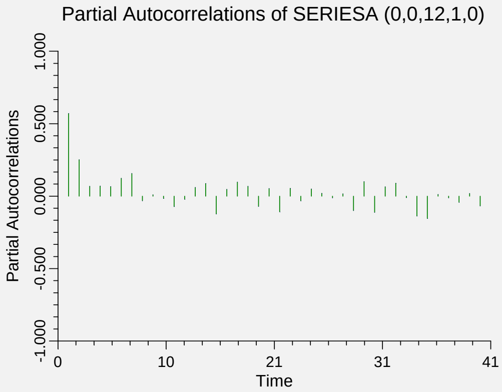
Time Series Data Plot



Notice that although the series moves around, it does not seem to follow a definite trend. The autocorrelation charts are shown next.

Series Autocorrelation Plots





The autocorrelations seem to die down fairly regularly after lag 1. The partial autocorrelations seem to be small after the first one, so we decide to fit an *ARIMA(1,0,1)* to these data.

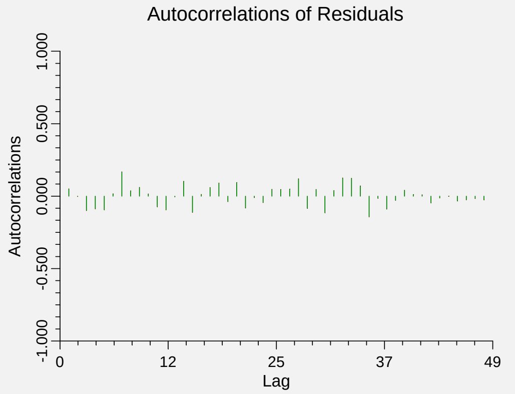
Model Estimation Reports

The following output shows the results of fitting the model.

```
Model Description Section
Series                          SERIESA-MEAN
Model                           Regular(1,0,1)   Seasonal(No seasonal parameters)
Mean                            1.706244

Observations                    197
Iterations                      11
Pseudo R-Squared                38.477242
Residual Sum of Squares         0.1922096
Mean Square Error               9.856902E-04
Root Mean Square                0.0313957

Model Estimation Section

Parameter       Parameter       Standard                     Prob
Name            Estimate        Error          T-Value       Level
AR(1)           0.9208993       4.111259E-02   22.3994       0.000000
MA(1)           0.5958619       8.240521E-02   7.2309        0.000000
```

The final step is to make the diagnostic checks of our model. The autocorrelation plot of the residuals are shown next.

Autocorrelation of Residuals Plot



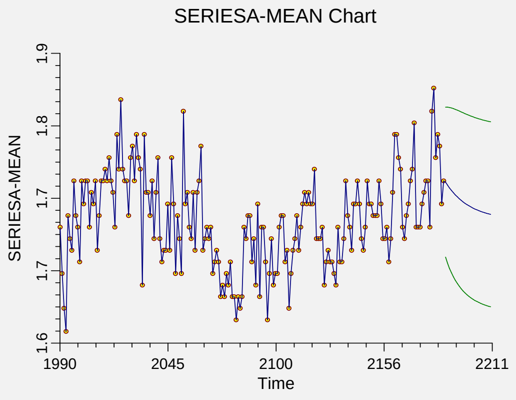
No action here. Finally, we take a look at the Portmanteau test results.

Portmanteau Test Report

Lag	DF	Portmanteau Test Value	Prob Level	Decision (0.05)
13	11	15.55	0.158664	Adequate Model
14	12	17.75	0.123437	Adequate Model
15	13	20.40	0.085570	Adequate Model
16	14	20.43	0.117064	Adequate Model
17	15	21.19	0.130966	Adequate Model
18	16	22.93	0.115544	Adequate Model
19	17	23.24	0.141718	Adequate Model
20	18	25.13	0.121460	Adequate Model
21	19	26.60	0.114351	Adequate Model
22	20	26.62	0.146230	Adequate Model
23	21	27.07	0.168631	Adequate Model
24	22	27.56	0.190707	Adequate Model

The diagnostic checking reveals no new patterns, so we can assume that our model is adequate. We generate the forecasts for the next few periods. These are shown next.

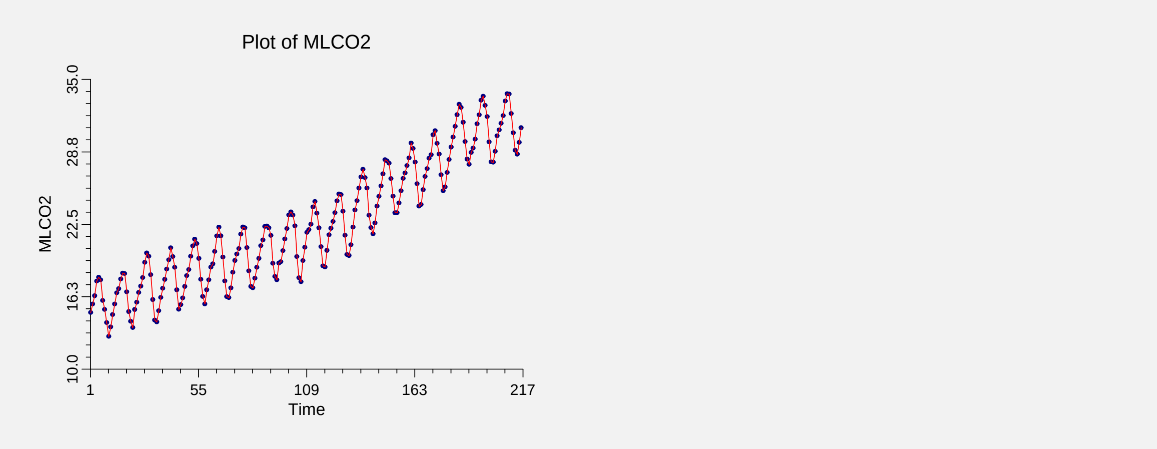
Time Series Plot Including Forecasts



Example 2 – Carbon Dioxide Above Mauna Loa, Hawaii

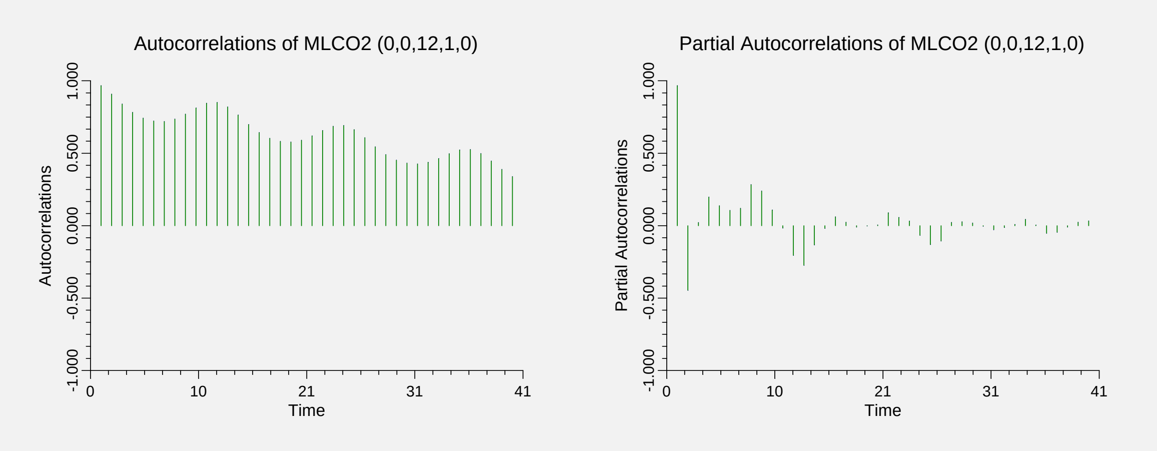
This example will an approach to data with a linear trend and seasonal variation. We will consider 216 monthly carbon dioxide measurements above Mauna Loa, Hawaii. The data was obtained from Newton (1988). It is stored in the data base named MLCO2.

Time Series Data Plot

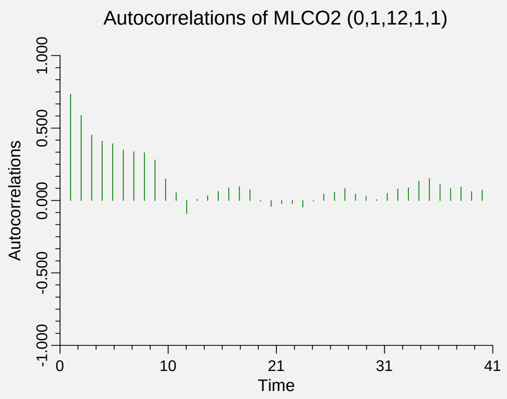


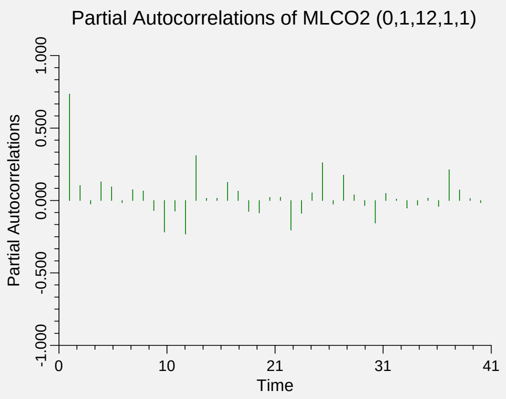
Note that the data are nonstationary on two counts: they show a trend and an annual cycle. The next step is to study the autocorrelations. The autocorrelation charts are shown next.

Series Autocorrelation Plots



Notice that the autocorrelations do not die out and they show a cyclical pattern. This points to nonstationarity in the data. The partial autocorrelations point to a value of 2 for p. However, because of the obvious nonstationarity, we first want to look at the autocorrelation functions of the first differences. Because these are monthly data, we use seasonal differences of length twelve. We also remove the trend in the data.





The autocorrelations die out fairly quickly. The partial autocorrelations are large around lags one and twelve. This suggests the multiplicative seasonal model: ARIMA(1,0,0) x (1,1,0)$_{12}$.

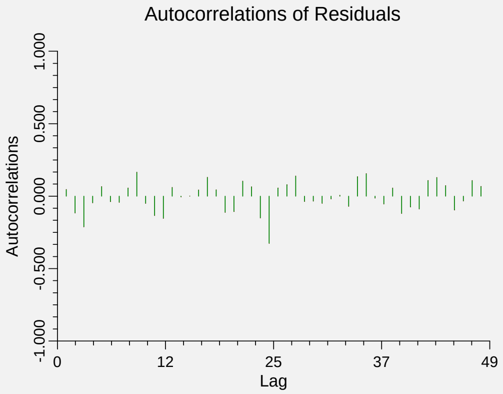
Model Estimation Reports

Following are the results of fitting this model.

```
Model Description Section
Series                        MLCO2-TREND
Model                         Regular(1,0,1)    Seasonal(1,1,0) Seasons = 12
Trend Equation                (14.07418)+(7.830546E-02)x(date)

Observations                  216
Iterations                    13
Pseudo R-Squared              99.500042
Residual Sum of Squares       30.3262
Mean Square Error             0.1508766
Root Mean Square              0.3884284


Model Estimation Section

Parameter      Parameter        Standard                      Prob
Name           Estimate         Error          T-Value        Level
AR(1)          0.9836381        1.274416E-02   77.1834        0.000000
SAR(1)         -0.4927093       5.991305E-02   -8.2237        0.000000
MA(1)          0.3183001        6.915411E-02   4.6028         0.000004
```

Everything appears fine here. The final step is to make the diagnostic checks of our model. The autocorrelation plot of the residuals is shown next.

Autocorrelation of Residuals Plot



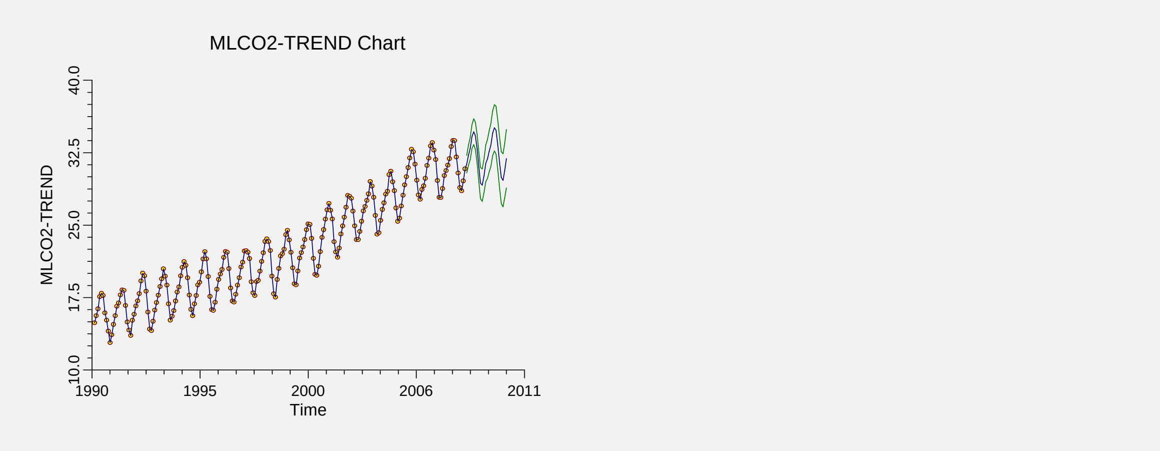
There appear to be some persistent autocorrelations at lag 25. We take a look at the Portmanteau test results.

Portmanteau Test Report

Lag	DF	Portmanteau Test Value	Prob Level	Decision (0.05)
13	10	32.78	0.000296	Inadequate Model
14	11	32.79	0.000570	Inadequate Model
15	12	32.79	0.001045	Inadequate Model
16	13	33.21	0.001585	Inadequate Model
17	14	37.13	0.000704	Inadequate Model
18	15	37.57	0.001044	Inadequate Model
19	16	40.51	0.000656	Inadequate Model
20	17	43.17	0.000453	Inadequate Model
21	18	45.72	0.000326	Inadequate Model
22	19	46.73	0.000391	Inadequate Model
23	20	52.17	0.000108	Inadequate Model
24	21	77.62	0.000000	Inadequate Model

The test points to additional information in the residual autocorrelations. We should refine our model further. We tried several other models, but could not find one that worked a lot better. Finally, we generate the forecasts from this model.

Time Series Plot Including Forecasts



As an exercise, you might try fitting this data with the Winters exponential smoothing algorithm.

Chapter 471

ARIMA (Box-Jenkins)

Introduction

Although the theory behind ARIMA time series models was developed much earlier, the systematic procedure for applying the technique was documented in the landmark book by Box and Jenkins (1976). Since then, ARIMA forecasting and Box-Jenkins forecasting usually refer to the same set of techniques. In this chapter, we will document the running of the ARIMA program. The methodology put forth by Box and Jenkins will be outlined in another chapter, since it uses several time series procedures.

ARIMA time series modeling is complex. You will want to become familiar with the details of the methodology before you place a lot of confidence in your forecasts. Our intent is to provide you with the tools you need to become proficient in the Box-Jenkins method.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by this algorithm. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable on which to run the analysis.

Time Series Variable

Time Series Variable

Specify the variable on which to run the analysis.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable.

Forecasting Options

Number of Forecasts

This option specifies the number of forecasts to be generated.

Data Adjustment Options

Remove Mean

Checking this option indicates that the series average should be subtracted from the data. This is almost always done.

Remove Trend

Checking this option indicates that the least squares trend line should be subtracted from the data. This is sometimes done, although differencing is often used to remove trends instead.

Regular Differencing

Specify the number of times to difference the series. You can enter 0, 1, or 2.

Seasonal Differencing

Specify the number of times to seasonally difference the series. The number of seasons per year is specified later.

Seasonality Options

Seasons

Specify the number of seasons per year in the series. Use ‘4’ for quarterly data or ‘12’ for monthly data.

First Season

Specify the first season of the series. This value is used to format the reports and plots. For example, if you have monthly data beginning with March, you would enter a ‘3’ here.

First Year

Specify the first year of the series. This value is used to format the reports and plots.

ARIMA Model Options

Regular AR

Specify the highest order of the autoregressive parameters. For example, if you specify '2' here, both the lag one and lag two autoregressive parameters will be included in the model.

Regular MA

Specify the highest order of the moving average parameters. For example, if you specify '2' here, both the lag one and lag two moving average parameters will be included in the model.

Seasonal AR

Specify the highest order of the seasonal autoregressive parameters. The number of seasons per year is specified later.

Seasonal MA

Specify the highest order of the seasonal moving average parameters. The number of seasons per year is specified later.

ARIMA Model Options

Max Iterations

The nonlinear estimation procedure will not converge for every model and data combination. This parameter lets you set the maximum number of iterations before the estimation algorithm is terminated.

Convergence

As the nonlinear estimation proceeds through each step, the residual sum of squares is calculated. When the ratio of the residual sum of squares from the current step to the residual sum of squares from the last step is less than this amount, the estimation procedure concludes. Hence, decreasing this amount will cause the procedure to go through more iterations, while increasing this amount will cause it to run fewer iterations.

Lambda

Lambda is a parameter from Marquart's nonlinear estimation procedure. This value of lambda was suggested by Marquart and we suggest that you leave it at the default value.

Reports Tab

The following options control which reports are displayed.

Select Additional Reports

Minimization Report - Portmanteau Test Report

Each of these options specifies whether the indicated report is displayed.

Forecast Report

This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Select Plots

Forecast Plot - Autocorrelation Plot

Each of these options specifies whether the indicated plot is displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic prediction limits of the forecasts. Usually, this number will range from 0.001 to 0.1. 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.

Decimals

Specifies the number of decimal places to use when displaying the forecasts.

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.

Plot Options

Large Plots

When checked, the plots displayed are larger (about five inches across) than normal (about two inches across).

Forecast Plot Tab

This section controls the forecast plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forecast 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

This option controls the attributes of the plotting 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.

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.

Autocorrelation Plot Tab

This section controls the autocorrelation plots.

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.

Uniform Scaling

Check this option to scale the autocorrelation plots the same.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Autocorrelation Plot

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

This option controls the attributes of the plotting 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.

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.

Storage Tab

The forecasts, prediction limits, 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.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Forecasts, Residuals, Lower Prediction Limits, and Upper Prediction Limits

The forecasts, residuals (Y-forecast), 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 an ARIMA Model

This section presents an example of how to fit an ARIMA model to a time series. The Intel_Close variable in the INTEL database will be fit with an ARMA(2,0,0) model.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the ARIMA (Box-Jenkins) window

1 Open the Intel dataset.

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **Intel.S0**.
- Click **Open**.

2 Open the ARIMA (Box-Jenkins) window.

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **ARIMA (Box-Jenkins)**. The ARIMA (Box-Jenkins) 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 ARIMA (Box-Jenkins) window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.

- Select **Intel_Close** from the list of variables and then click **Ok**.
- Enter **2** in the **Regular AR** box.
- Enter **0** in the **Regular MA** box.

4 Specify the reports.

- On the ARIMA (Box-Jenkins) window, select the **Reports tab**.
- Select **Data and Forecasts** in the **Forecast Report**.

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      AR(1)          AR(2)
0       85.89011      0.1         0.1            0.1
1       22.34787      0.1         0.9479776      -0.1168852
2       17.8811       0.04        1.292394       -0.4623865
3       17.57868      0.016       1.390741       -0.5737677
4       17.57362      0.0064      1.403895       -0.589241
5       17.57359      0.00256     1.40471        -0.5902494
Normal convergence.
```

This report displays the algorithms progress toward a solution.

Error Sum of Squares

The sum of the squared residuals. This is the value that is being minimized by the algorithm.

Lambda

The value of Marquart’s lambda parameter.

AR(...), MA(...)

The values of the autoregressive and moving average parameters. Note that if there are more parameters in the model than will fit on a single report line, only the first few parameters are displayed.

Model Description Section

```
Model Description Section

Series                     Intel_Close-MEAN
Model                      Regular(2,0,0)    Seasonal(No seasonal parameters)
Mean                       64.45625

Observations               20
Iterations                 5
Pseudo R-Squared           84.653036
Residual Sum of Squares    17.57359
Mean Square Error          0.9763107
Root Mean Square           0.9880844
```

This report displays summary information about the solution.

Series

The name of the variable being analyzed.

Model

The phrase *Regular (p,d,q)* gives the highest order of the regular ARIMA parameters. The *Seasonal(P,D,Q)* gives the highest order of the seasonal ARIMA parameters, if they were used.

p	Highest order autoregression parameter in the model.
d	Number of times the series was differenced.
q	Highest order moving average parameter in the model.
P	Highest order seasonal autoregression parameter in the model.
D	Number of times the series was seasonally differenced.
Q	Highest order seasonal moving average parameter in the model.

Mean

The average of the series.

Observations

The number of observations (rows) in the series.

Iterations

The number of iterations before the algorithm converged or was aborted.

Pseudo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Residual Sum of Squares

The sum of the squared residuals. This is the value that is being minimized by the algorithm.

Mean Square Error

The average squared residual (MSE) is a measure of how closely the forecasts track the actual data. The statistic is popular because it shows up in analysis of variance tables. However, because of the squaring, it tends to exaggerate the influence of outliers (points that do not follow the regular pattern).

Root Mean Square

The square root of MSE. This statistic is popular because it is in the same units as the time series.

Model Estimation Section

```
Model Estimation Section

Parameter       Parameter          Standard                      Prob
Name            Estimate           Error          T-Value        Level
AR(1)           1.40471            0.2065638      6.8004         0.000000
AR(2)           -0.5902494         0.2330099      -2.5332        0.011304
```

Parameter Name

The is the name of the parameter that is reported on this line.

AR(i)	The ith-order autoregressive parameter.
MA(i)	The ith-order moving average parameter.
SAR(i)	The ith-order seasonal autoregressive parameter.
SMA(i)	The ith-order seasonal moving average parameter.

Parameter Estimate

This is the estimated parameter value.

Standard Error

A large sample (N>50) estimate of the standard error of the parameter value.

T-Value

The t-test value testing whether the parameter is statistically significant (different from zero). The degrees of freedom is equal to the N minus the number of model parameters and differences.

Prob Level

The probability level for the above test. If you were testing at the alpha = 0.05 level of significance, this value would have to be less than 0.05 in order for the parameter to be considered statistically different from zero. When the highest order parameter is not significance, you should decrease the order by one and rerun. When a nonsignificant parameter is not the highest order, you should not delete it.

Asymptotic Correlation Matrix of Parameters

```
Asymptotic Correlation Matrix of Parameters

         AR(1)           AR(2)
AR(1)    1.000000        -0.881734
AR(2)    -0.881734       1.000000
```

This report gives the asymptotic estimates of the correlation between the parameter estimates. If some of the correlations are greater than 0.9999, you should consider removing appropriate parameters.

Parameter Name

The is the name of the parameter that is reported on this line.

AR(i)	The ith-order autoregressive parameter.

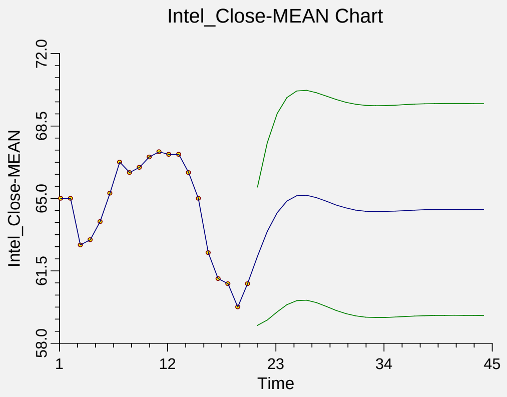
Forecast Section

Forecasts of SPOTS

Row	Date	Actual	Residual	Forecast	Lower 95% Limit	Upper 95% Limit
1	1	65.0	0.1	64.9	61.6	68.2
2	2	65.0	0.0	65.0	61.6	68.3
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
20	20	60.9	0.9	60.0	56.6	63.3
21	21			62.2	58.9	65.5
22	22			63.4	59.1	67.7
23	23			64.3	59.5	69.1
24	24			64.9	59.9	69.9
25	25			65.1	60.1	70.2

This section presents the forecasts, the residuals, and the 100(1-alpha)% prediction limits.

Forecast and Data Plot Section



This section displays a plot of the data values, the forecasts, and the prediction limits. It lets you determine if the forecasts are reasonable.

Autocorrelations of Residuals Section

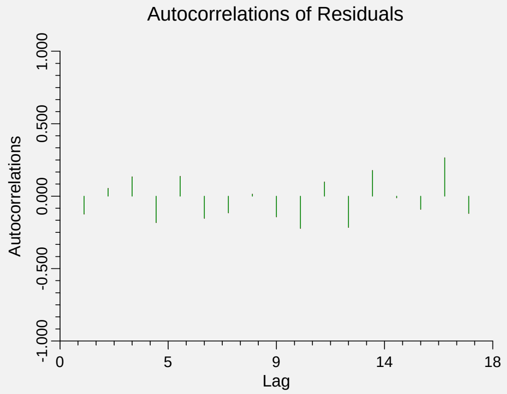
Autocorrelations of Residuals Section

Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
1	-0.124727	6	-0.152506	11	0.098453	16	0.264817
2	0.053877	7	-0.115625	12	-0.215545	17	-0.119435
3	0.133499	8	0.014388	13	0.177615		
4	-0.182724	9	-0.143282	14	-0.011190		
5	0.137313	10	-0.222115	15	-0.091297		

Significant if |Correlation|> 0.447214

If the residuals are white noise, these autocorrelations should all be nonsignificant. If significance is found in these autocorrelations, the model should be changed.

Autocorrelation Plot Section



This plot is the key diagnostic to determine if the model is adequate. If no pattern can be found here, you can assume that your model is as good as possible and proceed to use the forecasts. If large autocorrelations or a pattern of autocorrelations is found in the residuals, you will have to modify the model.

Portmanteau Test Section

```
Portmanteau Test Section Intel_Close-MEAN
                 Portmanteau       Prob
Lag     DF       Test Value        Level           Decision (0.05)
13      11       9.46              0.579765        Adequate Model
14      12       9.46              0.662805        Adequate Model
15      13       10.06             0.688621        Adequate Model
16      14       16.38             0.290932        Adequate Model
17      15       18.09             0.258059        Adequate Model
```

The Portmanteau Test (sometimes called the Box-Pierce-Ljung statistic) is used to determine if there is any pattern left in the residuals that may be modeled. This is accomplished by testing the significance of the autocorrelations up to a certain lag. In a private communication with Dr. Greta Ljung, we have learned that this test should only be used for lags between 13 and 24. The test is computed as follows:

$$Q(k) = N(N+2)\sum_{j=1}^{k}\frac{r_j^2}{N-j}$$

Q(k) is distributed as a Chi-square with *(K-p-q-P-Q)* degrees of freedom.

Chapter 472

Autocorrelations

Introduction

The correlation between X_t and X_{t+k} is called the kth order *autocorrelation* of X. The sample estimate of this autocorrelation, called r_k, is calculated using the formula:

$$r_k = \frac{\sum_{i=1}^{n-k}\left(X_i - \overline{X}\right)\left(X_{i+k} - \overline{X}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Autocorrelations are used extensively in time series analysis. When plotted, they become the *correlogram* which is used during the identification phase of the Box-Jenkins method. The large sample standard error of the sample autocorrelations is simply $1/\sqrt{n}$ so that large sample confidence limits are $\pm 2/\sqrt{n}$.

The kth order *partial autocorrelation* of X is the partial correlation between X_t and X_{t+k}, where the influence of $X_{t+1}, X_{t+2}, \cdots, X_{t+k-1}$ have been removed. We use the following recursive formulae to calculate the partial autocorrelations.

$$\hat{\phi}_{k+1,j} = \hat{\phi}_{k,j} - \hat{\phi}_{k+1,k+1}\hat{\phi}_{k,k-j+1}$$

$$\hat{\phi}_{k+1,k+1} = \frac{r_{k+1} - \sum_{j=1}^{k}\hat{\phi}_{k,j}r_{k+1-j}}{1 - \sum_{j=1}^{k}\hat{\phi}_{k,j}r_j}$$

The partial autocorrelations have the same large sample standard errors and confidence limits as do the autocorrelations. They are also used during the model identification phase of the Box-Jenkins method.

For this same reason, the filter is not used by this procedure.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by this algorithm. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Time Series Variable

Time Series Variable

Specify the variable on which to run the analysis.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable.

Data Adjustment Options

Remove Mean

Checking this option causes the series average to be subtracted from the data. This is almost always done.

Remove Trend

Checking this option causes the least squares trend line to be subtracted from the data. This is sometimes done, although differencing is usually used to remove trends.

Regular Differencing

This option lets you designate whether the original series, the first differences, or the second differences are analyzed. The first difference series, *W*, is calculated using the formula:

$$W_t = X_t - X_{t-1}$$

which may be written using the backshift operator, B, as:

$$W_t = (1 - B)X_t$$

The second difference series, *Z*, is the first difference of the *W* series. The formula is:

$$Z_t = W_t - W_{t-1}$$

which may be written using the backshift operator, B, as:

$$Z_t = (1 - B)^2 X_t$$

Seasonal Differencing

This option lets you designate whether the original series, the first seasonal differences, or the second seasonal differences are analyzed. Assuming the number of seasons is *s*, the first seasonal difference series, *W*, is calculated using the formula:

$$W_t = X_t - X_{t-s}$$

which may be written using the backshift operator, B, as:

$$W_t = \left(1 - B^s\right) X_t$$

The second seasonal difference series, *Z*, is the first seasonal difference of the *W* series. The formula is:

$$Z_t = W_t - W_{t-s}$$

which may be written using the backshift operator, B, as:

$$Z_t = \left(1 - B^s\right)^2 X_t$$

Seasonality Options

Seasons

Specify the number of seasons, *s*, in the series. Use ‘4’ for quarterly data or ‘12’ for monthly data. Note that this option is only used if seasonal differencing is used.

Reports Tab

The following options control which reports are displayed.

Select Reports

Autocorrelation Report - Partial Autocorrelation Report

Each of these options specifies whether the indicated report is displayed.

Select Plots

Autocorrelation Plot - Data Plot

Each of these options specifies whether the indicated plot is displayed.

Report / Plot Options

Number of Autocorrelations

Specify the number of autocorrelations that are calculated and displayed. Note that the number of autocorrelations must be less than the number of rows of data.

Number of Partial Autocorrelations

Specify the number of partial autocorrelations (PAC's) that are calculated and displayed. Note that the number of partial autocorrelations must be less than the number of rows of data.

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.

Autocorrelation Plot Tab

This section controls the autocorrelation plots.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Autocorrelation Plot

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

This option controls the attributes of the plotting 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.

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.

Data Plot Tab

A plot of the data may be displayed to help you assess the historical accuracy of the forecast method.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Data 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

This option controls the attributes of the plotting 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.

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.

Storage Tab

Both the autocorrelations and the partial autocorrelations 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.

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

Autocorrelations

The autocorrelations are stored in this variable.

Partial Autocorrelations

The partial autocorrelations are stored in this variable.

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 Autocorrelations of a Series

This section presents an example of how to generate autocorrelations of a series. The Spots variable in the SUNSPOT database will be used.

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

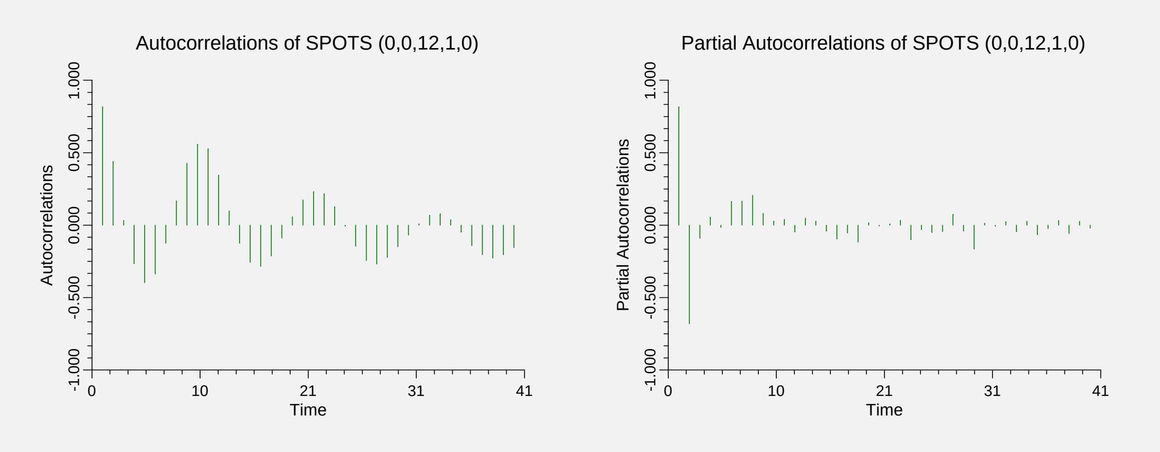
1. **Open the SUNSPOT dataset.**
   - From the File menu of the NCSS Data window, select **Open**.
   - Select the **Data** subdirectory of your NCSS directory.
   - Click on the file **SUNSPOT.S0**.
   - Click **Open**.

2. **Open the Autocorrelations window.**
   - On the menus, select **Analysis**, then **Forecasting/Time Series**, then **Autocorrelations**. The Autocorrelations 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 Autocorrelations window, select the **Variables tab**.
   - Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
   - Select **SPOTS** 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).

Autocorrelation Plots Section



This section displays the autocorrelations and partial autocorrelations in a plot format.

Autocorrelations Section

```
Autocorrelation Section

Autocorrelations of SPOTS (0,0,12,1,0)
Lag    Correlation      Lag    Correlation      Lag    Correlation      Lag    Correlation
1      0.816234         11     0.525615         21     0.230319         31     0.008908
2      0.439589         12     0.344871         22     0.216010         32     0.066994
3      0.031927         13     0.097503         23     0.126930         33     0.077056
4      -0.266327        14     -0.123974        24     -0.006907        34     0.037571
5      -0.395920        15     -0.256157        25     -0.143509        35     -0.048184
6      -0.335935        16     -0.283001        26     -0.243137        36     -0.141013
7      -0.124787        17     -0.211754        27     -0.268284        37     -0.204330
8      0.166522         18     -0.087193        28     -0.221329        38     -0.227475
9      0.426074         19     0.056621         29     -0.149028        39     -0.204460
10     0.558426         20     0.173492         30     -0.067392        40     -0.152599
Significant if |Correlation|> 0.136399
```

This section shows the values of the autocorrelations for the specified number of lags. The numbers in parentheses, (d,D,s,M,T), are defined as follows:

d is the regular differencing order.

D is the seasonal differencing order.

s is the number of seasons (ignored if D is 0).

M is 1 if the mean is subtracted, 0 otherwise.

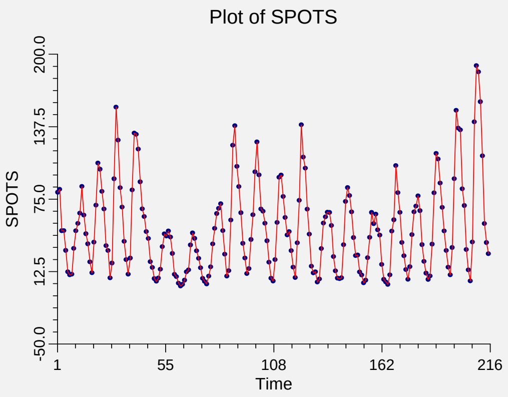
T is 1 if the trend is subtracted, 0 otherwise.

Partial Autocorrelation Section

```
Partial Autocorrelation Section

Partial Autocorrelations of SPOTS (0,0,12,1,0)
Lag     Correlation        Lag     Correlation        Lag     Correlation        Lag     Correlation
1       0.816234           11      0.038335           21      0.007443           31      -0.006587
2       -0.679072          12      -0.045487          22      0.032548           32      0.022903
3       -0.090380          13      0.047515           23      -0.100418          33      -0.042974
4       0.054429           14      0.027539           24      -0.029899          34      0.025960
5       -0.014413          15      -0.040822          25      -0.049139          35      -0.064242
6       0.163731           16      -0.095774          26      -0.043122          36      -0.021117
7       0.165977           17      -0.053699          27      0.074602           37      0.031502
8       0.205197           18      -0.114117          28      -0.040450          38      -0.056303
9       0.079963           19      0.016408           29      -0.164996          39      0.025227
10      0.026876           20      -0.003383          30      0.012685           40      -0.020340
Significant if |Correlation|> 0.136399
```

This section shows the values of the partial autocorrelations for the specified number of lags. The numbers in parentheses are defined above.

Data Plot Section



This section displays a plot of the data values.

Chapter 473

Cross-Correlations

Introduction

The cross correlation between X_t and Y_{t+k} is called the kth order *cross correlation* of *X* and *Y*. The sample estimate of this cross correlation, called r_k, is calculated using the formula:

$$r_k = \frac{\sum_{i=1}^{n-k}\left(X_i - \overline{X}\right)\left(Y_{i+k} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

The time index, *k*, is allowed to be either positive or negative. The large sample standard error of the sample cross correlations is simply $1/\sqrt{n}$ so that large sample confidence limits are $\pm 2/\sqrt{n}$.

Data Structure

The data are entered in two variables.

Missing Values

Missing values are not tolerated by this algorithm. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm. For this same reason, the filter is not used by this procedure.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Time Series Variables

Y Variable

Specify the first variable to be cross correlated.

X Variable

Specify the second variable to be cross correlated.

Reports Tab

The following options control which reports are displayed.

Select Reports

Cross-Correlation Report

This option specifies whether the indicated report is displayed.

Select Plots

Cross-Correlation Plot - Data Plots

Each of these options specifies whether the indicated plot is displayed.

Report Options

Number of Cross-Correlations

Specify the number of cross correlations that are calculated and displayed. Note that the number of cross correlations must be less than the number of rows of data.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

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.

Cross-Correlation Plot Tab

This section controls the cross-correlation plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Cross-Corr 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

This option controls the attributes of the plotting 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.

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.

Data Plot Tab

Characteristics of the plots of the two series across time are controlled here.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Data 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

This option controls the attributes of the plotting 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.

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.

Storage Tab

Both the autocorrelations and the partial autocorrelations 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.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Cross Correlations

The cross correlations are stored in this variable.

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 Cross-Correlations of Two Series

This section presents an example of how to generate cross correlations of two series. The Intel_Volume and Intel_Close variables in the INTEL database will be used.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Cross-Correlations window.

1 **Open the INTEL dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **INTEL.S0**.
- Click **Open**.

2 Open the Cross-Correlations window.

- On the menus, select **Analysis**, then **Forecasting/Time Series**, then **Cross-Correlations**. The Cross-Correlations procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

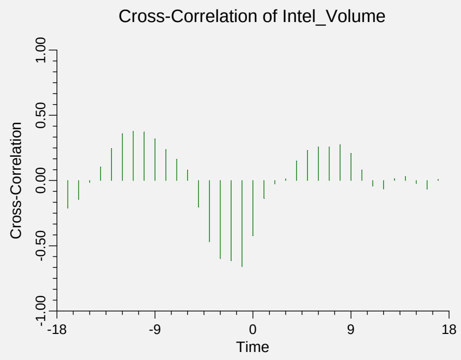
3 Specify the variables.

- On the Cross-Correlations window, select the **Variables tab**.
- Double-click in the **Y Variable** box. This will bring up the variable selection window.
- Select **Intel_Volume** 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 **Intel_Close** 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).

Cross Correlation Plot Section



This section displays the cross correlations from both positive and negative lags. The value at lag 0 is the simple correlation between these two variables.

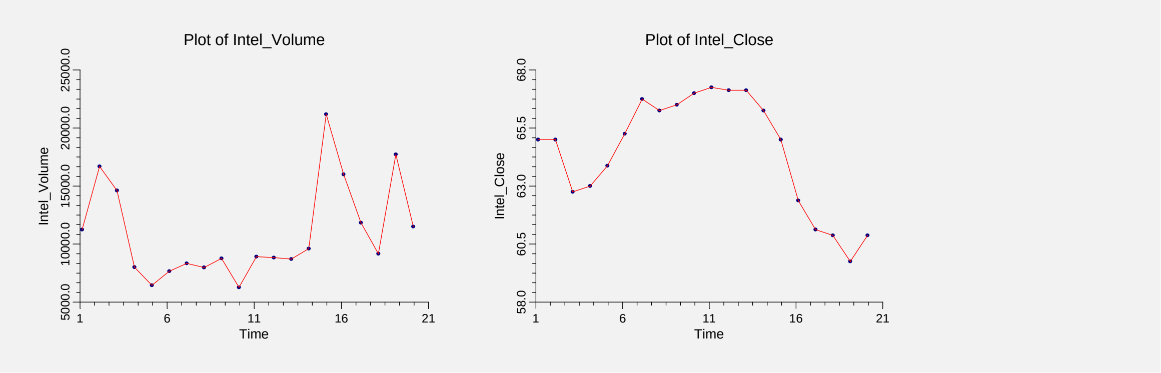
Cross Correlations Section

Cross-Correlations of Intel_Volume

Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
-17	-0.212089	-8	0.236876	1	-0.135799	10	0.080439
-16	-0.144536	-7	0.164298	2	-0.025696	11	-0.042123
-15	-0.012648	-6	0.080595	3	0.012290	12	-0.063397
-14	0.103407	-5	-0.202975	4	0.150682	13	0.013433
-13	0.244696	-4	-0.468207	5	0.229830	14	0.030845
-12	0.359095	-3	-0.596296	6	0.257356	15	-0.020010
-11	0.377600	-2	-0.615427	7	0.257654	16	-0.066433
-10	0.371336	-1	-0.657680	8	0.274812	17	0.009300
-9	0.318597	0	-0.422771	9	0.208603		

This section shows the values of the cross correlations for the specified number of lags.

Data Plot Section



This section displays plots of the data values.

473-8 Cross-Correlations

Chapter 474

Automatic ARMA

Introduction

The ARIMA (or Box-Jenkins) method is often used to forecast time series of medium (N over 50) to long lengths. It requires the forecaster to be highly trained in selecting the appropriate model. The procedure discussed here automates the ARIMA forecasting process by having the program select the appropriate model.

The Method

The Automatic ARMA program uses methodology from several authors to find and estimate an appropriate forecasting model. The method may be outlined as follows:

1. Using the model selection theory of Pandit and Wu (1983), any deterministic trend is removed from the series.
2. A set of models of increasing complexity is fit. These are *ARIMA(1,0,0)*, *ARIMA(2,0,1)*, *ARIMA(4,0,3)*, *ARIMA(6,0,5)*, and so on, increasing both *p* and *q* by two at each step. The most complex model tried is specified in the Maximum Order box. The residual sum of squares is calculated for each model and the minimum is noted.
3. Using the minimum residual sum of squares as the criterion, the models are again arranged from simpliest to most complex. The first model to be within the user-defined percentage of the minimum sum of squares is selected and used.
4. Once this model has been determined, one final attempt is made to find a model of smaller order that is within the specified percentage of the minimum. Suppose the previous steps lead to an *ARIMA(4,3)* model. This step would fit an *ARIMA(3,0,2)* model and check to see if the residual sum of squares was within the specified percentage. If it was, the *ARIMA(3,0,2)* model would be used. If not, the *ARIMA(4,3)* model would be used.

Because the procedure has to fit so many models, several of which are of large order, we use a sub-optimal (but much faster) model estimation algorithm. We chose the least squares modified Yule-Walker technique described in Marple (1987), section 10.4. This method is fast and seems to provide reasonable estimates of the residual sum of squares.

Data Structure

The data are entered in a single variable.

Missing Values

Missing values are not tolerated by this algorithm. When missing values are found in the series, they are replaced by the average of the nearest observation in the future and in the past. If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. This missing value replacement algorithm is particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variable on which to run the analysis.

Time Series Variable

Time Series Variable

Specify the variable on which to run the analysis.

Use Logarithms

Specifies that the log (base 10) transformation should be applied to the values of the variable.

Forecasting Options

Number of Forecasts

This option specifies the number of forecasts to be generated.

Data Adjustment Options

Remove Mean

Checking this option indicates that the series average should be subtracted from the data. This is almost always done.

Remove Trend

Checking this option indicates that the least squares trend line should be subtracted from the data. This option should be used if a trend is apparent in the data.

ARIMA Model Options

Maximum Order

The largest number of AR parameters that will be tried. If you are using seasonal data, this should be two more than the length of any seasonal pattern. Hence, for monthly data you would try fourteen, for quarterly data you would use six, and for annual data you would use four or six.

Even-order models are tried up to this size. For example, if you enter a six here, the program will fit the ARIMA models *ARIMA(2,0,1)*, *ARIMA(4,0,3)*, and *ARIMA(6,0,5)*. The residual sum of squares is noted, and the simplest model is used for forecasting.

Percent of Best

Once the program has found the residual sum of squares for each of the models designated by the Maximum Order, it finds the smallest of these values. It then searches through models, calculating the percent increase in the residual sum of squares of the current model over that of the best model. It selects the simplest (smallest in number of parameters) model that is less than this criterion.

Hence, the larger the percentage you enter here, the simpler will be the model. Normally, the value of five is sufficient.

Autoregressive Terms

When this value is greater than zero, no search is conducted. Instead, a model with this specific autoregressive order is calculated.

Moving Average Terms

When this value is greater than zero, no search is conducted. Instead, a model with this specific moving average order is calculated.

Seasonality Options

Seasons

Specify the number of seasons per year in the series. Use ‘4’ for quarterly data or ‘12’ for monthly data.

First Season

Specify the first season of the series. This value is used to format the reports and plots. For example, if you have monthly data beginning with March, you would enter a ‘3’ here.

First Year

Specify the first year of the series. This value is used to format the reports and plots.

Reports Tab

The following options control which reports are displayed.

Select Additional Reports

Search Report - Portmanteau Test Report

Each of these options specifies whether the indicated report is displayed.

Forecast Report

This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Select Plots

Forecast Plot - Autocorrelation Plot

Each of these options specifies whether the indicated plot is displayed.

Report Options

Alpha Level

The value of alpha for the asymptotic prediction limits of the forecasts. Usually, this number will range from 0.001 to 0.1. 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.

Decimals

Specifies the number of decimal places to use when displaying the forecasts.

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.

Plot Options

Large Plots

When checked, the plots displayed are larger (about five inches across) than normal (about two inches across).

Forecast Plot Tab

This section controls the forecast plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Forecast 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

This option controls the attributes of the plotting 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.

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.

Autocorrelation Plot Tab

This section controls the autocorrelation plots.

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.

Uniform Scaling

Check this option to scale the autocorrelation plots the same.

Minimum

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Autocorrelation Plot

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

This option controls the attributes of the plotting 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.

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.

Storage Tab

The forecasts, prediction limits, 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.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Forecasts, Residuals, Lower Prediction Limits, and Upper Prediction Limits

The forecasts, residuals (Y-forecast), 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 an Automatic ARMA Model

This section presents an example of how to fit an Automatic ARMA model. The SeriesA variable in the SERIESA database will be fit.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Automatic ARMA window.

1 **Open the SERIESA dataset.**

- From the File menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SeriesA.S0**.
- Click **Open**.

2 **Open the Automatic ARMA window.**

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Automatic ARMA**. The Automatic ARMA 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 Automatic ARMA window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **SeriesA** from the list of variables and then click **Ok**.

4 Specify the reports.

- On the Automatic ARMA window, select the **Reports tab**.
- Enter **3** in the **Decimals** box.

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).

Model Search Results Section

```
Model Search Results Section

        AR Order        MA Order      Sum of             Pseudo       Percent Change
No.     (P)             (Q)           Squares            R-Squared    From Last
0       0               0             0.3119473          0.00         0.00
1       1               0             0.2103325          32.57        -32.57
2       2               1             0.196571           36.99        -6.54
3       4               3             0.1952559          37.41        -0.67
4       6               5             0.1899079          39.21        -2.80
5       8               7             0.1826133          41.46        -3.76
```

This report displays information about the various models that were fit during the search. In this case, we note that the selected model is ARIMA(6,0,5). The individual definitions are as follows:

AR Order (P)

The number of autoregressive parameters in the model.

MA Order (Q)

The number of moving average parameters in the model.

Sum Squares

The sum of the squared residuals. The smaller this amount, the better the precision of the model.

Psuedo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

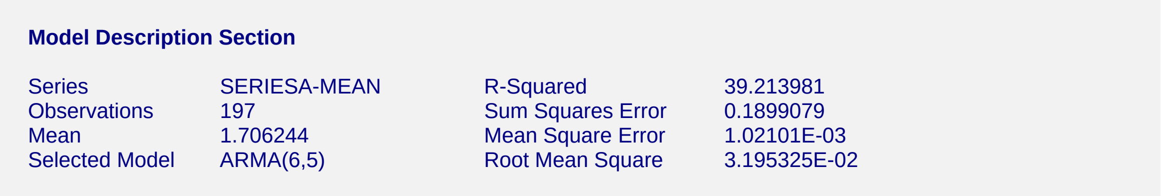
$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Percent Change From Last

The percent change in the sum of squares from model immediately above.

Model Description Section



```
Model Description Section

Series            SERIESA-MEAN          R-Squared            39.213981
Observations      197                   Sum Squares Error    0.1899079
Mean              1.706244              Mean Square Error    1.02101E-03
Selected Model    ARMA(6,5)             Root Mean Square     3.195325E-02
```

This report displays summary information about the solution.

Series

The name of the variable being analyzed.

Observations

The number of observations (rows) in the series.

Trend Equation

The trend equation that was fit and removed from the series before the ARMA models were fit.

Selected Model

The phrase *ARMA (p,q)* gives the highest order of the regular ARMA parameters.

p Number of autoregression parameters in the model.

q Number of moving average parameters in the model.

R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Sum of Squares Error

The sum of the squared residuals. This is the value that is being minimized by the algorithm.

Mean Square Error

The average squared residual (MSE) is a measure of how closely the forecasts track the actual data. The statistic is popular because it shows up in analysis of variance tables. However, because of the squaring, it tends to exaggerate the influence of outliers (points that do not follow the regular pattern).

Root Mean Square

The square root of MSE. This statistic is popular because it is in the same units as the time series.

Model Estimation Section

```
Model Estimation Section

Parameter        Parameter
Name             Estimate
AR(1)            0.3655652
AR(2)            0.1581511
AR(3)            0.0183087
AR(4)            3.503909E-02
AR(5)            1.653267E-02
AR(6)            0.1440598
MA(1)            1.811239E-02
MA(2)            -5.549401E-02
MA(3)            4.643534E-03
MA(4)            2.481859E-03
MA(5)            -2.905689E-02
```

Parameter Name

The is the name of the parameter that is reported on this line.

AR(i) The ith-order autoregressive parameter.

MA(i) The ith-order moving average parameter.

Parameter Estimate

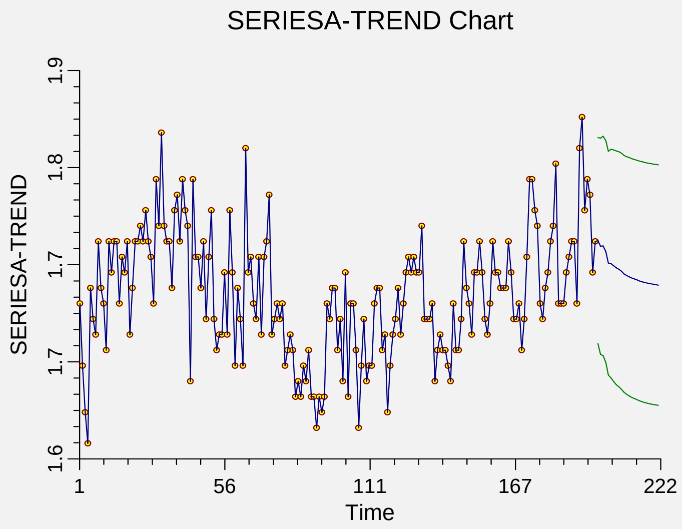
This is the estimated parameter value.

Forecast Section

```
Fourier Analysis of SPOTS (0,0,12,1,0)
Row      Date   Forecast   Lower 95% Limit   Upper 95% Limit
198      17 6   1.740      1.673             1.806
199      17 7   1.736      1.666             1.805
200      17 8   1.736      1.665             1.807
201      17 9   1.732      1.661             1.804
202      17 10  1.725      1.652             1.797
203      17 11  1.724      1.650             1.798
204      17 12  1.722      1.648             1.797
205      18 1   1.721      1.646             1.796
206      18 2   1.720      1.644             1.796
207      18 3   1.718      1.642             1.795
```

This section presents the forecasts, the residuals, and the 100(1-alpha)% prediction limits.

Forecast and Data Plot Section



This section displays a plot of the data values, the forecasts, and the prediction limits. It lets you determine if the forecasts are reasonable.

Autocorrelations of Residuals Section

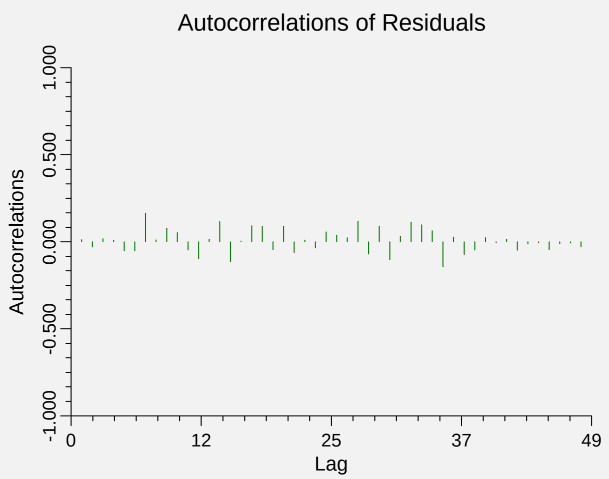
Autocorrelations of Residuals Section

Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
1	0.012330	13	0.013916	25	0.035760	37	-0.071720
2	-0.029505	14	0.115632	26	0.022786	38	-0.047004
3	0.017103	15	-0.114148	27	0.115409	39	0.023569
4	0.008808	16	0.003723	28	-0.070129	40	-0.002934
5	-0.051858	17	0.091939	29	0.086803	41	0.013134
6	-0.052182	18	0.088479	30	-0.101699	42	-0.047757
7	0.162349	19	-0.043432	31	0.029961	43	-0.011785
8	0.010603	20	0.088817	32	0.112346	44	-0.003118
9	0.076537	21	-0.060232	33	0.097306	45	-0.045457
10	0.052340	22	0.009222	34	0.062783	46	-0.010849
11	-0.048122	23	-0.035292	35	-0.142371	47	-0.007085
12	-0.095803	24	0.056370	36	0.027464	48	-0.028750

Significant if |Correlation|> 0.142494

If the residuals are white noise, these autocorrelations should all be nonsignificant. If significance is found in these autocorrelations, the model should be changed.

Autocorrelation Plot Section



This plot is the key diagnostic to determine if the model is adequate. If no pattern can be found here, you can assume that your model is as good as possible and proceed to use the forecasts. If large autocorrelations or a pattern of autocorrelations is found in the residuals, you will have to modify the model.

Portmanteau Test Section

```
Portmanteau Test Section Intel_Close-MEAN
                Portmanteau       Prob
Lag    DF       Test Value        Level         Decision (0.05)
13     2        11.12             0.003849      Inadequate Model
14     3        13.98             0.002927      Inadequate Model
15     4        16.79             0.002122      Inadequate Model
16     5        16.79             0.004908      Inadequate Model
17     6        18.63             0.004827      Inadequate Model
18     7        20.35             0.004862      Inadequate Model
19     8        20.76             0.007799      Inadequate Model
20     9        22.51             0.007390      Inadequate Model
21     10       23.32             0.009624      Inadequate Model
22     11       23.34             0.015825      Inadequate Model
23     12       23.62             0.022901      Inadequate Model
24     13       24.34             0.028143      Inadequate Model
```

The Portmanteau Test (sometimes called the Box-Pierce-Ljung statistic) is used to determine if there is any pattern left in the residuals that may be modeled. This is accomplished by testing the significance of the autocorrelations up to a certain lag. In a private communication with Dr. Greta Ljung, we have learned that this test should only be used for lags between 13 and 24. The test is computed as follows:

$$Q(k) = N(N+2)\sum_{j=1}^{k} \frac{r_j^2}{N-j}$$

Q(k) is distributed as a Chi-square with *(K-p-q)* degrees of freedom.

Chapter 475

Theoretical ARMA

Introduction

This procedure shows the theoretical characteristics of the autocorrelations, partial autocorrelations, and spectrum of user-specified ARMA models. Unlike the other time series programs, this one does not use data. Instead, it provides a theoretical analysis of various models. It also creates simulated series from these models.

We have found this program especially useful in training and model evaluation. While you are becoming familiar with the Box-Jenkins method, this program lets you study the characteristics of a large number of models. You will be able to see the sensitivity of the autocorrelation function to changes in the number of, and values of, parameters. You will be able to generate series from known models and see how difficult it is to identify the model that they came from.

For use in theoretical model evaluation, this program factors a model written as a polynomial in the backshift operator. This will let you compare several models that each seem adequate, but appear quite different. It will let you study the characteristics of various models in detail.

It is useful in model identification, because it will let you generate a catalog of possible autocorrelation patterns from known theoretical models that you can compare sample autocorrelation functions with.

All of the models treated by this program come from the general class defined by the model:

$$\phi_p(B)\Phi_P(B)X_t = \theta_q(B)\Theta_Q(B^s)a_t$$

(Refer to chapter on the Box-Jenkins method for more information on the interpretation of this equation.)

Procedure Options

This section describes the options available in this procedure.

Data Tab

The following options specify the model to be analyzed.

ARMA Model Specification

Regular AR (Phis)

The values of the autoregressive parameters, the phis.

This should be a list of values like **0.9 -0.2 0.3**. Note that the first value corresponds to lag one, the second corresponds to lag two, and so on.

Seasonal AR (Cap Phis)

The values of the seasonal autoregressive parameters, the capital phis. This should be a list of values like **0.9 -0.2 0.3**. Note that the first value corresponds to lag *s*, the second corresponds to lag *2s*, and so on.

Regular MA (Thetas)

The values of the moving average parameters, the thetas.

This should be a list of values like **0.9 -0.2 0.3**. Note that the first value corresponds to lag one, the second corresponds to lag two, and so on.

Seasonal MA (Cap Thetas)

The values of the seasonal moving average parameters, the capital thetas. This should be a list of values like **0.9 -0.2 0.3**. Note that the first value corresponds to lag *s*, the second corresponds to lag *2s*, and so on.

Seasonality Options

Seasons

Specify the number of seasons per year in the series. Use ‘4’ for quarterly data or ‘12’ for monthly data.

Simulation Options

Series Variance

This is the variance of the simulated series. It is the variance of the normal random numbers that are used in the simulation process.

Number of Rows

Specifies the number of rows of data that are generated in the simulated series from this model. This is also the number of rows that will be stored on the current database.

Reports Tab

The following options control which reports are displayed.

Select Reports

Autocorrelation Report - Coefficient Report

Each of these options specifies whether the indicated report is displayed.

Select Plots

Autocorrelation Plots - Data Plot

Each of these options specifies whether the indicated plot is displayed.

Report / Plot Options

Number of Autocorrelations

Specifies the number of autocorrelations that are reported.

Number of Frequencies

Specifies the number of frequencies displayed in the spectral density.

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.

Autocorrelation Plot Tab

This section controls the autocorrelation plots.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Autocorr 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

This option controls the attributes of the plotting 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.

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.

Spectrum Plot Tab

This section controls the spectrum plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Spectrum 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

This option controls the attributes of the plotting 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.

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.

Data Plot Tab

A plot of the data over time may be displayed. This tab controls that plot.

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

This option specifies the minimum value displayed on the corresponding axis. If left blank, it is calculated from the data.

Maximum

This option specifies the maximum value displayed on the corresponding axis. If left blank, it is 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.

Major Ticks - Minor Ticks

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

Show Grid Lines

This check box indicates whether the grid lines that originate from this axis should be displayed.

Autocorrelation Plot

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

This option controls the attributes of the plotting 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.

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.

Storage Tab

The generated series may be stored on the current database for further analysis. These options lets you designate which variable should receive this simulated series. The selected statistics are automatically stored to the current database when a variable name is entered.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Variables

Data Series

The simulated series is stored in this variable.

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 – Using the Theoretical ARMA Procedure

This section presents an example of how to use the theoretical ARMA program. You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Theoretical ARMA window.

1 **Open the Theoretical ARMA window.**

- On the menus, select **Analysis**, then **Forecasting / Time Series**, then **Theoretical ARMA**. The Theoretical ARMA procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

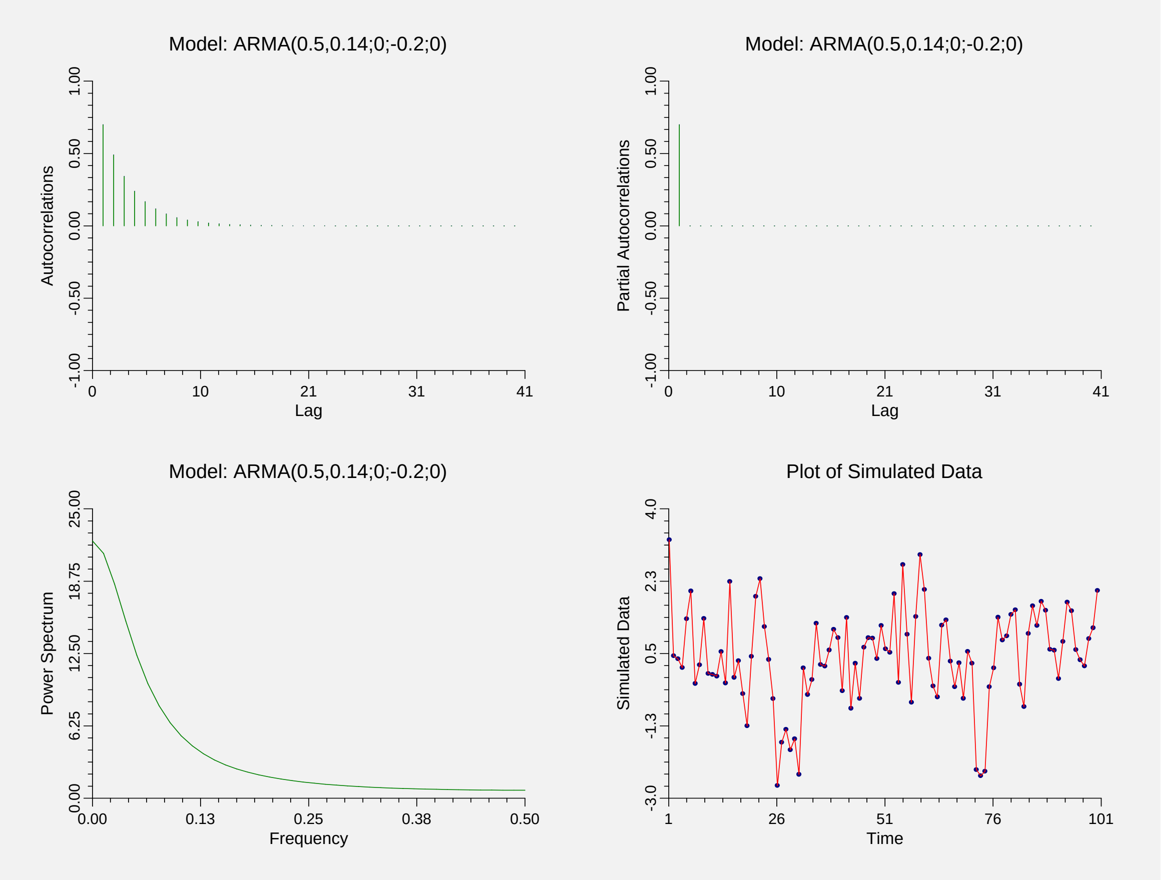
2 **Specify the model.**

- On the Theoretical ARMA window, select the **Data tab**.
- Enter **0.5 0.14** in the **Regular AR (Phis)** box.
- Enter **-0.2** in the **Regular MA (Thetas)** box.

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).

Plot Section



This section shows four plots of statistics generated from the model. The top two plots give the theoretical autocorrelations and partial autocorrelations. The bottom-left plot shows the power spectrum (see the chapter on spectral analysis) and the bottom-right plot is a plot of a simulated data series from the model.

Autocorrelation / Power Spectrum Section

```
Autocorrelations / Power Spectrum Section
                      Partial                                  Power
No.   Autocorrelations  Autocorrelations  Frequency            Spectrum
1     0.700000          0.700000          0.000000             22.22222
2     0.490000          0.000000          0.012821             21.15512
3     0.343000          0.000000          0.025641             18.49632
4     0.240100          0.000000          0.038462             15.30439
5     0.168070          0.000000          0.051282             12.34195
6     0.117649          0.000000          0.064103             9.89958
7     0.082354          0.000000          0.076923             7.988447
8     0.057648          0.000000          0.089744             6.52031
9     0.040354          0.000000          0.102564             5.394032
10    0.028248          0.000000          0.115385             4.524017
11    0.019773          0.000000          0.128205             3.844779
12    0.013841          0.000000          0.141026             3.308101
```

This section presents the numerical values associated with the plots of the last section.

Coefficient Analysis Section

```
Coefficient Analysis Section

Coefficient                Coefficient       Real              Imaginary
Name             Lag       Value             Root              Root
Phi(AR)          0         1.000000          -5.000000         0.000000
Phi(AR)          1         0.500000          1.428571          0.000000
Phi(AR)          2         0.140000
Theta(MA)        0         1.000000          -5.000000         0.000000
Theta(MA)        1         -0.20000
Model is stationary and model is invertible.
```

This report displays an analysis of the coefficients of the model. When the model is written in terms of the backshift operator, *B*, it may be thought of as polynomials in *B*. Hence in our current example, we have two equations to study, one for the autoregressive operator and one for the moving average operator. These are:

$$\left(1-0.5B-0.14B^{2}\right)=0$$

and

$$\left(1+0.2B\right)=0$$

As we will show, it can be useful to find and compare the roots of these two equations, since knowledge of the roots lets us factor the equations. We see from the report that the roots of the first polynomial are 1.4286 (which is 10/7) and -5. We would like to arrange these roots so that the factors may be displayed in a standard form. To do this, we perform the following algebra on each root:

$$B = -5 \qquad\qquad B = 1.4286$$

Move the constant to the right side.

$$5 + B = 0 \qquad\qquad -1.4286 + B = 0$$

Divide by the constant.

$$1 + .2B = 0 \qquad\qquad 1 - .7B = 0$$

These are now in the special form that we can easily use them as factors. We note that the polynomial may be factored as

$$(1 -.5B -.14B2) = (1 + .2B)(1 - .7B)$$

Hence, the model (9.3) may be rewritten as

$$(1+0.2B)(1-0.7B)X_t = (1+0.2B)a_t$$

Notice that the left and right sides of this equation have *(1+0.2B)* as a common factor. We can cancel this factor out, leaving the simpler model:

$$(1-0.7B)X_t = a_t$$

These models are equivalent. Now we can see why the partial autocorrelation plot indicated a single autoregressive parameter even though we had specified two.

A second purpose for studying these coefficients is to look for signals to difference a series. Note that if a root is approximately unity, the factor will be approximately *(1 - B)*, the difference operator. Hence, when we find roots on the autoregressive side close to one, we can simplify the model by differencing the series.

One criticism of the Box-Jenkins method is that two well-trained forecasters will most likely arrive at different models. We find that this criticism is not well-founded since often, by factoring the operator polynomials of the two models and studying their roots, we will find that the models are actually quite similar.

475-10 Theoretical ARMA

Chapter 480

Linear Programming

Introduction

Linear programming maximizes a linear objective function subject to one or more constraints. The technique finds broad use in operations research and is included here because it is occasionally of use in statistical work.

The mathematical representation of the linear programming (LP) problem is

Maximize

$$\mathrm{Z} = \mathrm{C}_1 X_1 + \mathrm{C}_2 X_2 + \cdots + \mathrm{C}_\mathrm{n} X_n$$

Subject to

$$X_1 \geq 0,\ X_1 \geq 0, \cdots,\ X_n \geq 0$$

$$\mathrm{a}_{i1} X_1 + \mathrm{a}_{i2} X_2 + \cdots + \mathrm{a}_{in} X_n \ \{\leq, =, \geq\}\ b_i \geq 0 \quad i = 1, \cdots, m$$

The X's are called *decision variables* (the unknowns), the first equation is called the *objective function* and the *m* inequalities (and equalities) are called *constraints*. The b_i's are often called *right-hand sides* (RHS).

The *simplex* algorithm, which solves this problem, was discovered by George Dantzig in 1947. We use a modified version of the revised simplex algorithm given by Press, Teukolsky, Vetterling, and Flannery (1992).

Example

We will solve the following problem using **NCSS**:

Maximize Z=X1+X2+2X3-2X4

subject to

$$\begin{aligned} \mathrm{X1 + 2X3} &\leq 700 \\ \mathrm{2X2 - 8X4} &\leq 0 \\ \mathrm{X2 - 2X3 + X4} &\geq 1 \\ \mathrm{X1 + X2 + X3 + X4} &= 10 \end{aligned}$$

The solution is X1 = 9, X2 = 0.8, X3 = 0, and X4 = 0.2 which results in Z = 9.4.

Data Structure

This technique requires a special data format. The coefficients of the object function are stored in one (usually the first) row. The constraints are stored one to a row. The type of constraint (less than, greater than, or equal to) is stored in a column. Following is an example of how to store the above example in an ***NCSS*** database. This particular database is called LP.S0.

LP dataset

X1	X2	X3	X4	Logic	RHS
1	1	2	-2	O	
1		2		LT	700
	2		-8	LT	0
	1	-2	1	GT	1
1	1	1	1	EQ	10

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Constraint (A) Variables

Specify the variables containing the A matrix (the matrix of coefficients). Each variable will be solved for during the running of the program. The coefficients can be either positive or negative. If a particular coefficient is zero, you may enter a zero or leave the value blank. Usually the objective function coefficients are entered as the first row.

Logic Variable

Specify the variable containing the logic values for the constraints. Use ‘LE’ for less than or equal, ‘EQ’ for equal, ‘GE’ for greater than or equal, and ‘O’ to designate the row containing the coefficients of the objective function.

Bounds (R.H.S.) Variable

Specify the variable that contains the right-hand sides (the b’s) for each constraint.

Zero

Zero

Because of the possibility of rounding error, a small value must be specified below which, all values will be treated as zero by the algorithm.

Reports Tab

Select Reports

Initial Tableau Report - Final Tableau Report

Indicate which reports you want to view.

Report Options

Variable Names

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

Precision

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

Report Options – Decimal Places

Initial Tableau - Final Tableau

This option lets you designate the number of decimal places to be displayed on each report.

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 – Linear Programming

This section presents an example of how to run the data presented in the example given above. The data are contained in the LP database.

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

1 Open the LP dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **LP.S0**.
- Click **Open**.

2 Open the Linear Programming window.

- On the menus, select **Analysis**, then **Operations Research**, then **Linear Programming**. The Linear Programming 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 Programming window, select the **Variables tab**.
- Double-click in the **Constraint (A) Variables** text box. This will bring up the variable selection window.
- Select variables **X1, X2, X3,** and **X4** from the list of variables and then click **Ok**. “X1-X4” will appear in this box.
- Double-click in the **Logic Variable** text box. This will bring up the variable selection window.
- Select **Logic** as the Logic Variable since this variable contains the logical sign of each constraint.
- Double-click in the **Bounds (R.H.S.) Variable** text box. This will bring up the variable selection window.
- Select **RHS** as the Bounds (R.H.S.) Variable since this variable contains the logical sign of each constraint.

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).

Initial Tableau Section

```
Initial Tableau Section

Row           X1        X2          X3          X4         RHS
1 Obj Fn  1.0000    1.0000      2.0000     -2.0000      0.0000
2 <=      1.0000    0.0000      2.0000      0.0000    700.0000
3 <=      0.0000    2.0000      0.0000     -8.0000      0.0000
4 >=      0.0000    1.0000     -2.0000      1.0000      1.0000
5 =       1.0000    1.0000      1.0000      1.0000     10.0000
```

This report lists the initial values so you can double check the input.

Optimal Solution Section

```
Optimal Solution Section

               Optimal    Original        Reduced
Variable         Value        Cost           Cost          Status
X1              9.0000      1.0000         0.0000           Basis
X2              0.8000      1.0000         0.0000           Basis
X3              0.0000      2.0000        -0.2000       Non Basis
X4              0.2000     -2.0000         0.0000           Basis
Obj. Fn.        9.4000
```

This report presents the solution. It shows the optimal value of each variable.

Variable

The variables that are being solved for.

Optimal Value

The values of the independent variables that results in a maximum value of the objective function. The maximum value of the objective function is given as the last line of the report.

Original Cost

These are the values of the coefficients of the objective functions. These are the C's.

Reduced Cost

The reduced costs are an additional output of the simplex method.

Status

This column gives the status of each independent variable in final solution. The solution is found by ignoring some variables (setting their values to zeros). When a variable is ignored, it is said to be a "non basis" variable. When a variable is not ignored, it is said to be a "basis" variable.

Constraint Section

```
Constraint Section

Row                                 Optimal
No.    Type            RHS              RHS    Constraint
2        <=       700.0000           9.0000    X1+2X3
3        <=         0.0000           0.0000    2X2-8X4
4        >=         1.0000           1.0000    X2-2X3+X4
5         =        10.0000          10.0000    X1+X2+X3+X4
```

This report presents an analysis of each constraint when the variables are set to their optimal values.

Row No.

The row of the database from which this constraint comes.

Type

The type of constraint that this row represents.

RHS

The original value of the right-hand side of the constraint.

Optimal RHS

The value of this constraint at the optimal solution.

Constraint

The first forty characters of the constraint.

Final Tableau Section

```
Final Tableau Section

Variables       X3       Slack2        Art1      Slack3         RHS
Z          -0.2000      -0.3000      1.0000     -0.6000      9.4000
Slack1     -1.0000       0.0000      1.0000     -1.0000    691.0000
X2         -1.6000       0.1000      0.0000     -0.8000      0.8000
X4         -0.4000      -0.1000      0.0000     -0.2000      0.2000
X1          3.0000       0.0000     -1.0000      1.0000      9.0000
```

This report presents the final values of the simplex tableau. The variables listed down the left side are the basis variables. These are the variables that are active in the solution. The variables listed across the top are the non-basis variables. These variables were not in the solution.

A slack variable is generated for each inequality constraint. An artificial variable is generated for each equality constraint. The values in the RHS column are the solution values.

Chapter 485

Appraisal Ratios

Introduction

An appraisal (or sales) ratio study measures the accuracy and equitability of mass appraisals by local government agencies. It compares the appraised value to the market value—measured by sales price—by creating the ratio of the two. These ratios, computed for each parcel that is sold, are compared across geographic areas (such as neighborhoods) to assess their level (central tendency) and uniformity (variation).

This topic is discussed in detail in Eckert (1990) and we refer you to that text for more details.

Data Structure

At the minimum, two variables are necessary to construct a ratio—the numerator and the denominator. In this case, the numerator is the appraised price of the property and the denominator is the sales price of the property. Other information is usually necessary such as the sales date, a geographic location such as neighborhood, and a property class. Note that the appraised price may be the total of two or more variables such as the price of the land and the price of the structure (home or building).

The dataset SALESRATIO.S0 contains an example of such a database. This database contains 360 rows, of which only five are displayed here. The sales price of the property is given in dollars. The sales date is in the format MMDDYY. Note that ***NCSS*** automatically handles the Year 2000 problem caused by the use of a two-digit year. The appraised value of the land is given in Land and that of the structure is given in Building. Three property classes are found on this database: 510, 511, and 512. Several neighborhoods are represented by an identification number.

SALESRATIO dataset (subset)

SaleDate	SalePrice	PClass	Land	Building	Neighborhood	YearSold
51195	0	512	3110	39200	1720	95
62096	250	512	5910	27370	1720	96
21696	628	511	12630	43710	0	96
41696	1500	512	11000	54740	0	96
52595	1900	512	10890	17830	0	95
32295	2000	511	11000	73710	0	95
90496	2000	512	0	5800	708	96
52595	2100	512	10890	21110	0	95
53195	2500	511	4890	72800	1455	95
12496	3000	511	11540	70250	0	96

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Appraisal Variables

The one or more variables listed here are summed to form the market valuation of the property. You might include land and building values. This value becomes the numerator of the sales ratio. It is assumed that all appraised values are current—that is, they represent the appraised value of the property today.

Sales Price Variables

These variables are summed to form the sales price. Usually, only one variable is specified. This value becomes the denominator of the sales ratio.

Since property sales may have occurred over several years, some adjustment of the sales price may be necessary. An automatic percentage increase adjustment is available in the Adjustment field below.

Break Variable

A separate line on the output report is created for each unique value of this variable. For example, you might want to compare the sales ratios by neighborhood.

Include Variable

This optional variable lets you restrict the analysis to certain rows of your database by listing the values to be included in the report. It works like a Filter variable.

For example, you might have several property types on your database and a variable that identifies the property type of each sale (row). Suppose you want to restrict the analysis to a few property classes. Specify the property class variable here and the values to be included in the Included Values option.

Included Values

These are the values of the Included Variable that are used in the analysis. Rows with values other than these are skipped. You may enter a single value or a comma-delimited list.

Min and Max Ratio

Min - Max Ratio Kept

The typical sales ratio database includes many records that are for specialized sales in which the listed sales price is not an accurate representation of the true market value of the parcel. An example would be sales among family members. Every effort should be taken to remove specialized sales from your database. However, this may not always be possible. This option lets you automatically omit records with sales ratios that are way out of bounds. The hope is that most of these are non-representative, specialized sales.

Specify a minimum and maximum ratio value here, using the metric defined by the Ratio Multiplier value. Only those properties with ratios in this range will be used. If the Ratio Multiplier is set to 100, typical values are Min Ratio equal to 50 and Max Ratio equal to 150. If the Ratio Multiplier is set to 1.0, typical values are Min Ratio equal to 0.5 and Max Ratio equal to 1.5.

Reports Tab

Select Reports

Count – Display COV x 100

These options let you indicate which items are displayed on the report. If more than five items are checked, you will need to put your printer into Landscape mode by selecting Printer Setup from the File Menu.

Star Items Based On Normality

This option causes certain items to be starred with an asterisk based on the results of the normality test. When checked and the normality test is rejected, the median, median confidence interval, and COD starred. When checked and the normality test is not rejected, the mean, mean confidence interval, and COV are starred.

Report Options

Normality Test Alpha

This option specifies the rejection probability for the normality test. If the normality test probability is less than this amount, reject the null hypothesis of normality. If the normality test probability is greater than this amount, there is not enough evidence in the data to reject the hypothesis of normality.

The possible range of this probability is 0.01 to 0.30. Most statisticians recommend testing normality at an alpha level higher than 0.05 such as 0.10, 0.15, or 0.20.

Confidence Coefficient

This option specifies the value of confidence coefficient for the confidence intervals of the mean and median. The possible range is from 0.800 to 0.999. The standard value is 0.95.

Ratio Multiplier

Sales Ratios are usually multiplied by 100 to scale them as a percentage. At other times, you may not wish to scale them. This option lets you specify the multiplier. Use 100 for percentages and 1 for regular ratios.

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.

Justification

This option specifies whether the data and labels in each column should be right or left justified.

Page Title

This option specifies a title that will be printed at the top of each page of the report.

Report Options – Decimal Places

Count - COD

These options let you specify the number of decimal places shown in the various items of the output report.

Report Options – Setup the ruler for Outputting a Line of the Report

These options let you specify the tab settings across the table. The output ruler is also modified by the settings of Justification.

First Tab

Specifies the position of the first cell in inches. Note that the left-hand label always begins at 0.5 inches. Hence, the distance between this tab and 0.5 is the width provided for the row label information.

Right Border

Specifies the right border of the table. The number of tabs is determined based on *First Tab*, the *Tab Increment*, and this option. If you set this value too large, your table may not be printed correctly.

Column Width

Specifies the width of a column in inches. We recommend a value of 0.6 when a lot of items are to be displayed.

Sale Date Tab

Sale Date Specification

Sale Date Variable

This optional variable contains the dates on which the sale took place. These date values may be used to select a range of dates for the report. For example, you may want to include only sales that occurred after 1994. These dates may also be used to adjust the sales price using the Adjustment value.

The date values may be in any one of several date formats. The date format is specified in the Date Format option.

Date Format

Specify the format of the date variable. Note that Y is for year, M is for month, and D is for day. The date format used must match the format used in the Min Date and Max Date. All date values are converted to Julian dates (number of days since January 1, 0000), so there is no year 2000 conversion problem.

Speaking of year 2000 conversion, two-digit year dates are converted to Julian dates using the conversion rule that is set as a system option. If the conversion factor is 30, then year values of 30 or more are converted to 1930 (or more) and year values 29 and less are set to 2029 (or less). For example, the year part of the date 010101 is assumed to be 2001; the year part of the date 010129 is assumed to be 2029; and the year part of the date 010130 is assumed to be 2030.

Minimum - Maximum Date Kept

Specify either the minimum date, the maximum date, or both. The format must match that specified in the Date Format box above. Only properties in the date range specified here will be used. Leave this option blank if you do not want to make a selection based on date.

Optional Price Adjustment

Adjust Prices To This Date

This is the date to which all sales price values are adjusted. Often this is January of the current year. The format is YYYY-MM. For example, you might enter 1999-01 to adjust all sales values to January 1, 1999.

Monthly Price Adjustment Factor

This is the monthly sales price adjustment factor. Each sales price is adjusted using the formula Adjusted Sales Price = (Actual Sales Price) $(1+a\ m)$ where m is the number of months between the actual sale date and the adjusted date and a is the factor entered here.

For example, suppose that a sales took place in August of 1997 and that you have set the Current Date to 1998-07 and the Adjustment value to 0.001. Since there are twelve months between from August to July, the sales value would be adjusted up using the factor (1+0.001*12) which is 1.012. That is, the sales price would be increased by 1.2 percent.

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 – Sales Ratio Study

This section presents a tutorial of a sales ratio study conducted on the SALESRATIO database. The analyst wants to limit the analysis to sales that occurred on or after January 1, 1994 and to property classes 510, 511, and 512.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Appraisal Ratios window.

1 Open the SALESRATIO dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **SALESRATIO.S0**.
- Click **Open**.

2 Open the Appraisal Ratios window.

- On the menus, select **Analysis**, then **Mass Appraisal**, then **Appraisal Ratios**. The Appraisal Ratios 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 Appraisal Ratios window, select the **Variables tab**.
- Set the **Appraisal Variables** box to **LAND-BUILDING**.
- Set the **Sales Price Variables** box to **SalePrice**.
- Double-click in the **Break Variable** text box. This will bring up the variable selection window.
- Set the **Break Variable** box to **Neighborhood**.
- Set the **Include Variable** box to **PClass**.
- Set the **Include Values** box to **510 511 512**.

4 Specify the dates.

- On the Appraisal Ratios window, select the **Sale Date tab**.
- Set the **Sale Date Variable** to **SaleDate**.
- Set the **Date Format** box to **MMDDYY**.
- Set the **Minimum Date Kept** to **010194**.

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).

Ratio Section

```
Ratio Section

Neighborhood      Count     Median      Mean    W Mean       PRD       COD
0                   190      83.62     88.96     85.56      1.04     18.00
110                   2      59.94     59.94     59.50      1.01      6.84
125                   4      63.21     62.50     62.16      1.01      8.60
.                     .          .         .         .         .         .
.                     .          .         .         .         .         .
.                     .          .         .         .         .         .
Total               275      83.76     88.71     85.39      1.04     18.37
```

This report gives various statistical values for each value of the break variable (which was Neighborhood in this example). Note the total line at the bottom of the report. This gives each statistical measure treating the whole database as a single group.

Neighborhood

This gives the value of the break variable. Note the total line at the bottom of the report.

Count

This is the number of rows included in the statistics reported on this line. The main use of this statistic is to be certain that a large enough sample was available to make the other statistics reliable. At a minimum, you would want at least thirty sales ratios included before you could have any faith in the accuracy and repeatability of the statistics.

Median

This is the median (middle) sales ratio. The individual sales ratios (R) are found using the formula

$$R = \frac{A}{S}$$

where *A* is the appraised value of the property and *S* is the sales price of the property. These values are ranked and the middle value is selected as the median.

The median is usually used as the measure of central tendency of sales ratios because of its resistance to distortion by outliers. Since outliers occur frequently in property sales databases, and since outliers often distort statistics that include them, their impact on the statistics must be realized and removed. Using the median is a good way to accomplish this.

The medians are scanned to determine if there are any groups (neighborhoods) that are way undervalued or way overvalued.

LCL and UCL Median

These are the confidence limits of the median sales ratio. These limits make no assumption of normality.

Mean

This is the averaged sales ratio. The individual sales ratios computed and averaged to calculate this value.

Unlike the median, the mean is easily distorted by outliers. For this reason, care should be exercised when using it. One possible use is to compare the mean with the median to determine if there were a lot of outliers present. The difference between the two is partially due to the presence of outliers.

LCL and UCL Mean

These are the confidence limits of the mean sales ratio.

W Mean

The weighted mean is the ratio of the total appraised values for the entire sample and the total sales prices of the entire sample. Hence, the formula is

$$WM = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} S_i}$$

The weighted mean weights each ratio by the sales price. Hence, high priced properties carry a larger weight than low priced properties. It is most appropriate for measuring the central tendency when you are most interested in total dollar value of the sample.

PRD

This is the price related differential. It measures the regressivity or progressivity of the assessments. Regressive appraisals occur when high-value properties are underappraised relative to low-value properties. Progressive appraisals occur when the opposite pattern occurs.

This statistic is the ratio of the mean and the weight mean. It is calculated using the formula

$$PRD = \frac{\sum_{i=1}^{n} \frac{A_i}{nS_i}}{WM}$$

A PRD greater than 1.0 indicates that high-value properties are underappraised, while a value less than 1.0 indicates that low-value properties are underappraised. As a general rule, except for computations involving small sample sizes, each PRD should be between 0.98 and 1.03.

COD

This is the coefficient of dispersion is usually used as the measure of uniformity in ratio studies. It is calculated using the formula

$$COD = \frac{\frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_i}{S_i} - Median\left(\frac{A_i}{S_i}\right)\right|}{Median\left(\frac{A_i}{S_i}\right)}$$

COD values of 15.0 or less tend to be associated with good appraisal uniformity.

COV

This is the coefficient of variation. See the Descriptive Statistics chapter for details.

Std. Dev.

This is the standard deviation of the sales-ratio values. See the Descriptive Statistics chapter for details.

Normality Prob

This is the probability level of the Shapiro-Wilk test of normality. If this value is less than some cutoff value, often 0.05 or 0.10, the hypothesis that the sales ratios are normally distributed is rejected.

Chapter 486

Comparables – Sales Price

Introduction

Appraisers often estimate the market value (current sales price) of a *subject* property from a group of *comparable* properties that have recently sold. Since sales data are considered the best evidence of market value, this is the preferred approach to market value estimation when sales data are available. This topic is discussed in detail in Eckert (1990) and we refer you to that text for more details.

This module allows you to select an appropriate subset of your sales database, such as properties in the same area of similar age and size, and specify a set of adjustment variables and their corresponding weights and adjustment values. The selected data are scanned to determine those properties (rows of the database) that are most comparable to the subject property. These comparable properties are then used to create an estimate of the market value (sales price) of the subject property.

Technical Details

The market value of subject property is estimated by adjusting the sales prices of comparable properties so that their attributes match those of the subject property. This adjustment may be a dollar amount, such as $50 per square foot, or a percentage, such as 1.5% decrease for each year since the property was constructed.

This sales comparison method has three general phases:

1. Select a pool of possible properties from a database of recent sales.
2. Rank these properties according to how close they are to the subject property.
3. Adjust the comparable properties so that their attributes match the subject property.

Phase I – Data Selection

The first step is to select a group of properties to work with from the database of sales data. Usually, this step involves selecting properties of similar age, size, and location. The program limits the search for comparables to properties that meet these selection criterion.

We cannot stress to much the importance of finding comparables that are as similar as possible to the subject. This greatly reduces the need for making many (sometimes controversial) adjustments to the sales prices.

Phase 2 – Rank the Properties

A distance measure is used quantify how close each comparable property is to the subject property. Suppose there are K attributes $X_1, X_2, \ldots, X_K$ on which the distance is to be measured. The value of the i^{th} attribute on the j^{th} comparable property is represented by X_{ij}. The distance between the j^{th} comparable property and the subject property is calculated using the Euclidean Distance formula:

$$D_j = \sqrt{\frac{1}{K}\sum_{i=1}^{K} w_i \frac{\left(X_{ij} - X_{is}\right)^2}{S_i}}$$

where S_i is the standard deviation of the X_{ij} for a particular attribute and w_i is an attribute importance weight scaled so that they sum to one.

This formula reduces all variables to unit-less index values by dividing each by its standard deviation. This allows us to combine the number of bedrooms (a small number) with the number of square feet (a relatively large number) in one formula. The differences are squared to put negative and positive values on an equal basis.

Note that if all attributes match, the distance will be zero. Typical values will be between zero and five.

Once these distances are calculated, they are sorted from lowest to highest. The properties with the smallest distances are closest to the subject property.

Phase 3 – Adjusting the Sales Price

Finally, an adjusted sales price is computed for each comparable. The magnitude of the adjustment depends on how well the property matches the subject property. There are three steps in this adjustment process:

Step 1 – Sales date adjustment

The sales price of each comparable property is first adjusted to a specified point in time using a monthly percentage adjustment.

For example, suppose the percentage adjustment is set at 0.2% per month, a comparable property sold for $100,000 in August of 1998, and the property as to be adjusted to August of 2000. The time adjusted sales price would be calculated as:

$$\$100{,}000\left(1 + 0.002(24)\right) = \$104{,}800$$

Step 2 – Dollar and percentage adjustments

The sales price of each comparable property next adjusted for differences in other attributes. The adjustments are either dollar (lump sum) or percentage adjustments. The adjustments are made in the same order that they are specified. Hence, if you want to make the percentage adjustments first, you should specify them first.

As an example of a dollar adjustment, suppose that a comparable property has 2,500 square feet while the subject property has 3,000 square feet. Obviously, the value of the comparable property must be adjusted up. The appraiser must set a dollar amount of adjustment for each unit difference. In this example, suppose the appraiser decides to add $50 per square foot. Since the difference in size is 500 square feet, $25,000 ($50 x 500) is added to the sales price of the comparable property.

As an example of a percentage adjustment, suppose that a comparable property has a quality rating of one while the subject property has a quality rating of three. Obviously, the value of the comparable property must be adjusted up so that it is on par with the subject property. The appraiser must set a percentage adjustment for each unit difference. In this example, suppose the appraiser decides to increase the property value by 2% per one unit difference in quality. Since the difference in quality is two units, the current adjusted sales price is multiplied by 1.04.

Step 3 – Estimate the sales price

Once the adjusted sales prices of the comparables have been calculated, the sales price of the subject property can be calculated. NCSS can calculate four different estimates:

1. **Closest**. The adjusted sales price of the closest comparable (using the Euclidean distance) is used as the estimate of the subject sales price.
2. **Least Absolute Dollar Change**. The adjusted sales price of the property that had the least amount of adjustments in absolute dollar amounts is used as the estimate of the subject sales price.
3. **Simple Average**. The average adjusted sales price of the closest four or five properties is used as the estimate of the subject sales price. It is hoped that averaging will help remove the influence of any anomalies that might occur with a single sale.
4. **Weighted Average**. A weighted average of the adjusted sales price of the closest four or five properties is used to estimate the subject sales price. The weights are based on the distances between the subject property and comparable property. Specifically, the weights are calculated using the formula:

$$W_j = 100\left(\frac{5 - D_j}{5}\right)$$

   with negative weights being reset to zero.

Data Structure

Each column of the spreadsheet (database) represents a variable and each row represents a property. The selection variables can be text or numeric, but the sales adjustment variables must be numeric. You must include a sales price variable and at least one adjustment variable.

The dataset Comparables.S0 contains an example of such a database. This database contains fifty-one rows, of which only a few are displayed here. There are also other variables that are not displayed here.

Note that the sales date is in the format YYYYMM. Also, the indicator variable, Subject, at the right-side of the database indicates which rows are to be treated as subjects (no-blank) or as comparables (blank).

COMPARABLES dataset (subset)

PropID	Neighborhood	SalePrice	SaleDate	SqFt	LotSize	Subjects
A-1	AAA	71589	199801	1165	4670	
A-2	AAA	50535	199907	735	3805	
A-3	AAA	134644	199902	2488	5249	
A-4	AAA	156865	199903	3149	4394	
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Subject1			199906	965	5502	1

Procedure Options

This section describes the options available in this procedure.

Selection Tab

Specify the variables used to select the rows of the database that are to be used as possible comparables.

Indicator Variable

Indicator Variable

This variable indicates which rows from the database are subject properties. A separate comparables report will be generated for each row with a non-blank value. These rows will not be candidates as comparables.

Normally, you would enter the information for each subject property at the bottom of the database, set the value of this variable to ‘1’ in that row, and run the analysis.

Note that this variable MUST be specified.

Report Specification

Report Variables

Specify additional variables to be displayed at the beginning of the Comparables Report.

Label Variable

Specify a variable used to label the rows of the reports. If this variable is omitted, the database row numbers are used.

Selection Variables and Ranges

Selection Variables

Specify variables whose values will be used to select specific records (rows or properties) from the database. Only those rows that meet the selection criterion given in the corresponding Selection Range box will be kept.

Only specify one variable per box.

Note that you do not have to specify any variables, in which case, the whole database will be used.

Selection Range

Specify a set of values used to specify which rows from the database are kept in the analysis. Note that all text values are changed to upper case before processing.

You can mixe the following types of entries together:

- **Value**

  You can enter a single value, either text or numeric. You do not have to enclose text in quote marks. For example, you might enter:

  AAA

  or

  1

- **List**

  You can enter one or more values in a comma-separated list. For example, you might enter:

  102,104,106

  or

  ABC,DEF,KJS

- **Minimum to Maximum**

  You can enter a minimum value and a maximum value, separated by the word ‘TO’. Only those rows whose values fall within the range are kept. The syntax for this command is:

  Minimum to Maximum

  For example, you might enter:

  100 to 500

  A TO C

  -900 To 1000, 1200, 1500, 1700 to 1900

Adjustment Tab

These options let you specify the attribute variables used to adjust the sales price of the comparable properties. These are the variables used to calculate the distance from a subject to a comparable.

Adjustment Specification

Adjustment Variables

Specify the adjustment variables in these boxes. Two types of adjustment variables may be specified: lump sum and percentage.

These variables are used in two phases of the analysis:

1. Similarity Phase. Calculate an index of how similar each candidate property is to the subject property.
2. Appraisal Phase. Calculate an adjusted sales price based on the difference between the candidate property and the subject property.

During the appraisal phase, the variables are applied in the order in which you have specified them. You should note that the order of these variables does change the estimated sales price because of the percentage variables. Since the percentage adjustment is applied to the current estimated sales price, the position of the variable in the list becomes import. Some schools of thought recommend that all percentage variables should be applied first. Others recommend that they should be applied after the lump sum adjustments are made. NCSS lets you decide. You could even put some at the beginning and some at the end.

Distance Weight

This value specifies the relative importance of this variable in calculating the Euclidean distance between the subject property and the candidate properties. You should specify a positive number for each adjustment variable.

Note that the weights are not used in adjusting the sales prices of the candidate properties.

One way to specify these weights is to set the least important variables weight to one. Then, other weights are specified relative to that. For example, suppose you decide to use three adjustment variables: Size, Age, and Number of Bedrooms. You decide that the least important variable is the Number of Bedrooms and assign that variable a weight of ‘1.’ You decide that Age is three times as important as bedrooms, so you assign Age a weight of ‘3.’ You decide that Size is four and one-half times as important, so you assign Size a weight of ‘4.5.’ NCSS totals these three weights, (which is 8.5) and divides each of your entries by that total. Thus, the actual weight for Number of Bedrooms is 1/8.5 = 0.1176, for Age is 3/8.5 = 0.3529, and for Size is 0.5882.

Although the actual weights for all variables must sum to one, you do not need to worry about that. In fact, your numbers can all be greater than one. The program will readjust your weights so that they sum to one.

Amount ($ or %)

This value specifies the amount that the sales price is changed for each unit change in the Adjustment Variable. You may use a percent sign to indicate a percentage or a dollar sign to indicate a dollar amount. If no sign is included, a dollar amount is assumed.

If you include a percent sign in your value, the adjustment will be applied as a percentage rather than an amount. For example, if you enter 2% here and the difference between the candidate property is 2 units less than the subject property on this variable, the overall sales price is increased by 4%.

As an example of a dollar adjustment, suppose that the Adjustment Variable is Size in square feet. Suppose that this Adjustment Amount is set at 30. That means that each additional square foot in size of the candidate property over that of the subject property will cause the estimated sales price of the candidate property to be decreased by $30.

Date

Sale Date Variable

This variable contains the date on which the sale took place. These date values are used to adjust the sales prices to the same point in time using a multiplicative adjustment (Date Adjustment).

The date values may be in any one of several date formats. The date format is specified in the Date Format option.

Date Format

Specify the format of the Sale Date Variable. This is the format of the dates that appear on your database.

Note that Y is for year, M is for month, and D is for day. All date values are converted to Julian dates (number of days since January 1, 0000), so there is no year 2000 conversion problem.

Speaking of year 2000 conversion, two-digit year dates are converted to Julian dates using the conversion rule that is set as a system option. If the conversion factor is 30, then year values of 30 or more are converted to 1930 (or more) and year values 29 and less are set to 2029 (or less). For example, the year part of the date 010101 is assumed to be 2001; the year part of the date 010129 is assumed to be 2029; and the year part of the date 010130 is assumed to be 2030.

Date Adjustment

This is the monthly sales price adjustment factor. Each sales price is adjusted using the formula Adjusted Sales Price = (Actual Sales Price) (1 + (a)(m)) where *m* is the number of months between the actual sale date and the adjusted date and *a* is the number entered here.

Note that this number is NOT a percentage!

For example, suppose that a sales took place in August of 1997 and that you have set the Current Date to 1998-07 and the Adjustment value to 0.001. Since there are twelve months between from August to July, the sales value would be adjusted up using the factor (1+0.001*12) which is 1.012. That is, the sales price would be increased by 1.2 percent.

Current Date

Specify the date to which all sales price values are adjusted. Often this is the current date.

The format is YYYY-MM. For example, you might enter ‘2001-01’ to adjust all sales values to January 1, 2001.

Sales Price

Sales Price Variable

Specify the variable that contains the sales prices (in dollars) of the comparable properties. Only those properties that have a positive sales price are used.

Reports Tab

Select Reports

Distance Report - Settings Report

Specify whether to display the report.

Show Dollar Signs

Specify whether to show dollar signs on the reports. The dollar sign may clutter the report, so it is optional.

Show Commas in Dollars

Specify whether to formal dollar amounts with a comma. The comma may need to be left out for formatting reasons.

Show Column Separator = |

The report may be formatted with or without the vertical bar. This option lets you decide whether you want to use it or not.

Report Options

Estimation Method

Specify the method to be used to calculate the estimated sales price of the subject property. We recommend that you use the Weighted Average method unless you have a good reason to use one of the other methods. The following methods are available:

- **Closest**

  Use the sales price of the comparable property that is closest (has the minimum distance) to the subject property.

- **Min |$| Change**

  Use the sales price of the comparable property that had the smallest absolute dollar value change. That is, this property was adjusted by the smallest dollar amount.

- **Simple Average**

  Use the average of the adjusted sales prices of the *M* comparables that have the smallest distance. The value of *M* is set on the Reports tab by the Properties in Averages option.

- **Weighted Average**

  Use the weighted average of the adjusted sales prices of the *M* comparables that have the smallest distance to the subject property. The weights are proportional to the distance from the comparable property to the subject property.

Report Options – Specify Number of Properties

Per Distance Report

Specify the number of properties (rows) displayed on the Distance report. Note that the database is sorted so that only the candidates closest to the subject are displayed. We recommend that you look at the best ten or twenty candidates.

Per Comparables Report

Specify the number of properties displayed on this report. Usually, three two five are used, but ten are displayed.

Note that the database is sorted so that only the candidates closest to the subject are displayed.

Per Page

Specify the number of comparables shown on each page of the comparables report. Each property requires two columns of the report. The maximum is four.

Used in Averages

Specify the number of properties in the averages used to estimate the subject's sales price. The simple average of the most similar properties (determined by the comparability index) is one sales price estimate. A weighted average of these same properties is another sales price estimate.

Report Options – Decimal Places

Dollars - Means

Specify the number of decimal places shown on the report for these items.

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 – Comparables Study

This section presents a tutorial of a comparables study conducted on the COMPARABLES database. The appraiser limits the analysis to those properties in neighborhood ‘AAA’ that were constructed from 1970 to 1980.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Comparables – Sales Price window.

1 Open the COMPARABLES dataset.

- From the **File** menu of the NCSS Data window, select **Open**.
- Select the **Data** subdirectory of your NCSS directory.
- Click on the file **COMPARABLES.S0**.
- Click **Open**.

2 Open the Comparables – Sales Price window.

- On the menus, select **Analysis**, then **Mass Appraisal**, then **Comparables – Sales Price**. The Comparables – Sales Price procedure will be displayed.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the selection variables.

- On the Comparables – Sales Price window, select the **Selection tab**.
- Double-click in the **Indicator Variable** text box. This will bring up the variable selection window.
- Select **Subjects** from the list of variables and then click **Ok**. “SUBJECTS” will appear in this box.
- Double-click in the **Report Variables** text box. This will bring up the variable selection window.
- Select **YearSold** and then click **Ok**. “YEARSOLD” will appear in this box.
- Double-click in the **Label Variable** text box. This will bring up the variable selection window.
- Select **PropId** and then click **Ok**. “PROPID” will appear in this box.
- Double-click in the first **Selection Variable** text box. This will bring up the variable selection window.
- Select **Neighborhood** and then click **Ok**. “NEIGHBORHOOD” will appear in this box.
- Double-click in the second **Selection Variable** text box. This will bring up the variable selection window.
- Select **YearBuild** and then click **Ok**. “YEARBUILD” will appear in this box.
- Enter **AAA** in the **first Selection Range** box.
- Enter **1970 to 1980** in the **second Selection Range** box.

4 Specify the adjustment variables.

- On the Comparables – Sales Price window, select the **Adjustment tab**.
- Double-click in the first **Adjustment Variable** text box. This will bring up the variable selection window.
- Select **Quality** and click **Ok**. “QUALITY” will appear in this box.
- Double-click in the second **Adjustment Variable** text box. This will bring up the variable selection window.
- Select **SQFT** and click **Ok**. “SQFT” will appear in this box.

- Double-click in the third **Adjustment Variable** text box. This will bring up the variable selection window.
- Select **LotSize** and click **Ok**. “LOTSIZE” will appear in this box.
- Double-click in the fourth **Adjustment Variable** text box. This will bring up the variable selection window.
- Select **Bedrooms** and click **Ok**. “BEDROOMS” will appear in this box.
- Set the first **Distance Weight** to **1**.
- Set the second **Distance Weight** to **4**.
- Set the third **Distance Weight** to **2**.
- Set the fourth **Distance Weight** to **1**.
- Set the first **Amount** to **3%**.
- Set the second **Amount** to **$50**.
- Set the third **Amount** to **$5**.
- Set the first **Amount** to **$400**.
- Double-click in the **Sale Date Variable** text box. This will bring up the variable selection window.
- Select **SaleDate** and click **Ok**. “SALEDATE” will appear in this box.
- Double-click in the **Sales Price Variable** text box. This will bring up the variable selection window.
- Select **SalePrice** and click **Ok**. “SALEPRICE” will appear in this box.
- Set the **Date Format** to **YYYYMM**.
- Set the **Date Adjustment** to **0.01**.
- Set the **Current Date** to **1999-10**.

5 Specify the reports.

- On the Comparables – Sales Price window, select the **Reports tab**.
- Check **Distance Report**.
- Check **Comparables Report**.
- Check **Settings Report**.
- Uncheck **Show Dollar Signs**.
- Uncheck **Show Commas in Dollars**.
- Check **Show Column Separator = |**.
- Set **Estimation Method** to **Weighted Average**.
- Set **Per Distance Report** to **10**.
- Set **Per Comparables Report** to **8**.
- Set **Per Page** to **4**.
- Set **Used in Averages** to **5**.
- Set **Dollars Decimals** to **0**.
- Set **Distances Decimals** to **3**.
- Set **Percents Decimals** to **0**.
- Set **Means Decimals** to **1**.

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).

Settings Report

```
Settings Report for Subject = Subject1

Selection            Selection
Variables            Criterion
Neighborhood         AAA
YearBuilt            1970 to 1980

Adjustment  Adjustment                         Standard
Variables        Value    Weight      Mean    Deviation        COV
SaleDate            1%
Quality             3%         1       2.2          0.7         33
SqFt               $50         4    1869.2       1001.3         54
LotSize             $5         2    5018.2       1153.3         23
Bedrooms          $400         1       2.5          1.2         47
Of the 50 properties on the database, 1 was a subject property and 37 were excluded
by the selection variable(s), leaving 12 comparables for consideration.
```

This report gives the settings you used to create the report. It is provided to let you document how the options where set. It also supplies summary statistics about the adjustment variables used in the analysis. The Mean, Standard Deviation, and Coefficient of Variation are computed on the rows selected for analysis.

Distance Report

```
Distance Report for Subject = Subject1

PropID      Distance    SalePrice  Quality      SqFt   LotSize     Bedrooms
Subject1                                 2       965      5502            2
A-14           0.240        88474        2      1309      6484            2
A-8            0.394        78728        2      1181      5350            4
A-1            0.484        71589        1      1165      4670            1
A-2            0.568        50535        2       735      3805            2
A-9            0.579        95660        1      1653      5715            1
A-10           0.597        50902        2       779      3745            2
A-15           1.083        60007        3       963      3506            3
A-3            1.503       134644        3      2488      5249            3
A-7            1.765       127109        3      2419      4086            3
A-13           2.866       168715        2      3202      6720            1
```

This report displays the values of the adjustment variables for the comparable properties that met the selection criterion. Note that the subject property is displayed first.

Distance

This value is the Euclidean distance, *D*, between the subject property and the comparable property. Values near zero are close. Values near five are very different.

Comparative Sales Price Report

```
Comparative Sales Price Adjustment Report for Subject = Subject1

PropID           | Subject1  | A-14            | A-8             | A-1             | A-2
                 | Value     | Value      $Adj | Value      $Adj | Value      $Adj | Value      $Adj
YearSold         |           | 1997            | 1996            | 1995            | 1996
Comparability    |           | 92%             | 87%             | 84%             | 81%
SalePrice        |           | 88474      6193 | 78728      3149 | 71589     15034 | 50535      1516
Quality          | 2         | 2             0 | 2             0 | 1          2599 | 2             0
SqFt             | 965       | 1309     -17200 | 1181     -10800 | 1165     -10000 | 735       11500
LotSize          | 5502      | 6484      -4910 | 5350        760 | 4670       4160 | 3805       8485
Bedrooms         | 2         | 2             0 | 4          -800 | 1           400 | 2             0
Net $Adj.        |           |          -15917 |           -7691 |           12192 |           21501
Sum |$Adj.|      |           |           22110 |           12360 |           17159 |           19985
Adj Sales Price  | 73197     |           72557 |           71037 |           83781 |           72036

Comparative Sales Price Adjustment Report for Subject = Subject1 (Continued)

PropID           | Subject1  | A-9             | A-10            | A-15            | A-3
                 | Value     | Value      $Adj | Value      $Adj | Value      $Adj | Value      $Adj
YearSold         |           | 1995            | 1996            | 1997            | 1994
Comparability    |           | 81%             | 80%             | 64%             | 50%
SalePrice        |           | 95660      2870 | 50902      2036 | 60007      3000 | 134644    10772
Quality          | 2         | 1          2956 | 2             0 | 3         -1890 | 3         -4362
SqFt             | 965       | 1653     -34400 | 779        9300 | 963         100 | 2488     -76150
LotSize          | 5502      | 5715      -1065 | 3745       8785 | 3506       9980 | 5249       1265
Bedrooms         | 2         | 1           400 | 2             0 | 3          -400 | 3          -400
Net $Adj.        |           |          -29239 |           20121 |           10790 |          -68876
Sum |$Adj.|      |           |           38821 |           18085 |           12370 |           82177
Adj Sales Price  | 73197     |           66421 |           71023 |           70797 |           65768
```

This report displays the values of the adjustment variables for the comparable properties that met the selection criterion. Note that the subject property is displayed first.

Comparability

This is an index have how close the comparable is to the subject. It is computed using the formula:

$$W_j = 100\left(\frac{5 - D_j}{5}\right)$$

When the value is negative, it is reset to zero.

This value is used to compute the weighted average estimate of the sales price.

Net $Adj

This row gives the net change in the price between the original sales price and the final adjusted sales price.

Sum |$Adj.|

This rows totals the absolute values of the dollar adjustments that are made. It is sometimes used as an indicator of how close the comparable is to the subject property. Unfortunately, its value depends on the order in which the variables are specified.

Adj Sales Price

The adjusted sales price is given in this row. The first column provides the estimated sales price of the subject property.

486-14 Comparables – Sales Price

Chapter 487

Hybrid Appraisal Models

Introduction

This procedure analyzes a special prediction model often used in mass appraisal and assessment. The model is referred to as the *hybrid* model. Although NCSS has a nonlinear regression module for solving this model, this module was added for several reasons:

1. It provides several methods of optimization for calibrating the model parameters. Some of these methods have the same goal as the 'feedback algorithm'.
2. It streamlines model setup, including the automatic generation of binary variables.
3. It provides many new reports and statistics that aid in the calibration of the parameter values.

The use of multiple and nonlinear regression in property appraisal and assessment administration has been encouraged by such organizations as the International Association of Assessing Officers (IAAO). They publish a book by Eckert (1990) and teach courses which use these regression procedures. This program was developed to automate the hybrid model that they propose.

The Hybrid Appraisal Model

The hybrid model is a combination of both additive and multiplicative models. It relates the sales price of a property to various characteristics such as size (in square feet), lot size, construction quality, location, number of bathrooms, etc. This model computes a market value for each 'structure' of the parcel, where a structure refers to an object like the land or a build. For example, the sales price may by the sum of the market values of the building, the lot, and a garage. Because of variance that may occur because of location, time, etc., this sum may be adjusted by one or more overall variables. This overall variables act as percentage adjustments.

The general form of the model is:

$$\text{Sales Price} = \text{Overall}\,(\text{Building} + \text{Land} \; + \text{Garage} \; + \ldots)$$

Each of these factors are modeled by one or more variables from the database. These factor models are made up of three types of variables: rate, binary, and amount. The amount variable usually represents the 'size' of the structure. Examples are acreage and square footage. The rate and binary variables are variables that adjust the size variables up or down, such as a quality index or an age adjustment.

Using this construction, the individual factors are modeled as follows.

$$Overall = R_1^{B_1} R_2^{B_2} B_3^{I_3} B_4^{I_4}$$

$$Building = R_5^{B_5} R_6^{B_6} B_7^{I_7} B_8^{I_8} (B_9 A_9 + B_{10} A_{10})$$

$$Land = R_{11}^{B_{11}} R_{12}^{B_{12}} B_{13}^{I_{13}} B_{14}^{I_{14}} (B_{15} A_{15} + B_{16} A_{16})$$

$$Garage = R_{17}^{B_{17}} R_{18}^{B_{18}} B_{19}^{I_{19}} B_{20}^{I_{20}} (B_{21} A_{21} + B_{22} A_{22})$$

In this model, R_i represent rates or multipliers (usually centered near one) which modify the whole factor. Examples of rate variables include sales date multiplier and depreciation. These are analogous to percentage adjustment variables.

The I_i's represent indicator (binary) variables. These are variables that have only two values: zero and one. Examples of these are neighborhood indicators and special feature indicators (such presence of a swimming pool). Usually, qualitative variables are broken down into individual binary variables. For example, suppose that the properties in your study come from three neighborhoods. You would designate one as the standard neighborhood, then create two binary variables, one for each of the other neighborhoods. The model will then adjust for differences among the three neighborhoods.

The A_i's represent amount variables like lot size or square footage of living area. These variables are entered in the normal linear fashion.

The B_i's represent the coefficients that are estimated from the data. Often, these coefficients are constrained to lie within specified limits.

Note that the number of variables in each group varies. Quite often you may have ten or fifteen binary variables and only one rate variable.

Differential Evolution

Introduction

One of the steps in using the hybrid model in mass appraisal is to calibrate (estimate) the unknown coefficient values, the *B*'s. Often, the method of least squares (MRA) is used to estimate the coefficients. Least squares finds the set of estimates that minimize the sum of the squared errors. That is, the objective of the method of least squares is to minimum the sum of squared errors. The sum of squared errors is called the *objective function*, and the problem is to minimize it. Least squares is one method of minimizing it. Another possibility is simple trial-and-error (of course, trial-and-error may be very time consuming).

Because of the distortion that a few anomalies in the data can cause when least squares is used, other methods have been proposed. One alternative is the *feedback* algorithm. This algorithm seeks to minimize a different objective function: the *average absolute percent error*. This objective function quantifies the percentage accuracy of the model. Both expensive and inexpensive parcels are modeled to the same percentage accuracy. Least squares, on the other hand, concentrates on fitting the most expensive properties. The feedback algorithm is used to minimize an specific objective function: the *average absolute percent error*. Recently, mathematicians have found other algorithms with the same goal of minimizing an objective function. One such algorithm is *differential evolution*.

Differential evolution is one of a group of *genetic algorithms* (see for example, the recent book by Haupt (1998)). By studying how generations respond over time to their environment, mathematicians have discovered new, more robust, algorithms for minimizing an objective function. Differential evolution is one of these algorithms. It can be outlined as follows. A *population* of about 20 individuals has certain traits (values of the unknown coefficients being estimated). The well-being of an individual is measured by his/her value of the objective function. Each individual gives birth to a new individual and then dies, thus forming a new generation. Each new individual *inherits* traits from their parents. The well-being of each child is computed, and if it is better than their parent, they take the parents place. Otherwise, the parent's traits pass directly to the child. Finally, occasionally, a mutation of a trait will occur.

The main point to remember is that the big goal is to minimize the objective function. Whether we use the feedback algorithm or the differential evolution algorithm makes little difference, as long as you find the minimum!

The Algorithm

A population consists of a small group (about 20) of individuals whose characteristics are the values of the unknown coefficients. Using these coefficients, the value of the objective function is computed for each individual. This value is an inverse measure of the wellbeing of an individual. The lower the value, the higher the well being.

To begin the algorithm, a small group of individuals must be formed. This is done by assigning the nonlinear regression coefficients to one individual and then randomly assigning the other individuals to a grid of values around this first individual. This is the initial population.

The next step is the evolution of the population. The population progresses through a series of *generations*. At each change in generation, depending on a member's wellbeing, each population member may move on to the next generation or be replaced by a better member. For each member, a trial replacement is constructed as follows. The best member of the population is found. The attributes of each replacement member are computed as a weighted average of those of the member and the best member. The amount of weight of the best member is controlled by the *inheritance factor*. This is a value between 0 and 1. The closer this value is to 1, the more the replacement member resembles the best member. The closer this value is to 0, the more the replacement member resembles their parent. The value of 0.85 seems to work in many cases.

As in real populations, *mutations* occur at a given rate. When a mutation occurs, a particular trait is changed randomly. This tends to maintain diversity in the population. A mutation rate of about 30% (0.30) seems to work well.

The algorithm proceeds from generation to generation until the population seems to converge to a single individual. The number of generations is arbitrary. Usually, about 100 generations are needed for the algorithm to converge.

Assumptions

The main assumption needed is that the data are well represented by the model.

Data Structure

The data are entered as one dependent variable and one or more independent variables. Each row of data represents a single parcel.

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 (1 to 6) Tabs

These panels specify the variables used in the analysis.

Estimation Specification

Estimation Method

This option specifies the method used to estimate the model coefficients. Several methods are available.

- **Minimize Squared Errors (Nonlinear Regression)**
  This is the classical approach often used by statisticians because it gives reasonable estimates relatively quickly. This method tends to emphasize relatively expensive properties as compared to less expensive properties.

- **Minimize the Average |Percent Error|**
  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize the average of the absolute percent errors. These percent errors are the difference between the actual and predicted sales prices divided by the actual price. This method treats all properties equally, irregardless of price.

  The function minimized is

$$MAPE = \frac{\sum_{properties} \left| \frac{100(actual - predicted)}{actual} \right|}{N}$$

  where |X| represents the absolute (positive) value of X and $\sum_{properties} X$ is interpreted as the sum of the values of all properties. Note that this is the quantity minimized by the *feedback* algorithm.

  If you want to use a solution that minimizes the percent errors, this is the solution that we recommend.

- **Minimize the Maximum |Percent Error|**

  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize the maximum of the absolute percent errors. These percent errors are the difference between the actual and predicted sales prices divided by the actual price.

- **Minimize the Median |Percent Error|**

  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize the median of the absolute percent errors. These percent errors are the difference between the actual and predicted sales prices divided by the actual price.

- **Minimize the Percentile |Percent Error|**

  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize a designated percentile of the absolute percent errors. These percent errors are the difference between the actual and predicted sales prices divided by the actual price. The percentile is specified in the Min Percentile box.

- **Minimize the Average |Error|**

  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize the average of the absolute errors. These absolute errors are the difference between the actual and predicted sales prices. These percent errors are the difference between the actual and predicted sales prices divided by the actual price. The percentile is specified in the Min Percentile box.

- **Minimize the Median |Error|**

  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize the median of the absolute errors. These errors are the difference between the actual and predicted sales prices.

- **Minimize the Percentile |Error|**

  Using the genetic search algorithm called *differential evolution*, this method finds estimates that minimize a designated percentile of the absolute errors. These errors are the difference between the actual and predicted sales prices. The percentile is specified in the Min Percentile box.

Min Percentile

This option specifies the percentile minimized when the Estimation Method is set to minimize a percentile. Although a value between 1 and 99 is legal, the routine should use a value between 50 and 95.

Dependent (Sales Price) Variable

Y - Dependent Variable (Sales Price)

This option specifies the sales price variable. This is the value that will be estimated by the hybrid model.

If you want to estimate the sales price of other properties, just add their values to the end of the database, but leave the sales price blank.

Note that the prices may be in any scale you want—dollars, hundreds of dollars, or thousands of dollars. When the resulting predicted sales prices are displayed, they will be in the same scale as these values.

Variable Specification

Factor

Specify the factor (object) that this parameter belongs to. Possible choices of factor are overall, building (up to 10), land, shed, garage, or pool.

Variable

Specify one or more variables with the same factor, type, minimum, maximum, and starting values. Usually, you will specify only one variable per line, but more are allowed if you desire.

- **Binaries**

  If this is a type D (discrete) variable, you can specify a reference value in parentheses after the variable name. If you do not specify a reference value, the program sorts the values and picks the last value as the reference value. The reference value is that value for which no binary variable is generated. The number of binary (0-1) variables generated is always one less than the number of unique values.

  For example, suppose you will use a variable called ExtType that has three possible values: B for brick, U for stucco, or S for siding. Further suppose that in your area, siding is the most common exterior type. Hence, siding would be the most obvious choice for the reference value. You would enter *ExtType*(*S*) for this variable. The program would generate two binary variables: one for brick and the other for stucco.

- **Single Binary**

  It is possible to specify that only a single binary be generated for a type D (discrete) variable. This is done by adding a comma and an ‘I’ after the reference value. When you do this, only a single binary variable is generated for the value indicated.

  For example, using the exterior type example given above, the statement *ExtType(S,I)* would cause the program to generate a single indicator variable that is ‘1’ when the value of *ExtType* is S and ‘0’ otherwise.

Type

This option specifies the variable type. Possible choices are: (A)mount, (D)iscrete, and (R)ate.

- **(A)mount**

  Amount variables represent the factor size. Examples are square footage and lot size. These enter into the prediction model as linear variables such as $B_{21} X_{21} + B_{22} X_{22}$, where the *B*’s are the estimated coefficients and the *X*’s are the variables.

- **(D)iscrete**

  A discrete variable is one taking on only a view unique values, such as exterior type or neighborhood. In fact, discrete variables are not necessarily numeric. A set of indicator (binary) variables is generated for a discrete variable. This set of binary variables enters into the prediction model as $B_7^{X_7} B_8^{X_8}$, where the *B*’s are the estimated coefficients and the *X*’s are the binary variables. Note that since $B^0 = 1$ for any *B*, these variables make an adjustment when the binary variable is true (1), but have no effect when the binary variable is false (0)

- **(R)ate**

  A rate variable is a multiplier (usually centered near one) which modifies the whole factor. Examples of rate variables include sales date multiplier and depreciation. These are analogous to percentage adjustment variables. They enter into the prediction model as multipliers using the construction $X_1^{B_2} X_2^{B_2}$, where the *B*'s are the estimated coefficients and the *X*'s are the rates.

Min Start Max

Enter the word 'Defaults' or an ascending set of three numbers separated by blanks or commas. If you enter the word 'Defaults', the corresponding default values as entered under the Defaults tab will be used.

If you enter a set of numbers, the first number is the minimum, the second is the starting value, and the third is the maximum value that the coefficient associated with the variable can take on. For example, the triplet '0.1, 1.0, 2.0' sets the minimum at 0.1, the maximum at 2.0, and the starting value at 1.0. The minimum and maximum allow you to define a range of possible values (hopefully, a realistic range) in which the search for the final estimate is to take place.

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 -1,000,000,000. You can enter a -B (for negative one billion) if you want to leave the value free to vary.

Starting Value

This is the beginning value of the parameter. The algorithm searches for a value between the Minimum and the Maximum, beginning with this number. The closer this value is to the final value, the quicker the algorithm will converge.

Following are suggestions for selecting starting values.

1. Make sure that the starting values you supply are legitimate.
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.

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 1,000,000,000. You can enter a B (for positive one billion) if you want to leave the value free to vary.

Options Tab

Nonlinear Regression Options

This set of options controls the nonlinear regression algorithm. Note that when you are estimating using differential evolution, nonlinear regression is used to find appropriate starting values. Hence, these nonlinear regression options are always used.

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.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the nonlinear regression algorithm is aborted. 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.

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 Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Min Iterations

This sets the minimum number of iterations that the nonlinear regression algorithm must run before it can be terminated. Sometimes, poor starting values make the algorithm think it is finished when it is not. By setting this value to a reasonable number such as 6 or 8, the algorithm is forced to continue even when it thinks it should finish. This reduces the chance of early termination.

Differential Evolution Options

This set of options controls the differential evolution search algorithm.

Max Generations

Specify the maximum number of iterations used by the differential evolution algorithm. Usually, a value between 100 and 200 is adequate.

Individuals

This is the population size (number of trial points) used by the differential evolution algorithm at each iteration. A value between 15 and 25 is recommended. More points may dramatically increase the running time. Fewer points may not allow the algorithm to converge.

Inheritance

This value controls the amount of movement of each individual toward the current best. Usually, a value between 0.5 and 1.0 is used. We suggest 0.85. A larger value accelerates movement toward the current best, but reduces the chance of locating the global maximum. A smaller value improves the chances of finding the global, rather than a local, solution, but increases the number of iterations until convergence.

Mutation Rate

This value controls the mutation rate of the differential evolution algorithm. This is the probability that a random adjustment is made of a coefficient—which is a *mutation* in the algorithm. Values between 0 and 1 are allowed. A value of 0.3 is recommended.

Grid Range

This is the initial range about each of the initial coefficients that is sampled during the differential evolution algorithm. The algorithm is not limited to this range, but specifying a value large enough to include the solution will increase the probability of convergence.

Min Percent

This option stops the differential evolution iterations when the objective function, defined in terms of absolute percent error, is lower than this amount.

Min Amount

This option stops the differential evolution iterations when the objective function, defined in terms of absolute error, is lower than this amount.

Seed

This option specifies a random seed for the random number generator used by the differential evolution algorithm. Possible values are all integers between 1 and 32000. If you want to obtain the same results, 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’.

‘Min Start Max’ Default Options

Default ‘Min Start Max’ for Type = (Amounts, Discrete, and Rate)

Enter the default values to be used for the 'Min Start Max' settings when the word 'Defaults' is entered for that option. A separate set of defaults is required for each variable type. Suggested values are:

(A)mount:	'0.001 1 B'
(D)iscrete:	'0 1 5'
(R)ate:	'-5 0 5'.

Reports Tab

The following options control which reports are displayed.

Specify Reports

Run Summary Report - Exception Report

Each of these options specifies whether the indicated report is displayed.

Report Options

Exception Percentage

Only rows whose percent prediction error is greater than this will be displayed in the Exception Report.

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.

Ratio and Sales Decimals

These options control how many decimal places are displayed on the Ratio and Sales values that are shown on the reports.

Labels

These options control the labels that are displayed for each factor. You might want to change them for specific studies. Note that the names are completely arbitrary—they are only used to make the output more readable.

Storage Tab

The predicted values, residuals, and prediction ratios 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.

Storage Variables

Store Predicted Values in Variable

The predicted values (Yhat) are stored in this variable, if a variable is selected.

Store Residuals in Variable

The residuals (Y-Yhat) are stored in this variable, if a variable is selected.

Store Ratios in Variable

The sales ratios Y/Yhat are stored in this variable.

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 – Hybrid Appraisal Model

This section presents an example of how to estimate a prediction equation from the sales price data stored in the ASSESS database. In this example, a hybrid model of the form

$$\text{Overall(Land+Building+Garage)}$$

will be estimated.

You may follow along here by making the appropriate entries or load the completed template **Example1** from the Template tab of the Hybrid Appraisal Models window.

1 Open the ASSESS dataset.

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

2 Open the Hybrid Appraisal Models window.

- On the menus, select **Analysis**, then **Mass Appraisal**, then **Hybrid Appraisal Models**. The Hybrid Appraisal 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 model.

- On the Hybrid Appraisal Models window, select the **Variables 1 tab**.
- Set the **Estimation Method** to **Minimize the AVERAGE |PERCENT error|**.
- Set the **Y - Dependent Variable (Sales Price)** box to **Sale**.

4 Specify the overall factor.

- On the first line, set **Factor** to **Ov**, **Variable** to **Date**, and **Type** to **R**.
- On the next line, set **Factor** to **Ov**, **Variable** to **Neighborhood(5)**, and **Type** to **D**. The "5" in parentheses will cause indicator variables to be generated for all neighborhoods except neighborhood 5.

5 Specify the building factor.

- On the next line, set **Factor** to **B1**, **Variable** to **GradeLinear**, and **Type** to **R**.
- On the next line, set **Factor** to **B1**, **Variable** to **SqFt1stFlr-Baths**, and **Type** to **A**.

6 Specify the land factor.

- On the next line, set **Factor** to **Ln**, **Variable** to **LotAdjusted**, and **Type** to **R**.
- On the next line, set **Factor** to **Ln**, **Variable** to **LotSize**, and **Type** to **A**.

7 Specify the garage factor.

- On the next line, set **Factor** to **B2**, **Variable** to **GarageSqFt**, and **Type** to **A**.

8 Specify the options.

- Select the **Options tab**.
- Set the **Max Generations** to **50**.
- Set the **Min Percent** to **10**.
- Set the **Seed** to **12346**. Note that normally, you would set this to RANDOM SEED.

9 Specify the reports.

- Select the **Reports tab**.
- Check all reports.
- Set the **Exception Percentage** to **25**.

10 Specify the labels.

- Select the **Reports tab**.
- Set the **Ov** label to **Overall**.
- Set the **Ln** label to **Land**.
- Set the **B1** label to **Building**.
- Set the **B2** label to **Garage**.

11 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 Report

```
Item                            Value
Model                           Sale=Overall(Land+Building+Garage)
Estimation Method               Average |Percent Error| (AAPE)
Final Value of AAPE             11.77
R-Squared (from NonLinReg)      0.8394111
Random Number Seed              -12346

Number of Variables Used        10
Number of Parameters in Model   10
Number of Rows Used             76
Number of N.R. Iterations       10
Number of D.E. Iterations       50
```

This report displays summary information about the analysis such as the model that was fit, the number of rows and variables, the number of iterations, and the random number seed.

Model

This shows the model that was estimated. It gives you a quick overview.

Estimation Method

This is the estimation method that was used.

Final Value of AAPE

This shows the final (minimum) value of the objective function.

R-Squared (from NonLinReg)

This is the R-Squared that was achieved by the nonlinear regression routine. 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:

$$R\text{-}Squared = (ModelSS - MeanSS)/(TotalSS - MeanSS)$$

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 Number Seed

The random number seed is shown so that if you want to duplicate these results, you can be entering this random number seed.

Number of Variables Used

This is the number of variables from the database that were used in the analysis.

Number of Parameters in Model

This is the number of parameters that were estimated by the model.

Number of Rows Used

This is the number of rows from the database that were used in the analysis.

Number of N.R. Iterations

This is the number of iterations used by the nonlinear regression procedure.

Number of D.E. Iterations

This is the number of iterations used by the differential evolution algorithm.

Nonlinear Regression Iteration Section

```
Itn     Sum of
No.     Squared Errors          Message
0       87346.32
1       20104.99
2       5566.038
3       5076.756
4       5044.927
5       5044.661
6       5044.651
7       5044.647
8       5044.647
Convergence criterion met.
```

This report displays the sum of squared errors which is the objective function of the nonlinear regression routine. It allows you to observe the algorithm's progress toward the solution. If you do not see the message 'Convergence criterion met' at the bottom, it means that the algorithm did not terminate normally and you should take corrective action—which usually means that you should increase the maximum number of iterations or provide different starting values.

Differential Evolution Iteration Section

```
Itn     Objective Function
No.     (AAPE)
0       12.08115
1       12.08115
2       12.08115
3       12.08115
4       12.08115
5       12.08115
10      12.08115
15      12.08115
20      11.93821
25      11.92409
30      11.80285
35      11.78862
40      11.77882
45      11.77437
50      11.76717
```

This report displays the value of the objective function that is being minimum by the differential evolution algorithm. In this example, it is the average absolute percent error between the actual and predicted sales price. This allows you to see the progress of the algorithm towards a solution. If it looks like the objective function is still shrinking, you may want to rerun the procedure with a larger value for the Max Generations parameter.

Model Specification and Estimation Section

```
Parm   Variable          Variable          Parameter      Starting   Parameter
Name   Type              Name              Estimate       Value      Bounds
B1     Overall Rate      Date              4.662152E-02   0          -5 to 5
B2     Overall Binary    Neighborhood=4    0.9451685      1          0 to 5
B3     Overall Binary    Neighborhood=6    1.006669       1          0 to 5
B4     Building Rate     GradeLinear       1.776111       0          -5 to 5
B5     Building Amount   SqFt1stFlr        2.225528       1          0.001 to B
B6     Building Amount   SqFtOthFlr        1.634559       1          0.001 to B
B7     Building Amount   Baths             5.391271       1          0.001 to B
B8     Land Rate         LotAdjusted       3.064366       0          -5 to 5
B9     Land Amount       LotSize           0.4990667      1          0.001 to B
B10    Garage Amount     GarageSqFt        2.325238       1          0.001 to B
```

This report displays the details of the model that was fit.

Parm Name

The name of the parameter shown on this line.

Variable Type

This shows the type of the variable as well is the factor that it belongs to.

Variable Name

The name of the variable on the database. Note that for discrete variables that were expanded into a set of binary indicator variables, a separate line is given for each generated variable. The value indicated by this binary variable is shown after an equals sign. For example, a binary variable was generated that is one when the Neighborhood value is ‘4’ and zero otherwise. This variable is called ‘Neighborhood=4’.

Parameter Estimate

This is the estimated value of the parameter in the hybrid model. Note that these values should not be analyzed separately, but together as a group. If you change the model in any way (such as including other variables), these values will change—perhaps substantially!

If you are going to use these values to predict sales prices, you should use the double-precision version of these numbers. These are obtained by setting the Precision value to ‘Double’ in the Report tab.

Starting Value

These are the values used by the nonlinear regression algorithm in the first iteration. Since the differential evolution algorithm uses the results of a nonlinear regression as its starting values, these values have little influence on the results of the differential evolution algorithm.

Parameter Bounds

These are the limits that were provided for the parameter estimates. If you notice an estimate that is equal to one of its bounds, you should analyze the situation carefully to determine if bound should be relaxed to allow the parameter a wider range.

Model

Factor	Detail
Overall	Date^(B1) *(B2)^(Neighborhood=4) *(B3)^(Neighborhood=6)
Land	LotAdjusted^(B8) *((B9) *LotSize)
Building	GradeLinear^(B4) *((B5) *SqFt1stFlr +(B6) *SqFtOthFlr +(B7) *Baths)
Garage	(B10) *GarageSqFt

This report displays the variables that make up each factor. The parameters that will be estimated are B1, B2, …

Factor

The name of the factor in the hybrid equation.

Detail

The details of the construction of the factor.

Estimated Model

```
Estimated Model
Date^(4.66215216533015E-02) *(0.945168518786127)^(Neighborhood=4)
*(1.00666919375535)^(Neighborhood=6)*((LotAdjusted^(3.0643663225902) *((0.499066748801895)
*LotSize))+(GradeLinear^(1.77611087858318) *((2.22552851585402) *SqFt1stFlr +(1.63455938516463) *SqFtOthFlr
+(5.39127079611148) *Baths))+((2.32523766293606) *GarageSqFt))
```

This is the model with the parameter names replaced with the parameter estimates. This expression may be copied onto the Clipboard and pasted into the transformation section of the database to allow you to predict other observations. Note that this expression is always provided in double precision.

Assessment Ratio Section

Statistic Name	Actual Sale	Predicted Sale	Ratio	Percent Error
Number of Cases	76	76	76	76
Mean	55.9	55.1	1.01	11.77
Minimum	22.5	29.3	0.77	0.00
Lower Quartile	41.7	42.0	0.90	2.87
Median	51.9	49.4	0.99	9.43
Upper Quartile	64.3	62.6	1.05	16.77
Maximum	117.5	130.4	1.58	57.73
Range	95.0	101.1	0.80	57.72
I. Q. Range	22.6	20.6	0.15	13.90
Variance	418.8	379.8	0.03	150.84
Std. Deviation	20.5	19.5	0.17	12.28
Ave |Dev. from Median|	14.7	13.5	0.12	8.39
Coef. of Variation x 100	36.62	35.37	16.92	104.37
Coef. of Dispersion x 100	28.38	27.31	11.87	88.96
Weighted Mean			0.99	
Price Related Differential			1.02	

This report provides information that assessors have found useful in analyzing the performance of the estimated model. The four columns of the report represent the actual sales price, the predicted sales, and the ratio of the two (predicted over actual), and the percent error in the predicted. Most of the statistics are defined in the Descriptive Statistics procedure. Uncommon terms are defined next.

Ave |Dev - Median|

The average of the absolute values of the deviations of the variable from its median.

Coef. of Dispersion (COD)

This is 100 times the average absolute deviation about the median divided by the median.

IAOO standards recommend that for single-family residences, COD's of the ratios should be 15.0 or less.

Coef. of Variation (COV)

This is 100 times the standard deviation divided by the mean.

Weighted Mean

The weight ratio mean is the mean of the predicted values divided by the mean of the actual values.

Price Related Differential (PRD)

The price related differential is the mean ratio divided by the weighted mean ratio. It provides an a measure of assessment regressivity or progressivity. A PRD greater than 1.0 indicates that the more expensive properties are underappraised. A PRD less than one indicates that the more expensive properties are overappraised. Experience indicates that this value is normal when it is in the range 0.98 to 1.03.

Predicted Values of New Rows Section

```
Row            Predicted
No.                 Sale
77                  86.3
78                  44.6
79                  82.4
```

The section shows the predicted sales price for rows in which values for all variables except the sales price are given.

Using the Model to Predict for New Parcels

You can use your model to predict sales for new values of the model variables. Here is how. Add new rows to the bottom of your database containing the values of the independent variables that you want to create predictions from. Leave the sales price 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 sales price variable is available.

Exception Report Section

Row No.	Actual Sale	Predicted Sale	Actual-Predicted (Residual)	Predicted/Actual (Ratio)	Percent |Error|
3	22.5	29.6	-7.1	1.32	31.62
23	26.0	38.7	-12.7	1.49	48.89
26	78.5	106.2	-27.7	1.35	35.29
47	33.0	52.1	-19.1	1.58	57.73
53	48.0	67.8	-19.8	1.41	41.35
64	36.0	50.7	-14.7	1.41	40.81
75	30.0	44.6	-14.6	1.49	48.60

This report shows those rows that had a large percentage prediction error. These are the parcels that were poorly predicted by the model. You should analyze them to determine is there is some explanation as to why they were not fit well. You may find that the explanation is as simple as an error in data entry. It may be worth while to rerun the analysis without these rows, especially if there is a reasonable explanation as to why they did not fit the pattern shown by most of the data.

Note that the actual cutoff value for inclusion on this report is set in the Exception Percentage box under the Reports tab.

Predicted Values and Residuals Section

Row No.	Actual Sale	Predicted Sale	Actual-Predicted (Residual)	Predicted/Actual (Ratio)	Percent |Error|
1	26.0	30.5	-4.5	1.17	17.38
2	53.0	43.8	9.2	0.83	17.31
3	22.5	29.6	-7.1	1.32	31.62
4	85.0	74.3	10.7	0.87	12.57
5	48.0	44.2	3.8	0.92	7.87
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This reports shows the actual and predicted sales as well as various measures of there disagreement. Assessors commonly study the Ratio and the Percent Error for individual predictions to determine the goodness of a mass appraisal.

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

3

3D Scatter Plots, I - 170
3D Surface Plots, I - 171

A

All Possible Regressions, III - 312
Analysis of Two-Level Designs, II - 213
Analysis of Variance
    Analysis of Two-Level Designs, II - 213
    Analysis of Variance for Balanced Data, II - 211
    General Linear Models (GLM), II - 212
    Mixed Models, II - 220
    One-Way Analysis of Variance, II - 210
    Repeated Measures Analysis of Variance, II - 214
Analysis of Variance for Balanced Data, II - 211
Appraisal Ratios, IV - 485
Area Under Curve, III - 390
ARIMA (Box-Jenkins), IV - 471
Attribute Charts, II - 251
Autocorrelations, IV - 472
Automatic ARMA, IV - 474
Axis-Line Selection Window, I - 184

B

Balanced Incomplete Block Designs, II - 262
Bar Charts, I - 141
Beta Distribution Fitting, V - 551
Binary Diagnostic Tests
    Clustered Samples, V - 538
    Paired Samples, V - 536
    Single Sample, V - 535
    Two Independent Samples, V - 537
Box-Jenkins Method, IV - 470
Box Plots, I - 152

C

Canonical Correlation, III - IV - 400
Circular Data Analysis, II - 230
Clustering
    Double Dendrograms, IV - 450
    Fuzzy Clustering, IV - 448
    Hierarchical Clustering / Dendrograms, IV - 445
    K-Means Clustering, IV - 446
    Medoid Partitioning, IV - 447
    Regression Clustering, IV - 449
Color Selection Window, I - 180
Comparables - Sales Price, IV - 486
Contour Plots, I - 172
Correlation Matrix, IV - 401
Correspondence Analysis, IV - 430
Cox Regression, V - 565
Creating / Loading a Database, I - 2
Cross Tabs on Summarized Data, I - 16
Cross Tabulation, V - 501
Cross-Correlations, IV - 473
Cross-Over Analysis Using T-Tests, II - 235
Cumulative Incidence, V - 560
Curve Fitting
    Area Under Curve, III - 390
    Curve Fitting - General, III - 351
    Growth and Other Models, III - 360
    Introduction to Curve Fitting, III - 350
    Nonlinear Regression, III - 315
    Piecewise Polynomial Models, III - 365
    Ratio of Polynomials Fit
        Many Variables, III - 376
        One Variable, III - 375
    Ratio of Polynomials Search
        Many Variables, III - 371
        One Variable, III - 370
    Sum of Functions Models, III - 380
    User-Written Models, III - 385
Curve Fitting - General, III - 351

D

Data Matching – Optimal and Greedy, I - 123
Data Report, I - 117
Data Screening, I - 118
Data Simulation, I - 15
Data Simulator, I - 122
Data Stratification, I - 124
Data Transformation, I - 3
Data Window, I - 7
Database Subsets, I - 14
Databases, I - 102
    Merging Two Databases, I - 104
Decomposition Forecasting, IV - 469
Dendrograms
    Double Dendrograms, IV - 450
    Hierarchical Clustering / Dendrograms, IV - 445
Descriptive Statistics, II - 200
Descriptive Tables, II - 201
Design of Experiments

Analysis of Two-Level Designs, II - 213
Balanced Incomplete Block Designs, II - 262
Design Generator, II - 268
D-Optimal Designs, II - 267
Fractional Factorial Designs, II - 261
Latin Square Designs, II - 263
Response Surface Designs, II - 264
Screening Designs, II - 265
Taguchi Designs, II - 266
Two-Level Designs, II - 260
Design Generator, II - 268
Diagnostic Tests
Binary
Clustered Samples, V - 538
Paired Samples, V - 536
Single Sample, V - 535
Two Independent Samples, V - 537
ROC Curves, V - 545
Discriminant Analysis, IV - 440
Distribution (Weibull) Fitting, V - 550
D-Optimal Designs, II - 267
Dot Plots, I - 150
Double Dendrograms, IV - 450

E

Equality of Covariance, IV - 402
Error-Bar Charts, I - 155
Exponential Smoothing - Horizontal, IV - 465
Exponential Smoothing - Trend, IV - 466
Exponential Smoothing - Trend / Seasonal, IV - 467
Exporting Data, I - 116

F

Factor Analysis, IV - 420
Filter, I - 121
Filters, I - 10
Forecasting / Time Series
ARIMA (Box-Jenkins), IV - 471
Autocorrelations, IV - 472
Automatic ARMA, IV - 474
Cross-Correlations, IV - 473
Decomposition Forecasting, IV - 469
Exponential Smoothing
Horizontal, IV - 465
Trend, IV - 466
Trend / Seasonal, IV - 467
Spectral Analysis, IV - 468
The Box-Jenkins Method, IV - 470
Theoretical ARMA, IV - 475
Fractional Factorial Designs, II - 261
Frequency Tables, V - 500
Function Plots, I - 160
Fuzzy Clustering, IV - 448

G

Gamma Distribution Fitting, V - 552
General Linear Models (GLM), II - 212
Graphics
Introduction to Graphics, I - 140
Settings Windows
Axis-Line, I - 184
Color, I - 180
Grid / Tick, I - 185
Heat Map, I - 187
Line, I - 183
Symbol, I - 181
Text, I - 182
Tick Label, I - 186
Single-Variable Charts
Bar Charts, I - 141
Histograms, I - 143
Pie Charts, I - 142
Probability Plots, I - 144
Three-Variable Charts
3D Scatter Plots, I - 170
3D Surface Plots, I - 171
Contour Plots, I - 172
Grid Plots, I - 173
Two-Variable Charts
Box Plots, I - 152
Dot Plots, I - 150
Error-Bar Charts, I - 155
Function Plots, I - 160
Histograms - Comparative, I - 151
Percentile Plots, I - 153
Scatter Plot Matrix, I - 162
Scatter Plot Matrix for Curve Fitting, I - 163
Scatter Plots, I - 161
Violin Plots, I - 154
Greedy Data Matching, I - 123
Grid Plots, I - 173
Grid / Tick Selection Window, I - 185
Growth and Other Models, III - 360

H

Heat Map Selection Window, I - 187
Hierarchical Clustering / Dendrograms, IV - 445
Histograms, I - 143
Histograms - Comparative, I - 151
Hotelling's One-Sample T2, IV - 405
Hotelling's Two-Sample T2, IV - 410
Hybrid Appraisal Models, IV - 487

I

If-Then Transformations, I - 120
Importing Data, I - 115

Importing Data, I - 12
Installation, I - 100
Installation and Basics, I - 1
Introduction
Data
Data Matching – Optimal and Greedy, I - 123
Data Report, I - 117
Data Screening, I - 118
Data Simulator, I - 122
Data Stratification, I - 124
Exporting Data, I - 116
Filter, I - 121
If-Then Transformations, I - 120
Importing Data, I - 115
Merging Two Databases, I - 104
Transformations, I - 119
Essentials
Databases, I - 102
Installation, I - 100
Macros, I - 130
Merging Two Databases, I - 104
Navigator, I - 107
Output, I - 106
Procedures, I - 105
Spreadsheets, I - 103
Tutorial, I - 101
Introduction to Curve Fitting, III - 350
Introduction to Graphics, I - 140
Item Analysis, V - 505
Item Response Analysis, V - 506

K

Kaplan-Meier Curves (Logrank Tests), V - 555
K-Means Clustering, IV - 446

L

Latin Square Designs, II - 263
Levey-Jennings Charts, II - 252
Life-Table Analysis, V - 570
Line Selection Window, I - 183
Linear Programming, IV - 480
Linear Regression and Correlation, III - 300
Logistic Regression, III - 320
Loglinear Models, V - 530
Logrank Tests, V - 555

M

Macros, I - 130
Mantel-Haenszel Test, V - 525
Mass Appraisal
Appraisal Ratios, IV - 485
Comparables - Sales Price, IV - 486
Hybrid Appraisal Models, IV - 487
Matching – Optimal and Greedy, I - 123
Medoid Partitioning, IV - 447
Merging Two Databases, I - 104
Meta-Analysis
Correlated Proportions, IV - 457
Hazard Ratios, IV - 458
Means, IV - 455
Proportions, IV - 456
Mixed Models, II - 220
Multidimensional Scaling, IV - 435
Multiple Regression, III - 305
Multiple Regression with Serial Correlation Correction, III - 306
Multivariate Analysis
Canonical Correlation, III - IV - 400
Correlation Matrix, IV - 401
Correspondence Analysis, IV - 430
Discriminant Analysis, IV - 440
Equality of Covariance, IV - 402
Factor Analysis, IV - 420
Hotelling's One-Sample T2, IV - 405
Hotelling's Two-Sample T2, IV - 410
Multidimensional Scaling, IV - 435
Multivariate Analysis of Variance (MANOVA), IV - 415
Principal Components Analysis, IV - 425
Multivariate Analysis of Variance (MANOVA), IV - 415

N

Navigator, I - 107
Nondetects Analysis, II - 240
Nondetects Regression, III - 345
Nonlinear Regression, III - 315

O

One Proportion, V - 510
One-Way Analysis of Variance, II - 210
Operations Research
Linear Programming, IV - 480
Optimal Data Matching, I - 123
Output, I - 106
Output Window, I - 9

P

Parametric Survival (Weibull) Regression, V - 566
Pareto Charts, II - 253
Percentile Plots, I - 153
Pie Charts, I - 142
Piecewise Polynomial Models, III - 365
Poisson Regression, III - 325

Principal Components Analysis, IV - 425
Principal Components Regression, III - 340
Probability Calculator, I - 135
Probability Plots, I - 144
Probit Analysis, V - 575
Procedure Window, I - 8
Procedures, I - 105
Proportions
    Loglinear Models, V - 530
    Mantel-Haenszel Test, V - 525
    One Proportion, V - 510
    Two Correlated Proportions (McNemar), V - 520
    Two Independent Proportions, V - 515

Q

Quality Control
    Attribute Charts, II - 251
    Levey-Jennings Charts, II - 252
    Pareto Charts, II - 253
    R & R Study, II - 254
    Xbar R (Variables) Charts, II - 250
Quick Start & Self Help
    Creating / Loading a Database, I - 2
    Cross Tabs on Summarized Data, I - 16
    Data Simulation, I - 15
    Data Transformation, I - 3
    Data Window, I - 7
    Database Subsets, I - 14
    Filters, I - 10
    Importing Data, I - 12
    Installation and Basics, I - 1
    Output Window, I - 9
    Procedure Window, I - 8
    Running a Regression Analysis, I - 6
    Running a Two-Sample T-Test, I - 5
    Running Descriptive Statistics, I - 4
    Value Labels, I - 13
    Writing Transformations, I - 11

R

R & R Study, II - 254
Ratio of Polynomials Fit - Many Variables, III - 376
Ratio of Polynomials Fit - One Variable, III - 375
Ratio of Polynomials Search - Many Variables, III - 371
Ratio of Polynomials Search - One Variable, III - 370
Regression
    Cox Regression, V - 565
    Linear Regression and Correlation, III - 300
    Logistic Regression, III - 320
    Multiple Regression, III - 305
    Multiple Regression with Serial Correlation Correction, III - 306
    Nondetects Regression, III - 345
    Nonlinear Regression, III - 315
    Poisson Regression, III - 325
    Principal Components Regression, III - 340
    Response Surface Regression, III - 330
    Ridge Regression, III - 335
    Variable Selection
        Variable Selection for Multivariate Regression, III - 310
        Stepwise Regression, III - 311
        All Possible Regressions, III - 312
Regression Clustering, IV - 449
Reliability See *Survival*
Repeated Measures Analysis of Variance, II - 214
Response Surface Designs, II - 264
Response Surface Regression, III - 330
Ridge Regression, III - 335
ROC Curves, V - 545
Running a Regression Analysis, I - 6
Running a Two-Sample T-Test, I - 5
Running Descriptive Statistics, I - 4

S

Scatter Plot Matrix, I - 162
Scatter Plot Matrix for Curve Fitting, I - 163
Scatter Plots, I - 161
Screening Designs, II - 265
Settings Windows
    Axis-Line, I - 184
    Color, I - 180
    Grid / Tick, I - 185
    Heat Map, I - 187
    Line, I - 183
    Symbol, I - 181
    Text, I - 182
    Tick Label, I - 186
Spectral Analysis, IV - 468
Spreadsheets, I - 103
Stepwise Regression, III - 311
Stratification of Data, I - 124
Sum of Functions Models, III - 380
Survival / Reliability
    Beta Distribution Fitting, V - 551
    Cox Regression, V - 565
    Cumulative Incidence, V - 560
    Distribution (Weibull) Fitting, V - 550
    Gamma Distribution Fitting, V - 552
    Kaplan-Meier Curves (Logrank Tests), V - 555
    Life-Table Analysis, V - 570
    Parametric Survival (Weibull) Regression, V - 566
    Probit Analysis, V - 575
    Time Calculator, V - 580
    Tolerance Intervals, V - 585
Symbol Selection Window, I - 181

T

Tabulation
- Cross Tabulation, V - 501
- Frequency Tables, V - 500

Taguchi Designs, II - 266
Text Selection Window, I - 182
Theoretical ARMA, IV - 475
Tick Label Selection Window, I - 186
Time Calculator, V - 580
Time Series, See *Forecasting*
Tolerance Intervals, V - 585
Tools
- Data Matching – Optimal and Greedy, I - 123
- Data Simulator, I - 122
- Data Stratification, I - 124
- Macros, I - 130
- Probability Calculator, I - 135

Transformations, I - 119
T-Tests
- One-Sample or Paired, II - 205
- Two-Sample, II - 206
- Two-Sample (From Means and SD's), II - 207

Tutorial, I - 101
Two Correlated Proportions (McNemar), V - 520
Two Independent Proportions, V - 515
Two-Level Designs, II - 260

U

User-Written Models, III - 385

V

Value Labels, I - 13
Variable Selection for Multivariate
 Regression III - 310
Violin Plots, I - 154

W

Writing Transformations, I - 11

X

Xbar R (Variables) Charts, II - 250

Index

2

2BY2 dataset, 320-62

3

3D scatter plot, 140-10, 170-1
- depth, 170-8
- elevation, 170-7
- perspective, 170-6
- projection method, 170-8
- rotation, 170-7

3D surface plot, 140-10, 171-1
- depth, 171-7
- elevation, 171-6
- perspective, 171-6
- projection method, 171-7
- rotation, 171-7

A

Ability data points
- item response analysis, 506-4

Abs transformation, 119-7

Absolute residuals
- multiple regression, 305-78

Accelerated testing
- parametric survival regression, 566-1

Access exporting, 116-1

Access importing, 115-1

Accuracy
- double-precision, 102-4

Accuracy, 101-2

Active colors, 180-3

Add output to log, 106-2

Adding a datasheet, 103-2

Additive constant, 585-4
- descriptive statistics, 200-5
- tolerance intervals, 585-4

Additive seasonality
- exponential smoothing, 467-1

Adjacent values
- box plot, 152-2

Adjusted average distance
- medoid partitioning, 447-13

Adjusted R-squared
- linear regression, 300-46

A-efficiency
- D-optimal designs, 267-13

AIC
- mixed models, 220-7
- Poisson regression, 325-24

Akaike information criterion
- mixed models, 220-7
- Poisson regression, 325-24

Algorithms
- hierarchical cluster analysis, 450-2

Alias
- two level designs, 260-2
- two-level designs, 213-7

All possible regressions, 312-1

Alone lambda
- discriminant analysis, 440-13

Alpha
- Cronbach's, 401-6, 505-2
- hierarchical clustering, 445-8
- multiple regression, 305-32

Alpha Four exporting, 116-1

Alpha level of C.I.'s
- linear regression, 300-26

Alpha of assumptions
- linear regression, 300-26

Alphas
- Cox regression, 565-9, 565-38

Amplitude
- spectral analysis, 468-1

Analysis of covariance
- GLM, 212-25

Analysis of two-level designs, 213-1

Analysis of variance, 211-2
- balanced data, 211-1
- GLM, 212-1
- linear regression, 300-46
- one-way, 210-1
- repeated measures, 214-1

ANCOVA
- GLM, 212-25
- mixed models, 220-1
- multiple regression, 305-86

ANCOVA dataset, 212-25, 305-86

ANCOVA example
- mixed models, 220-85

And
- if-then transformation, 120-2

Anderson and Hauck's test
- cross-over analysis using t-tests, 235-8

Anderson-Darling test
- descriptive statistics, 200-22
- linear regression, 300-49

Andrew's sine
- multiple regression, 305-26

Angular data, 230-1

ANOVA
- balanced data, 211-1
- multiple regression, 305-49

ANOVA balanced
- assumptions, 211-2

ANOVA detail report
- multiple regression, 305-50

Answer variable
- item response analysis, 506-2

Appraisal models
- hybrid, 487-1

Appraisal ratios, 485-1

Appraisal variables, 485-2

Appraisers
- R & R, 254-11

AR order (P)
- automatic ARMA, 474-8

Arc sine transformation, 119-17

Arc tangent transformation, 119-17

ArCosh transformation, 119-17

ArcSine-square root hazard
- Weibull fitting, 550-4

Area charts, 140-1, 141-1

Area under curve, 390-1
- ROC curves, 545-26

ARIMA
- automatic ARMA, 474-1
- Box-Jenkins, 470-1, 471-1

ARMA
- theoretical, 475-1

ARMA model
- Box Jenkins, 470-2

Armitage proportion trend test
- cross tabulation, 501-5

Armitage test
- cross tabulation, 501-16

ARSENIC dataset, 240-16

Arsine transformation, 119-17

ArSinh transformation, 119-17

ArTan transformation, 119-17

ArTanh transformation, 119-17

ASCII dataset, 12-1

ASCII delimited exporting, 116-1

ASCII files

  - importing fixed format, 115-3
- ASCII fixed format exporting, 116-1
- Aspin-Welch, 206-2
- ASSESS dataset, 487-11
- Assessment models
  - hybrid appraisal, 487-1
- Assignable causes
  - presence of, 250-9
- Association
  - partial and marginal, 530-5
- Assumption tests
  - linear regression, 300-48
- Assumptions
  - analysis of variance, 210-2
  - Kruskal-Wallis test, 210-2
  - linear regression, 300-3
  - multiple regression, 305-6
  - one-sample t-test, 205-2
  - one-way ANOVA, 210-28
  - t-test, 205-22
  - two-sample t-test, 206-18, 206-27
  - two-sample t-tests, 206-1
- Asymmetric-binary variables
  - fuzzy clustering, 448-5
  - hierarchical clustering, 445-7
  - medoid partitioning, 447-2
- Asymmetry
  - probability plots, 144-3
- Attribute chart, 251-1
- AUC, 390-1
  - ROC curves, 545-1, 545-6
- AUC dataset, 390-2, 390-6
- AUC1 dataset, 390-2
- Autocorrelation, 472-1
  - multiple regression, 305-7
  - residuals, 305-53
  - type of, 300-56
- Autocorrelation function
  - Box Jenkins, 470-1
- Autocorrelation plot, 472-8
  - ARIMA, 471-12
  - automatic ARMA, 474-12
- Automatic ARMA, 474-1
- Autoregressive parameters
  - ARIMA, 471-3
  - theoretical ARMA, 475-1
- Average absolute percent error
  - multiple regression, 305-45
- Average difference plot
  - t-test, 205-20
- Average distance
  - medoid partitioning, 447-13
- Average silhouette
  - fuzzy clustering, 448-9
  - medoid partitioning, 447-13
- Average squared loadings
  - canonical correlation, 400-4
- Average transformation, 119-15
- Axis-line settings window, 184-1

B

- Backcasting
  - exponential smoothing, 465-2, 466-3, 467-3
- Backward links
  - double dendrograms, 450-2
  - hierarchical clustering, 445-3
- Balanced incomplete block designs, 262-1
- Band
  - linear regression, 300-6
- Bar charts, 140-1, 141-1
  - depth, 141-13
  - elevation, 141-13
  - gap between bars, 141-14
  - gap between sets of bars, 141-15
  - perspective, 141-12
  - projection method, 141-14
  - rotation, 141-13
- Barnard’s test of difference
  - two proportions, 515-14
- Bartlett test
  - factor analysis, 420-14
  - principal components analysis, 425-17
  - T2, 410-10
- Bartlett’s test, 402-1
- Baseline
  - area under curve, 390-3
- Baseline cumulative survival
  - Cox regression, 565-38
- Baseline survival
  - Cox regression, 565-8
- Basic palette, 180-2
- Basics, 1-1
- BBALL dataset, 445-5, 445-12, 446-2, 446-6, 447-6, 447-12
- BEAN dataset, 220-79, 220-82
- Best model
  - all possible regressions, 312-4
- Beta
  - hierarchical clustering, 445-8
- BETA dataset, 551-2, 551-11
- Beta distribution
  - probablility calculator, 135-1
  - simulation, 122-3
- Beta distribution fitting, 551-1
- Beta trace
  - PC regression, 340-14
- BetaProb transformation, 119-8
- BetaValue transformation, 119-8
- Between subject
  - repeated measures, 214-2
- Bias
  - R & R, 254-22
- BIB designs, 262-1
- Bimodal data
  - simulation, 122-23
- Binary diagnostic tests
  - clustered samples, 538-1
  - paired samples, 536-1
  - two independent samples, 537-1
- Binary response variables, 320-1
- Binary test
  - 1-sample binary diagnostic test, 535-1
- BINCLUST dataset, 538-3, 538-7
- Binomial distribution
  - probablility calculator, 135-2
  - simulation, 122-5
- BinomProb transformation, 119-8
- BinomValue transformation, 119-8
- Binormal
  - ROC curves, 545-2
- Bioequivalence
  - cross-over analysis using t-tests, 235-5
- Bisquare weights
  - linear regression, 300-14
- Bivariate normal distribution
  - probablility calculator, 135-2
- Biweight
  - Weibull fitting, 550-17
- Biweight estimator of scale, 200-22
- Biweight kernel
  - Kaplan-Meier, 555-9
  - Weibull fitting, 550-35
- Blackwelder test
  - correlated proportions, 520-5
- Bleasdale-Nelder model
  - curve fitting, 351-5
  - growth curves, 360-3
- Block size
  - balanced incomplete block designs, 262-3
  - fractional factorial designs, 261-2
- Block variable
  - fractional factorial designs, 261-1
  - response surface designs, 264-2
- Blocking
  - two level designs, 260-2
- BMDP exporting, 116-1
- BMT dataset, 555-43
- Bonferroni
  - one-way ANOVA, 210-4
- Bonferroni adjustment
  - mixed models, 220-14
- Bonferroni C.I.’s
  - T2, 405-9, 410-9
- Bootstrap
  - linear regression, 300-42
- Bootstrap C.I. method
  - linear regression, 300-30
  - two proportions, 515-28
- Bootstrap C.I.’s
  - multiple regression, 305-31
- Bootstrap confidence coefficients
  - linear regression, 300-30
- Bootstrap histograms

linear regression, 300-31, 300-44, 305-42
multiple regression, 305-75
Bootstrap percentile type
linear regression, 300-30
two proportions, 515-28
Bootstrap report
multiple regression, 305-74
Bootstrap retries
linear regression, 300-30
two proportions, 515-28
Bootstrap sample size
linear regression, 300-29
two proportions, 515-28
Bootstrap sampling method
linear regression, 300-30
Bootstrapping
curve fitting, 351-14
linear regression, 300-22
multiple regression, 305-21
t-test, 205-3
two-sample t-test, 206-3
Bootstrapping example
multiple regression, 305-72, 305-76
Box plot
adjacent values, 152-2
fences, 152-6
interquartile range, 152-1
whiskers, 152-5
Box plot style file, 152-13
Box plots, 140-5, 152-1
multiple comparisons, 152-2
Box's M, 214-1
Box's M test, 402-1, 402-7
Hotelling's T2, 410-2
repeated measures, 214-22
T2, 410-10
BOX320 dataset, 213-6
BOX402 dataset, 213-12
Box-Behnken designs, 264-1
Box-Jenkins
ARIMA, 471-1
automatic ARMA, 474-1
Box-Jenkins analysis, 470-1
Box-Pierce-Ljung statistic
automatic ARMA, 474-12
Box's M test
MANOVA, 415-5
BRAIN WEIGHT dataset, 2-2
Breslow ties
Cox regression, 565-6

C

C.I.method
multiple regression, 305-41
Calibration
linear regression, 300-6, 300-41
Caliper matching, 123-4
Caliper radius, 123-5
Candidate points
D-optimal designs, 267-14
Canonical correlation, 400-1
Canonical variate
MANOVA, 415-13
Capability analysis
Xbar R, 250-11
Capacities
Xbar R, 250-30
Carryover effect
cross-over analysis using t-tests, 235-3
Cascade, 106-5
Categorical IV's
Cox regression, 565-20
logistic regression, 320-20
multiple regression, 305-29
Poisson regression, 325-9
Categorical variables
multiple regression, 305-3, 305-87
Cauchy distribution
simulation, 122-5
Cbar
logistic regression, 320-15
C-chart, 251-2
Cell edit box, 103-10
Cell reference, 103-10
Censor variable
parametric survival regression, 566-4
Censored
Cox regression, 565-17
Kaplan-Meier, 555-15
Weibull fitting, 550-11
Censored regression, 566-1
Centering
Cox regression, 565-19
Central moments
descriptive statistics, 200-11
Central-composite designs, 264-1
Centroid
double dendrograms, 450-2
hierarchical clustering, 445-3
Charts
pareto, 253-1
variables, 250-1
Checklist
one sample tests, 205-21
one-way ANOVA, 210-26
two-sample tests, 206-25
Chen's method
two proportions, 515-20
Chi
loglinear models, 530-20
Chi-square
cross tabulation, 501-10
frequency tables, 500-11
Poisson regression, 325-26
Chi-square distribution
probablility calculator, 135-2
Chi-square test
cross tabulation, 501-1
two proportions, 515-6
Chi-square test example, 16-1
CHOWLIU73 dataset, 235-9, 235-15
Circular correlation, 230-12
Circular data analysis, 230-1
Circular histogram, 230-17
Circular histograms, 230-1
Circular statistics, 230-1
Circular uniform distribution, 230-3
CIRCULAR1 dataset, 230-22
Circularity
repeated measures, 214-3, 214-23
Clear, 103-5
Cluster analysis
double dendrograms, 450-1
K-means, 446-1
Cluster centers
K-means clustering, 446-1
Cluster cutoff
hierarchical clustering, 445-8
Cluster means
K-means clustering, 446-8
Cluster medoids section
fuzzy clustering, 448-9
medoid partitioning, 447-14
Cluster randomization
clustered binary diagnostic, 538-1
Cluster variables
K-means clustering, 446-3
Clustering
centroid, 445-7
complete linkage, 445-7
flexible strategy, 445-7
fuzzy, 448-1
group average, 445-7
hierarchical, 445-1
median, 445-7
medoid, 447-1
regression, 449-1
simple average, 445-7
single linkage, 445-7
Ward's minimum variance, 445-7
Cochran's Q test
meta analysis of hazard ratios, 458-4
meta-analysis of correlated proportions, 457-4
meta-analysis of means, 455-3
meta-analysis of proportions, 456-4
Cochran's test
two proportions, 515-7
Cochrane-Orcutt procedure, 306-1
COD
appraisal ratios, 485-8
descriptive statistics, 200-20
hybrid appraisal models, 487-17

Code cross-reference, 310-7
Coefficient alpha
  item analysis, 505-2
Coefficient of dispersion
  appraisal ratios, 485-8
  descriptive statistics, 200-18
  hybrid appraisal models, 487-17
Coefficient of variation
  descriptive statistics, 200-18
  linear regression, 300-38
  multiple regression, 305-45
Coefficients
  regression, 305-47
  stepwise regression, 311-8
Collate transformation, 119-12
COLLETT157 dataset, 565-55
COLLETT266 dataset, 320-73
COLLETT5 dataset, 555-42
Collinearity
  MANOVA, 415-5
Color
  mixer, 180-2
  model, 180-2
  wheel, 180-3
Color selection window, 180-1
Column widths, 103-15
Communality
  factor analysis, 420-3, 420-12, 420-16
  principal components analysis, 425-16
Communality iterations
  factor analysis, 420-8
Comparables
  sales price, 486-1
COMPARABLES dataset, 486-10
Competing risks
  cumulative incidence, 560-1
Complete linkage
  double dendrograms, 450-2
  hierarchical clustering, 445-3
Compound symmetry
  repeated measures, 214-3
CONCENTRATION dataset, 240-21
Concordance
  Kendall's coefficient, 211-15
Condition number
  multiple regression, 305-58
  PC regression, 340-13
  ridge regression, 335-17
Conditional tests
  two proportions, 515-5
Confidence band
  linear regression, 300-6, 300-33, 300-60
Confidence coefficient
  multiple regression, 305-32
  T2, 410-5
Confidence interval
  descriptive statistics, 200-13
  multiple regression, 305-14
  Poisson regression, 325-26
Confidence intervals
  Cox regression, 565-11
  curve fitting, 350-4
  linear regression, 300-6
  T2, 405-9, 410-9
  two proportions, 515-18
Confidence intervals of odds ratio
  two proportions, 515-23
Confidence intervals of ratio
  two proportions, 515-21
Confidence limits, 200-2
  linear regression, 300-33
  Nelson-Aalen hazard, 550-4
Confounding
  two level designs, 260-2
Confounding size, 213-3
Constant distribution
  simulation, 122-6
Constraint section
  linear programming, 480-5
Constraints
  linear programming, 480-1
Contains transformation, 119-17
Contaminated normal simulation, 122-21
Continuity correction
  two proportions, 515-7
Contour plots, 140-11, 172-1
  response surface regression, 330-19
Contrast type
  multiple regression, 305-29
  Poisson regression, 325-9
Contrast variables
  multiple regression, 305-4
Control charts
  attribute, 251-1
  formulas, 250-5
  Xbar R, 250-1
Control limits
  Xbar R, 250-2
Cook's D
  linear regression, 300-20, 300-62, 300-63, 300-65, 300-66
  multiple regression, 305-20, 305-64
Cook's distance
  logistic regression, 320-15
Cophenetic correlation
  hierarchical clustering, 445-14
Cophenetic correlation coefficient, 445-4
Copy, 103-4
Copy output, 106-3
Copying data, 7-2
COR
  correspondence analysis, 430-14
Correlation, 300-1
  canonical, 400-1
  confidence limits, 300-12
  cross, 473-1
  linear regression, 300-2, 300-11, 300-45
  Pearson, 300-45
  Spearman, 300-45
  Spearman rank, 401-1
  Spearman's rank, 300-12
Correlation coefficient
  linear regression, 300-9
Correlation coefficient distribution
  probablility calculator, 135-3
Correlation matrices
  factor analysis, 420-5
  principal components analysis, 425-8
Correlation matrix, 401-1
Correlation matrix report
  multiple regression, 305-46
Correlations
  medoid partitioning, 447-10
  partial, 401-3
  principal components analysis, 425-17
Correlogram
  autocorrelation, 472-1
CORRES1 dataset, 430-6, 430-10, 430-16
Correspondence analysis, 430-1
  eigenvalues, 430-12
CorrProb transformation, 119-8
CorrValue transformation, 119-8
Cos transformation, 119-17
Cosh transformation, 119-17
Cosine transformation, 119-17
Cost benefit analysis
  ROC curves, 545-22
Count tables, 500-1
Count transformation, 119-15
Covariance
  analysis of, 212-25
  multiple regression, 305-86
Covariance matrices, 402-1
Covariance matrix
  repeated measures, 214-3
Covariance pattern models
  mixed models, 220-5
Covariates
  GLM, 212-3
  mixed models, 220-9
  response surface regression, 330-5
CovRatio
  linear regression, 300-21, 300-63
  multiple regression, 305-20, 305-64
Cox model
  Cox regression, 565-1
Cox proportional hazards regression model, 565-1
Cox regression, 565-1
Cox test

- circular data, 230-9
- Cox-Mantel logrank test
  - Kaplan-Meier, 555-41
- COXREG dataset, 565-51
- COXSNELL dataset, 123-23
- Cox-Snell residual
  - parametric survival regression, 566-19
- Cox-Snell residuals
  - Cox regression, 565-13, 565-39
  - nondetects regression, 345-13
- Cp
  - all possible regressions, 312-8
  - multiple regression, 305-55
  - Xbar R, 250-12
- Cp variable plot
  - all possible regressions, 312-10
- Cpk
  - Xbar R, 250-12, 250-31
- Cramer's V
  - cross tabulation, 501-14
- Creating a database, 2-1
- Creating a new database
  - tutorial, 101-2
- Creating data
  - simulation, 122-1
- Cronbach's alpha
  - item analysis, 505-2, 505-6
- Cronbachs alpha
  - correlation matrix, 401-6
- CROSS dataset, 220-101
- Cross tabulation, 501-1
  - summarized data, 16-1
- Cross-correlations, 473-1
- Crossed factors
  - design generator, 268-1
- Crossover analysis, 220-1
- Cross-over analysis using t-tests, 235-1
- Crossover data example
  - mixed models, 220-101
- Crosstabs, 501-1
- CsProb transformation, 119-9
- CsValue transformation, 119-9
- CTR
  - correspondence analysis, 430-14
- Cubic fit
  - curve fitting, 351-2
- Cubic terms
  - response surface regression, 330-7
- Cum transformation, 119-7
- Cumulative hazard
  - Cox regression, 565-2
- Cumulative hazard function
  - Kaplan-Meier, 555-2
  - Weibull fitting, 550-2
- Cumulative incidence analysis, 560-1
- Cumulative survival
  - Cox regression, 565-2
- Curve equivalence
  - curve fitting, 351-16
- Curve fitting, 351-1
  - introduction, 350-1
- Curve inequality test
  - curve fitting, 351-32
- Custom model
  - Cox regression, 565-26
  - multiple regression, 305-34
- CUSUM chart, 250-4, 250-8
- CUSUM Charts, 250-37
- Cut, 103-4
- Cut output, 106-3
- Cycle-input variable
  - decomposition forecasting, 469-5

D

- D'Agostino kurtosis
  - descriptive statistics, 200-24
- D'Agostino kurtosis test
  - linear regression, 300-49
- D'Agostino omnibus
  - descriptive statistics, 200-25
- D'Agostino omnibus test
  - linear regression, 300-49
- D'Agostino skewness
  - descriptive statistics, 200-23
- D'Agostino skewness test
  - linear regression, 300-49
- DAT exporting, 116-1
- Data
  - entering, 2-1
  - estimating missing, 118-1
  - importing, 12-1
  - numeric, 102-1
  - printing, 2-7, 103-3, 117-1
  - saving, 2-6
  - simulation, 15-1
  - simulation of, 122-1
  - text, 102-1
- Data features, 200-1
- Data imputation, 118-1
- Data matching
  - caliper, 123-4
  - caliper radius, 123-5
  - distance calculation method, 123-3
  - forced match variable, 123-4
  - full (variable), 123-3
  - greedy, 123-1, 123-2
  - optimal, 123-1, 123-2
  - propensity score, 123-2
  - standardized difference, 123-15
- Data orientation
  - bar charts, 141-2
- Data report, 103-6, 117-1
- Data screening
  - T2 alpha, 118-3
- Data screening, 118-1
- Data screening, 200-3
- Data simulator, 122-1
- Data stratification, 124-1
- Data transformation, 3-1
- Data type, 102-10
- Data window, 1-4, 7-1
- Database, 102-1
  - clearing, 2-9
  - creating, 2-1, 101-2
  - Excel compatible, 102-1
  - exporting, 115-1, 116-1
  - introduction, 101-1
  - limits, 102-1
  - loading, 2-1, 2-10, 7-1
  - opening, 101-3
  - printing, 2-7
  - S0, 102-1
  - s0 and s1 files, 2-6
  - S0-type, 2-9
  - S0Z (zipped), 102-1
  - S0Z-type, 2-9
  - saving, 101-2
  - size, 102-1
  - sorting, 103-6
  - subsets, 14-1
- Database/spreadsheet comparison, 102-4
- Databases
  - merging two, 104-1
- Dataset
  - 2BY2, 320-62
  - ANCOVA, 212-25, 305-86
  - ARSENIC, 240-16
  - ASCII, 12-1
  - ASSESS, 487-11
  - AUC, 390-2, 390-6
  - AUC1, 390-2
  - BBALL, 445-5, 445-12, 446-2, 446-6, 447-6, 447-12
  - BEAN, 220-79, 220-82
  - BETA, 551-2, 551-11
  - BINCLUST, 538-3, 538-7
  - BMT, 555-43
  - BOX320, 213-6
  - BOX402, 213-12
  - BRAIN WEIGHT, 2-2
  - CHOWLIU73, 235-9, 235-15
  - CIRCULAR1, 230-22
  - COLLETT157, 565-55
  - COLLETT266, 320-73
  - COLLETT5, 555-42
  - COMPARABLES, 486-10
  - CONCENTRATION, 240-21
  - CORRES1, 430-6, 430-10, 430-16
  - COXREG, 565-51
  - COXSNELL, 123-23
  - CROSS, 220-101
  - DCP, 345-2, 345-9
  - DIOXIN, 240-2, 240-11

DOPT_MIXED, 267-22
DOPT3, 267-20
DRUGSTUDY, 501-19
DS476, 315-2, 315-9, 385-2, 385-9
EXAMS, 450-12
EXERCISE, 214-6, 214-16
FANFAILURE, 550-49
FISH, 220-90
FISHER, 143-14, 144-15, 150-8, 151-13, 152-12, 153-8, 154-8, 170-2, 170-9, 173-7, 402-2, 402-5, 440-4, 440-10, 440-20, 440-22
FNREG1, 360-15, 380-7
FNREG2, 365-11
FNREG3, 163-4, 370-6, 375-8
FNREG4, 371-6, 376-8
FNREG5, 351-30
FRUIT, 141-1, 141-17
FUZZY, 448-3, 448-8
HAIR, 220-103
HEART, 212-23
HOUSING, 306-4, 306-10
INTEL, 465-7, 466-9, 471-7, 473-5
IQ, 305-27, 305-43, 305-72, 305-76, 305-79
ITEM, 505-2, 505-5, 506-2, 506-6
KLEIN6, 555-45
KOCH36, 325-7, 325-21
LACHIN91, 320-71
LATINSQR, 212-22
LEAD, 240-19
LEE91, 570-4, 570-15
LEUKEMIA, 320-18, 320-34, 320-57
LINREG1, 300-24, 300-37
LOGLIN1, 530-7, 530-11
LP, 480-2, 480-4
LUNGCANCER, 565-15, 565-31, 565-48
MAMMALS, 3-1, 4-1, 10-1
MAMMALS1, 5-1, 6-1
MANOVA1, 410-3, 410-6, 415-5, 415-10
MARUBINI, 560-3, 560-9
MDS2, 435-6, 435-10
MDS2, 435-15
METACPROP, 457-6, 457-14
METAHR, 458-6, 458-12
MLCO2, 470-11
MOTORS, 566-3, 566-11
NC CRIMINAL, 320-64, 320-68
NONDETECTS, 240-4
ODOR, 330-3, 330-11
PAIN, 220-51
PCA2, 420-5, 420-11, 425-9, 425-15
PCA2, 118-4
PET, 538-11
PIE, 142-6
PLANT, 212-27
POISREG, 325-37
POLITIC, 13-1, 14-1
PREPOST, 305-87
PROPENSITY, 123-5, 123-12, 124-4
QATEST, 250-14, 250-27, 250-33, 250-35, 250-37, 251-3, 251-11, 253-2, 253-7, 253-9
RCBD, 220-94
REACTION, 214-29
REACTION, 214-6
READOUT105, 550-47
REGCLUS, 449-2, 449-5
RESALE, 117-4, 151-14, 155-1, 155-7, 201-1, 201-11, 201-12, 201-14, 201-15, 201-17, 201-19, 201-21, 305-81, 500-1, 500-9, 500-10, 500-12, 500-14, 501-1, 501-8, 501-11, 501-17
RIDGEREG, 335-7, 335-15, 340-3, 340-11
RMSF, 545-3
RNDBLOCK, 211-4, 211-11, 212-3, 212-12
ROC, 545-19
RRSTUDY, 254-1, 254-10
RRSTUDY1, 254-24
SALES, 467-9, 469-9
SALESRATIO, 485-1, 485-6, 486-4
SAMPLE, 101-3, 161-20, 162-5, 171-9, 172-7, 200-4, 200-10, 205-12, 206-12, 210-16, 310-3, 310-6, 311-3, 311-6, 312-2, 312-6, 400-8, 401-2, 401-5, 585-8
SERIESA, 470-8, 474-7
SMOKING, 525-2, 525-5
SUNSPOT, 468-9, 472-7
SURVIVAL, 555-14, 555-37, 575-1, 575-5
SUTTON 22, 456-6, 456-14
SUTTON30, 455-6, 455-13
T2, 405-3, 405-5, 405-10
TIMECALC, 580-3
TUTOR, 220-98
TWOSAMPLE, 220-69, 220-72
TWOSAMPLE2, 220-70, 220-73
TWOSAMPLECOV, 220-76
WEIBULL, 550-12, 550-27, 550-44, 552-3, 552-12, 555-27
WEIBULL2, 144-17
WEIGHTLOSS, 220-85
WESTGARD, 252-9
ZHOU 175, 545-33
ZINC, 345-15
Datasheet, 101-1
Datasheets, 102-1
Date formats, 102-8
Date function transformations, 119-6
Day format, 102-8
Day transformation, 119-6
DB, 115-1
Dbase importing, 115-1
DBF exporting, 116-1
DBF importing, 115-1
DCP dataset, 345-2, 345-9
Death density
life-table analysis, 570-3
Decision variables
linear programming, 480-1
Decomposition forecasting, 469-1
Default template, 105-1
Defects/defectives variable, 251-4
D-efficiency
D-optimal designs, 267-12
Degrees of freedom
factor analysis, 420-14
two-sample t-test, 206-13
Delta
cluster goodness-of-fit, 445-4
loglinear models, 530-8
Mantel-Haenszel test, 525-4
Dendrogram
hierarchical clustering, 445-15
Dendrograms, 445-1
double, 450-1, 450-3
Density trace
histograms, 143-1
histograms – comparative, 151-2
violin plot, 154-1
Dependent variable
linear regression, 300-25
multiple regression, 305-1
Poisson regression, 325-8
Depth
3D scatter plot, 170-8
3D surface plot, 171-7
bar charts, 141-13
Derivatives
Weibull fitting, 550-16
Descriptive statistics, 4-1, 200-1
additive constant, 200-5
Anderson-Darling test, 200-22
central moments, 200-11
COD, 200-20
coefficient of dispersion, 200-18
coefficient of variation, 200-18
confidence interval, 200-13
D'Agostino kurtosis, 200-24
D'Agostino omnibus, 200-25
D'Agostino skewness, 200-23
dispersion, 200-16
EDF, 200-7
Fisher's g1, 200-18
Fisher's g2, 200-18
geometric mean, 200-14
harmonic mean, 200-14

Histogram, 200-25
interquartile range, 200-17
IQR, 200-17
Kolmogorov-Smirnov, 200-23
kurtosis, 200-18
Lilliefors' critical values, 200-23
MAD, 200-20
Martinez-Iglewicz, 200-22
mean, 200-13
mean absolute deviation, 200-20
mean deviation, 200-20
mean-deviation, 200-20
median, 200-14
mode, 200-15
moment, 200-11
Normal probability plot, 200-26
normality, 200-21
normality tests, 200-21
percentile type, 200-6
Probability plot, 200-26
quartiles, 200-21
range, 200-17
Shapiro-Wilk test, 200-22
skewness, 200-17
Skewness test, 200-24
standard deviation, 200-16
standard error, 200-13
Stem-leaf plot, 200-27
trim-mean, 200-19
trimmed, 200-19
trim-std dev, 200-19
unbiased Std Dev, 200-17
variance, 200-15
Descriptive statistics report
multiple regression, 305-45
Descriptive tables, 201-1
Design generator, 268-1
Designs
analysis of, 213-1
Box-Behnken, 264-1
central-composite, 264-1
design generator, 268-1
factorial, 260-3
fractional factorial, 261-1
Plackett-Burman, 265-1
response surface, 264-1
screening, 265-1
Taguchi, 266-1
two-level factorial, 260-1, 268-1
Determinant
D-optimal designs, 267-13
Determinant analysis
D-optimal designs, 267-11
Deviance
Cox regression, 565-10
logistic regression, 320-8
Poisson regression, 325-4, 325-5
Deviance residuals
Cox regression, 565-14, 565-40
logistic regression, 320-13
Poisson regression, 325-31
Deviance test
Poisson regression, 325-3
DFBETA
logistic regression, 320-14
DFBETAS
linear regression, 300-21, 300-63
multiple regression, 305-20, 305-65
DFCHI2
logistic regression, 320-15
DFDEV
logistic regression, 320-15
Dffits
linear regression, 300-63
DFFITS
linear regression, 300-20
multiple regression, 305-19, 305-64
Diagnostic test
1-sample binary diagnostic test, 535-1
2-sample binary diagnostic, 537-1
paired binary diagnostic, 536-1
DIF exporting, 116-1
Differencing
ARIMA, 471-2
autocorrelation, 472-2
Box Jenkins, 470-7
spectral analysis, 468-4
Differential evolution
hybrid appraisal models, 487-2
Weibull fitting, 550-11
Digamma
beta distribution fitting, 551-12
Dimensions
multidimensional scaling, 435-4
DIOXIN dataset, 240-2, 240-11
Directional test
meta analysis of hazard ratios, 458-3
meta-analysis of correlated proportions, 457-4
meta-analysis of proportions, 456-4
Disabling the filter, 121-4
Discriminant analysis, 440-1
logistic regression, 320-1
Discrimination parameter
item response analysis, 506-8
Dispersion
descriptive statistics, 200-16
Dissimilarities
medoid partitioning, 447-1
multidimensional scaling, 435-4
Distance
multidimensional scaling, 435-2
Distance calculation
medoid partitioning, 447-2
Distance calculation method
data matching, 123-3
Distance method
fuzzy clustering, 448-5
hierarchical clustering, 445-8
Distances
medoid partitioning, 447-10
Distinct categories
R & R, 254-3, 254-19
Distribution
circular uniform, 230-3
Von Mises, 230-5
Distribution fitting
Weibull fitting, 550-1
Distribution statistics, 200-1
Distributions
combining, 122-13
exponential, 550-1
extreme value, 550-1
logistic, 550-1
log-logistic, 550-1
lognormal, 550-1
mixing, 122-13
simulation, 122-1
Weibull, 550-1
Dmn-criterion value, 206-23
DOPT_MIXED dataset, 267-22
DOPT3 dataset, 267-20
D-optimal designs, 267-1
Dose
probit analysis, 575-1
Dose-response plot
probit analysis, 575-9
Dot plots, 140-4, 150-1
jittering, 150-1
Double dendrograms, 450-1
Double exponential smoothing, 466-1
Double-precision accuracy, 101-2, 102-4
DRUGSTUDY dataset, 501-19
DS476 dataset, 315-2, 315-9, 385-2, 385-9
Dummy variables
multiple regression, 305-3
Duncan’s test
one-way ANOVA, 210-5
Dunn’s partition coefficient
fuzzy clustering, 448-2
Dunn’s test
one-way ANOVA, 210-7
Dunnett’s test
one-way ANOVA, 210-6
Duplicates
D-optimal designs, 267-5
Durbin-Watson
linear regression, 300-17
multiple regression, 305-17
Durbin-Watson test
multiple regression, 305-53
multiple regression with serial correlation, 306-3

E

e - using
  Cox regression, 565-4
E notation, 102-4
EDF
  descriptive statistics, 200-7
EDF plot, 240-15
Edit
  clear, 103-5
  copy, 103-4
  cut, 103-4
  delete, 103-5
  fill, 103-6
  find, 103-6
  insert, 103-5
  paste, 103-4
  undo, 103-4
Efron ties
  Cox regression, 565-7
Eigenvalue
  MANOVA, 415-14
  PC regression, 340-13
Eigenvalues, 425-17
  correspondence analysis, 430-12
  factor analysis, 420-14
  multidimensional scaling, 435-11
  multiple regression, 305-58, 305-59
  principal components analysis, 425-12
  ridge regression, 335-17
Eigenvector
  multiple regression, 305-58, 305-60
Eigenvectors
  factor analysis, 420-15
Elapsed time
  time calculator, 580-1
Elevation
  3D scatter plot, 170-7
  3D surface plot, 171-6
  bar charts, 141-13
Ellipse (probability)
  linear regression, 300-8
Else
  if-then transformation, 120-4
EM algorithm
  principal components analysis, 425-5
Empirical
  ROC curves, 545-2
Empty cells, 102-5
Entry date
  time calculator, 580-2
Entry time
  Cox regression, 565-17
  Kaplan-Meier, 555-15
Epanechnikov
  Weibull fitting, 550-17
Epanechnikov kernel
  Kaplan-Meier, 555-8
  Weibull fitting, 550-34
Epsilon
  Geisser-Greenhouse, 214-4
  repeated measures, 214-20
Equal slopes
  multiple regression, 305-86
Equality of covariance matrices, 402-1
Equivalence
  2-sample binary diagnostic, 537-9
  clustered binary diagnostic, 538-8
  cross-over analysis using t-tests, 235-1
  paired binary diagnostic, 536-7
  ROC curves, 545-30
Equivalence test
  correlated proportions, 520-8
  two proportions, 515-17
  two-sample, 207-1
Equivalence tests
  two proportions, 515-38
Error-bar charts, 140-6, 155-1
Euclidean distance
  medoid partitioning, 447-2
Event date
  time calculator, 580-2
EWMA chart, 250-4, 250-35
EWMA chart limits, 250-8
EWMA parameter, 250-19
Exact test
  two proportions, 515-12
Exact tests
  two proportions, 515-4, 515-36
EXAMS dataset, 450-12
Excel exporting, 116-1
EXERCISE dataset, 214-6, 214-16
Exiting NCSS, 101-4
Exp transformation, 119-7
Experiment (Run)
  two level designs, 260-2
Experimental design, 260-1
  two level designs, 260-2
Experimental error
  two level designs, 260-2
Experimentwise error rate, 210-3
Exponential
  curve fitting, 351-10
  using, 565-4
Exponential distribution
  simulation, 122-6
  Weibull fitting, 550-8
Exponential model
  curve fitting, 351-6
  growth curves, 360-4
Exponential regression, 566-1
Exponential smoothing
  double, 466-1
  horizontal, 465-1
  simple, 465-1
  trend, 466-1
  trend and seasonal, 467-1
ExpoProb transformation, 119-9
Export, 103-3
Export limitations, 116-1
Exporting data, 116-1
Exposure
  Poisson regression, 325-1
Exposure variable
  Poisson regression, 325-12
ExpoValue transformation, 119-9
Extract transformation, 119-18
Extreme value distribution
  Weibull fitting, 550-8

F

F distribution
  probablility calculator, 135-3
  simulation, 122-7
Factor analysis, 420-1
Factor loadings
  factor analysis, 420-16
  principal components analysis, 425-2
Factor rotation
  factor analysis, 420-7
Factor scaling
  D-optimal designs, 267-2
Factorial designs
  two level designs, 260-3
  two-level designs, 260-1
Factors
  how many, 420-3, 425-6
Failed
  parametric survival regression, 566-2
  Weibull fitting, 550-11
Failure
  Cox regression, 565-16
  Kaplan-Meier, 555-15
Failure distribution
  Weibull fitting, 550-37
Familywise error rate, 210-3
FANFAILURE dataset, 550-49
Farazdaghi and Harris model
  curve fitting, 351-5
  growth curves, 360-3
Farrington-Manning test
  two proportions, 515-10
Fast Fourier transform
  spectral analysis, 468-3
Fast initial restart, 250-9
Feedback model, 487-1
Fences
  box plot, 152-6
File function transformation, 119-15
Files

Access, 115-1
ASCII, 115-3
BMDP, 115-1
creating text, 115-1
Dbase, 115-1
Excel, 115-1
NCSS 5.0, 115-1
Paradox, 115-1
SAS, 115-1
SPSS, 115-1
text, 115-1
Fill, 103-6
Fill functions transformations, 119-6
Filter, 121-1
disabling, 10-4
specifying, 103-7
Filter statements, 103-7
Filters, 10-1
Final Tableau section
linear programming, 480-6
Find, 103-6
Find a procedure, 107-1
Find in output, 106-4
Find next in output, 106-4
FIR, 250-9
FISH dataset, 220-90
FISHER dataset, 143-14, 144-15, 150-8, 151-13, 152-12, 153-8, 154-8, 170-2, 170-9, 173-7, 402-2, 402-5, 440-4, 440-10, 440-20, 440-22
Fisher information matrix
beta distribution fitting, 551-14
gamma distribution fitting, 552-15
Weibull fitting, 550-32
Fisher’s exact test, 501-1, 501-13
cross tabulation, 501-17
Fisher’s Z transformation
linear regression, 300-11
Fisher's exact test
cross tabulation, 501-11
Fisher's g1
descriptive statistics, 200-18
Fisher's g2
descriptive statistics, 200-18
Fisher's LSD
one-way ANOVA, 210-6
Fixed effects
mixed models, 220-9
Fixed effects model
meta-analysis of correlated proportions, 457-5
meta-analysis of hazard ratios, 458-4
meta-analysis of means, 455-4
meta-analysis of proportions, 456-5
Fixed effects models
mixed models, 220-4
Fixed factor
ANOVA balanced, 211-5
GLM, 212-4
repeated measures, 214-8
Fixed sigma
Xbar R, 250-19
Fixed Xbar
Xbar R, 250-18
Fleiss Confidence intervals
two proportions, 515-24
Fleming-Harrington tests
Kaplan-Meier, 555-12
Flexible strategy
double dendrograms, 450-3
hierarchical clustering, 445-4
Flipping constant, 240-2
FNREG1 dataset, 360-15, 380-7
FNREG2 dataset, 365-11
FNREG3 dataset, 163-4, 370-6, 375-8
FNREG4 dataset, 371-6, 376-8
FNREG5 dataset, 351-30
Follow-up
life-table analysis, 570-2
Forced match variable, 123-4
Forced points
D-optimal designs, 267-5
Forced X’s
variable selection, 310-4
Forecast
ARIMA, 471-11
automatic ARMA, 474-10
decomposition forecasting, 469-10
exponential smoothing, 465-8, 466-12, 467-10
Forecasts
multiple regression with serial correlation, 306-3
Forest plot
meta analysis of hazard ratios, 458-17
meta-analysis of correlated proportions, 457-20
meta-analysis of means, 455-17
meta-analysis of proportions, 456-20
Format, 102-6
Forward selection
Cox regression, 565-23
logistic regression, 320-17
Poisson regression, 325-6
Forward selection with switching
logistic regression, 320-18
multiple regression, 305-24
Poisson regression, 325-7
Forward variable selection
multiple regression, 305-23
Fourier plot
spectral analysis, 468-10
Fourier series
spectral analysis, 468-2
Fprob transformation, 119-9
Fraction transformation, 119-7
Fractional-factorial designs, 261-1
F-ratio
linear regression, 300-47
Freeman-Tukey standardized residual
loglinear models, 530-20
Frequency
spectral analysis, 468-1
Frequency polygon
histograms, 143-13
Frequency tables, 500-1
Frequency variable
linear regression, 300-25
Poisson regression, 325-8
Friedman’s Q statistic, 211-15
Friedman’s rank test, 211-3
FRUIT dataset, 141-1, 141-17
F-test
multiple regression, 305-50
FT-SR
loglinear models, 530-20
Full matching, 123-3
Function plots, 160-1
Functions
nonlinear regression, 315-4
Fuzz factor
filter, 121-2
in filter comparisons, 103-8
Fuzzifier
fuzzy clustering, 448-5
Fuzzy clustering, 448-1
FUZZY dataset, 448-3, 448-8
Fvalue transformation, 119-9

G

G statistic test
Poisson regression, 325-3
Gamma
hierarchical clustering, 445-8
Gamma distribution
probablility calculator, 135-4
simulation, 122-7
Gamma distribution fitting, 552-1
GammaProb transformation, 119-9
GammaValue transformation, 119-9
Gap between bars
bar charts, 141-14
Gap between sets of bars
bar charts, 141-15
Gart-Nam test
two proportions, 515-11
Gehan test
Kaplan-Meier, 555-12
nondetects analysis, 240-3
Geisser-Greenhouse adjustment, 214-1, 214-5

Geisser-Greenhouse epsilon, 214-4, 214-20
General linear models, 212-1
Generating data, 122-1
Generations
- hybrid appraisal models, 487-8

Geometric mean
- descriptive statistics, 200-14

Gleason-Staelin redundancy measure
- principal components analysis, 425-17

GLM
- checklist, 212-18

Gompertz model
- curve fitting, 351-7
- growth curves, 360-5

Goodness of fit
- loglinear models, 530-4
- Poisson regression, 325-3

Goodness-of-fit
- hierarchical clustering, 445-4
- K-means clustering, 446-2
- multidimensional scaling, 435-3
- ratio of polynomials, 370-2

Goto in output, 106-4
Graeco-Latin square designs, 263-1
Greedy matching, 123-1, 123-2
Greenwood's formula
- Kaplan-Meier, 555-3, 555-29, 555-33
- Weibull fitting, 550-3

Grid / tick settings window, 185-1
Grid lines, 185-1
Grid plot style file, 173-8
Grid plots, 140-11, 173-1
- response surface regression, 330-19

Grid range
- hybrid appraisal models, 487-9

Group average
- double dendrograms, 450-2
- hierarchical clustering, 445-4

Group variables
- logistic regression, 320-19

Growth curves, 360-1

H

HAIR dataset, 220-103
Harmonic mean
- descriptive statistics, 200-14

Hat diagonal
- linear regression, 300-19, 300-62
- multiple regression, 305-18, 305-64

Hat matrix
- linear regression, 300-18
- logistic regression, 320-14
- multiple regression, 305-18
- Poisson regression, 325-34

Hat values
- Poisson regression, 325-5

Hazard
- baseline, 565-8
- cumulative, 565-3
- Nelson-Aalen, 555-4

Hazard function
- beta distribution fitting, 551-2
- Cox regression, 565-2
- gamma distribution fitting, 552-2

Hazard function plot
- Kaplan-Meier, 555-36

Hazard rate
- Kaplan-Meier, 555-2
- life-table analysis, 570-3
- Weibull fitting, 550-2, 550-36

Hazard rate plot
- Kaplan-Meier, 555-36

Hazard ratio
- confidence interval, 555-40
- Kaplan-Meier, 555-40

Hazard ratio test
- Kaplan-Meier, 555-41

Hazard ratios
- meta analysis, 458-1

Hazard-baseline
- Cox regression, 565-38

HEART dataset, 212-23
Heat map colors, 187-5
Heat map settings window, 187-1
Help system, 1-10, 100-1
Heterogeneity test
- meta-analysis of proportions, 456-4

Heteroscedasticity
- linear regression, 300-3

Hierarchical cluster analysis, 450-1
- dendrograms, 450-3

Hierarchical clustering, 445-1
Hierarchical models
- Cox regression, 565-23
- loglinear models, 530-3
- multiple regression, 305-32
- response surface regression, 330-1

Hierarchical-classification designs, 212-27
Histogram
- bootstrap, 300-31, 305-42
- definition, 140-2
- density trace, 143-1
- descriptive statistics, 200-25
- linear regression, 300-34
- multiple regression, 305-67
- t-test, 205-20
- Xbar R, 250-32

Histogram style file, 143-16
Histograms, 140-2, 143-1
Histograms - comparative, 140-4, 151-1
Histograms – comparative
- density trace, 151-2

Holliday model
- curve fitting, 351-5
- growth curves, 360-4

Holt's linear trend, 466-1
Holt-Winters forecasting
- exponential smoothing, 467-1

Hotelling's one sample T2, 405-1
Hotelling's T2, 410-1
- 1-Sample, 405-1

Hotelling's T2 distribution
- probablility calculator, 135-4

Hotelling's T2 value, 410-7
Hotelling's two-sample T2, 410-1
Hour format, 102-8
HOUSING dataset, 306-4, 306-10
Hsu's test
- one-way ANOVA, 210-6

Huber's method
- multiple regression, 305-26

Huynh Feldt epsilon, 214-20
Huynh-Feldt adjustment, 214-1
Hybrid appraisal models, 487-1
Hybrid model, 487-1
HYP(z)
- piecewise polynomial models, 365-6

Hypergeometric distribution
- probablility calculator, 135-4

HypergeoProb transformation, 119-9
Hypothesis tests
- linear regression, 300-6
- multiple regression, 305-13

I

Identicalness
- curve fitting, 350-6

IEEE format, 102-4
If-then transformations, 120-1
Import limitations, 115-1
Importing, 103-2
Importing data, 12-1, 115-1
Imputation, 118-1
- principal components analysis, 425-4

Imputing data values, 118-1
Incidence
- Poisson regression, 325-1

Incidence rate
- Poisson regression, 325-34

Inclusion points
- D-optimal designs, 267-6

Incomplete beta function ratio
- beta distribution fitting, 551-2

Independence tests
- cross tabulation, 501-1

Independent variable

  - linear regression, 300-25
- Independent variables
  - logistic regression, 320-20
  - multiple regression, 305-1
  - multiple regression, 305-28
  - Poisson regression, 325-8
- Indicator variables
  - creating, 119-19
  - multiple regression, 305-3
- Individuals
  - hybrid appraisal models, 487-8
- Individuals chart, 250-4
  - Xbar R, 250-33
- Inertia
  - correspondence analysis, 430-13
- Influence
  - multiple regression, 305-17
- Influence report
  - linear regression, 300-66
- Influence detection
  - linear regression, 300-65
- Information matrix
  - Cox regression, 565-7
- Inheritance
  - hybrid appraisal models, 487-9
  - Weibull fitting, 550-15
- Initial communality
  - factor analysis, 420-3
- Initial Tableau section
  - linear programming, 480-4
- Initial values
  - backcasting, 465-2, 466-3, 467-3
- Insert, 103-5
- Installation, 1-1, 100-1
  - folders, 1-1
- Int transformation, 119-7
- INTEL dataset, 465-7, 466-9, 471-7, 473-5
- Interaction
  - two level designs, 260-3
- Interactions
  - multiple regression, 305-4
- Intercept
  - linear regression, 300-25, 300-39
  - multiple regression, 305-34
  - Poisson regression, 325-15
- Interquartile range
  - box plot, 152-1
  - descriptive statistics, 200-17
- Interval censored
  - parametric survival regression, 566-3
  - Weibull fitting, 550-11
- Interval data
  - Cox regression, 565-17
- Interval failure
  - Kaplan-Meier, 555-15
- Interval variables
  - fuzzy clustering, 448-4
  - hierarchical clustering, 445-6
  - medoid partitioning, 447-1
- Intervals
  - tolerance, 585-1
- Inverse prediction
  - linear regression, 300-6, 300-41, 300-67, 300-68
- IQ dataset, 305-27, 305-43, 305-72, 305-76, 305-79
- IQR
  - descriptive statistics, 200-17
- Isolines, 140-11
  - contour plot, 172-1
- Item analysis, 505-1
- ITEM dataset, 505-2, 505-5, 506-2, 506-6
- Item response analysis, 506-1

J

- Jittering
  - dot plots, 150-1
- Join transformation, 119-18
- Julian date transformation, 119-6

K

- K analysis
  - ridge regression, 335-22
- K values
  - ridge regression, 335-8
- Kaplan-Meier
  - Weibull fitting, 550-1
- Kaplan-Meier estimates, 555-1
- Kaplan-Meier product limit estimator
  - Weibull fitting, 550-3
- Kaplan-Meier product-limit, 555-32
  - beta distribution fitting, 551-14
  - gamma distribution fitting, 552-16
  - nondetects analysis, 240-14
  - Weibull fitting, 550-33
- Kaplan-Meier product-limit estimator
  - beta distribution fitting, 551-2
- Kappa reliability test
  - cross tabulation, 501-15
- Kaufman and Rousseeuw
  - medoid partitioning, 447-4
- Kendall’s coefficient
  - concordance, 211-15
- Kendall's tau-B
  - cross tabulation, 501-15
- Kendall's tau-C
  - cross tabulation, 501-15
- Kenward and Roger method
  - mixed models, 220-28
- Kernel-smoothed estimators
  - Kaplan-Meier, 555-9
  - Weibull fitting, 550-35
- Keyboard
  - commands, 103-11
- KLEIN6 dataset, 555-45
- K-means cluster analysis, 446-1
- KOCH36 dataset, 325-7, 325-21
- Kolmogorov-Smirnov
  - descriptive statistics, 200-23
- Kolmogorov-Smirnov test
  - two-sample, 206-1, 206-23
- Kruskall-Wallis test statistic, 210-21
- Kruskal-Wallis test, 210-1
- Kruskal-Wallis Z test
  - one-way ANOVA, 210-7
- Kurtosis, 200-2
  - descriptive statistics, 200-18
  - t-test, 205-15

L

- L’Abbe plot
  - meta-analysis of correlated proportions, 457-22
  - meta-analysis of means, 455-18
  - meta-analysis of proportions, 456-22
- Labeling values, 102-10
- Labeling variables, 2-4
- Labels
  - values, 13-1
- LACHIN91 dataset, 320-71
- Lack of fit
  - linear regression, 300-16
- Lack-of-fit test
  - response surface regression, 330-1
- Lagk transformation, 119-16
- Lambda
  - canonical correlation, 400-10
  - discriminant analysis, 440-12
  - loglinear models, 530-18
- Lambda A
  - cross tabulation, 501-14
- Lambda B
  - cross tabulation, 501-15
- Latin square designs, 263-1
- LATINSQR dataset, 212-22
- Latin-square
  - GLM, 212-21
- Lawley-Hotelling trace
  - MANOVA, 415-3
- Lcase transformation, 119-18
- LEAD dataset, 240-19
- Least squares
  - linear regression, 300-5
  - multiple regression, 305-13
- Least squares trend, 466-1
- Ledk transformation, 119-16

LEE91 database, 570-15
LEE91 dataset, 570-4
Left censored
 parametric survival regression, 566-3
 Weibull fitting, 550-11
Left transformation, 119-18
Length transformation, 119-18
LEUKEMIA dataset, 320-18, 320-34, 320-57
Levenberg-Marquardt algorithm, 385-1
Levene test
 linear regression, 300-27
 modified, 206-20
 modified (multiple-groups), 210-18
Levene test (modified)
 linear regression, 300-50
Levey-Jennings control charts, 252-1
Life-table analysis, 570-1
Like. ratio chi-square
 loglinear models, 530-13
Likelihood
 Cox regression, 565-5
Likelihood ratio
 1-sample binary diagnostic test, 535-3
 logistic regression, 320-8
 ROC curves, 545-24
Likelihood ratio test
 Cox regresion, 565-10
Likelihood ratio test of difference
 two proportions, 515-8
Likelihood-ratio statistic
 loglinear models, 530-4
Likert-scale
 simulation, 122-8, 122-22
Lilliefors' critical values
 descriptive statistics, 200-23
Limitations
 exporting, 116-1
Line charts, 140-1, 141-1
Line granularity
 linear regression, 300-33
Line settings window, 183-1
Linear discriminant functions
 discriminant analysis, 440-2
Linear model, 212-1
Linear programming, 480-1
Linear regression, 300-1
 assumptions, 300-3
Linearity
 MANOVA, 415-5
 multiple regression, 305-6
Linear-linear fit
 curve fitting, 351-11
Linear-logistic model, 320-1
Linkage type
 hierarchical clustering, 445-7
LINREG1 dataset, 300-24, 300-37
Ljung statistic
 automatic ARMA, 474-12
LLM, 530-1
Ln(X) transformation, 119-7
Loading a database, 2-1, 2-10, 7-1
Loess
 robust, 300-14
LOESS
 linear regression, 300-13
LOESS %N
 linear regression, 300-33
LOESS curve
 linear regression, 300-33
LOESS order
 linear regression, 300-33
LOESS robust
 linear regression, 300-34
Loess smooth
 scatter plot, 161-14
Log document, 106-1
Log file
 tutorial, 101-4
Log likelihood
 Poisson regression, 325-23
 Weibull fitting, 550-30
Log odds ratio transformation
 logistic regression, 320-2
Log of output, 9-6
Log transformation, 119-7
Logarithmic fit
 curve fitting, 351-8
LogGamma transformation, 119-9
Logistic distribution
 Weibull fitting, 550-10
Logistic item characteristic curve
 item response analysis, 506-1
Logistic model
 curve fitting, 351-6
 growth curves, 360-5
Logistic regression, 320-1
 parametric survival regression, 566-1
Logit transformation, 119-7
 logistic regression, 320-1
LOGLIN1 dataset, 530-7, 530-11
Loglinear models, 530-1
Log-logistic distribution
 Weibull fitting, 550-10
Log-logistic regression, 566-1
Lognormal
 curve fitting, 351-10, 351-11
 growth curves, 360-9
Lognormal distribution
 nondetects regression, 345-2
 Weibull fitting, 550-5
Lognormal regression, 566-1
Logrank test
 Kaplan-Meier, 555-41
Log-rank test
 Kaplan-Meier, 555-12
 nondetects analysis, 240-3
 randomization, 555-1
Log-rank tests
 Kaplan-Meier, 555-38
Longitudinal data example
 mixed models, 220-51
Longitudinal data models
 mixed models, 220-4
Longitudinal models, 220-1
Lookup transformation, 119-14
Lotus 123 exporting, 116-1
Lotus 123 importing, 115-1
Lowess smooth
 scatter plot, 161-14
LP dataset, 480-2, 480-4
LUNGCANCER dataset, 565-15, 565-31, 565-48

M

MA order (Q)
 automatic ARMA, 474-8
Macros, 130-1
 command list, 130-25
 commands, 130-6
 examples, 130-26
 syntax, 130-2
MAD
 descriptive statistics, 200-20
MAD constant
 multiple regression, 305-40
MAE
 exponential smoothing, 466-4, 467-2
Mallow's Cp
 variable selection and, 312-8
Mallow's Cp statistic
 multiple regression, 305-55
MAMMALS dataset, 3-1, 4-1, 10-1
MAMMALS1 dataset, 5-1, 6-1
Manhattan distance
 medoid partitioning, 447-3
Mann-Whitney U test, 206-1, 206-20
MANOVA, 415-1
 multivariate normality and Outliers, 415-4
MANOVA1 dataset, 410-3, 410-6, 415-5, 415-10
Mantel Haenszel test
 two proportions, 515-7
Mantel-Haenszel logrank test
 Kaplan-Meier, 555-41
Mantel-Haenszel test, 525-1
MAPE
 exponential smoothing, 466-4, 467-2
Maps
 contour plots, 172-1
 contour plots, 140-11
Mardia-Watson-Wheeler test

circular data, 230-10
Marginal association
loglinear models, 530-6
Martinez-Iglewicz
descriptive statistics, 200-22
Martingale residuals
Cox regression, 565-13, 565-39
Cox regression, 565-40
MARUBINI dataset, 560-3, 560-9
Mass
correspondence analysis, 430-13
Matched pairs
correlated proportions, 520-1
Matching
caliper, 123-4
caliper radius, 123-5
distance calculation method, 123-3
forced match variable, 123-4
full (variable), 123-3
greedy, 123-1, 123-2
optimal, 123-1, 123-2
propensity score, 123-2
standardized difference, 123-15
Mathematical functions
transformations, 119-7
Matrix determinant
equality of covariance, 402-8
Matrix type
principal components analysis, 425-11
Mauchley’s test of compound symmetry, 214-5
Mavk transformation, 119-16
Max % change in any beta
multiple regression, 305-78
Max terms
multiple regression, 305-33
Max transformation, 119-16
Maximum likelihood
Cox regression, 565-5
mixed models, 220-17
Weibull fitting, 550-10
Maximum likelihood estimates
beta distribution fitting, 551-12
McHenry's select algorithm, 310-1
McNemar test
correlated proportions, 520-1, 520-6
cross tabulation, 501-16
McNemar's tests, 501-1
MDB exporting, 116-1
MDB importing, 115-1
MDS, 435-1
MDS2 dataset, 435-6, 435-10, 435-15
Mean
confidence interval for, 200-13
descriptive statistics, 200-13
deviation, 200-20
geometric, 200-14
harmonic, 200-14
standard error of, 200-13
Mean absolute deviation
descriptive statistics, 200-20
Mean deviation
descriptive statistics, 200-20
estimate of standard error of, 200-20
Mean square
linear regression, 300-47
Mean squared error
linear regression, 300-19
multiple regression, 305-19
Mean squares
multiple regression, 305-50
Mean-deviation
descriptive statistics, 200-20
Means
meta-analysis of means, 455-1
Measurement error
R & R, 254-19
Measurement error ratio
R & R, 254-3
Median
cluster method, 445-4
confidence interval, 200-14
descriptive statistics, 200-14
Median cluster method
double dendrograms, 450-2
Median remaining lifetime
life-table analysis, 570-4, 570-22
Median smooth
scatter plot, 161-15
Median survival time
Kaplan-Meier, 555-30
Medoid clustering, 447-1
Medoid partitioning, 447-1
Membership
fuzzy clustering, 448-1
Merging two databases, 104-1
M-estimators
multiple regression, 305-25
Meta-analysis
correlated proportions, 457-1
Meta-analysis of hazard ratios, 458-1
Meta-analysis of means, 455-1
Meta-analysis of proportions, 456-1
METACPROP dataset, 457-6, 457-14
METAHR dataset, 458-6, 458-12
Method of moments estimates
beta distribution fitting, 551-12
Metric multidimensional scaling, 435-5
Michaelis-Menten
curve fitting, 351-1, 351-4
Miettinen - Nurminen test
two proportions, 515-8
Mill’s ratio
Kaplan-Meier, 555-2
Weibull fitting, 550-2
Min transformation, 119-16
Minimum Percent Beta Change, 305-40
Minute format, 102-8
Missing
if-then transformation, 120-4
Missing value estimation
factor analysis, 420-7
Missing values, 102-5, 320-8, 320-18, 425-4
cross tabs, 501-4
descriptive tables, 201-7
estimating, 118-1
GLM, 212-19
principal components analysis, 425-3
Missing-value imputation
principal components analysis, 425-4
Mixed model
defined, 220-2
Mixed models, 220-1
AIC, 220-7
Bonferroni adjustment, 220-14
covariates, 220-9
differential evolution, 220-29
F test, 220-28
Fisher scoring, 220-29
fixed effects, 220-9
G matrix, 220-18
Kenward and Roger method, 220-28
L matrix, 220-26
likelihood formulas, 220-17
maximum likelihood, 220-17
MIVQUE, 220-29
model building, 220-13
multiple comparisons, 220-14
Newton-Raphson, 220-29
R matrix, 220-19
random vs repeated error, 220-7
restricted maximum likelihood, 220-18
technical details, 220-16
time, 220-11
types, 220-4
zero variance estimate, 220-8
Mixture design
D-optimal designs, 267-22
MLCO2 dataset, 470-11
Mod transformation, 119-7
Mode
descriptive statistics, 200-15
Model
Bleasdale-Nelder, 351-5, 360-3
exponential, 351-6, 360-4
Farazdaghi and Harris, 351-5, 360-3
four-parameter logistic, 351-7, 360-5
Gompertz, 351-7, 360-5

  - Holliday, 351-5, 360-4
  - Kira, 351-4, 360-2
  - monomolecular, 351-6, 360-4
  - Morgan-Mercer-Floding, 351-8, 360-6
  - multiple regression, 305-33
  - reciprocal, 351-4, 360-2
  - Richards, 351-8, 360-7
  - Shinozaki, 351-4, 360-2
  - three-parameter logistic, 351-6, 360-5
  - Weibull, 351-7, 360-6
- Model size
  - all possible regressions, 312-8
- Models
  - growth curves, 360-1
  - hierarchical, 530-3
  - multiphase, 365-1
  - multiple regression, 305-35
  - piecewise polynomial, 365-1
  - ratio of polynomials, 370-1, 375-1
  - sum of functions, 380-1
  - user written, 385-1
- Modified Kuiper's test
  - circular data, 230-4
- Moment
  - descriptive statistics, 200-11
- Monomolecular model
  - curve fitting, 351-6
  - growth curves, 360-4
- Monte Carlo samples
  - 1-Sample T2, 405-4
  - linear regression, 300-31
- Monte Carlo simulation, 122-1
- Month format, 102-8
- Month transformation, 119-6
- Morgan-Mercer-Floding model
  - curve fitting, 351-8
  - growth curves, 360-6
- MOTORS dataset, 566-3, 566-11
- Moving average chart, 250-4
- Moving average chart limits, 250-8
- Moving average parameters
  - ARIMA, 471-3
  - theoretical ARMA, 475-2
- Moving data, 103-14
- Moving range
  - Xbar R, 250-33
- Moving range chart, 250-4
- MSEi
  - multiple regression, 305-19
- Multicollinearity
  - canonical correlation, 400-2
  - discriminant analysis, 440-4
  - MANOVA, 415-5
  - multiple regression, 305-7
  - ridge regression, 335-1
  - stepwise regression, 311-2
- Multicollinearity report
  - multiple regression, 305-57
- Multidimensional scaling, 435-1
  - metric, 435-1
- Multinomial chi-square tests
  - frequency tables, 500-1
- Multinomial distribution
  - simulation, 122-8
- Multinomial test
  - frequency tables, 500-10
- Multiple comparisons
  - Bonferroni, 210-4
  - box plots, 152-2
  - Duncan's test, 210-5
  - Dunn's test, 210-7
  - Dunnett's test, 210-6
  - Fisher's LSD, 210-6
  - Hsu's test, 210-6
  - Kruskal-Wallis Z test, 210-7
  - mixed models, 220-14
  - Newman-Keuls test, 210-8
  - one-way ANOVA, 210-3
  - recommendations, 210-8
  - Scheffe's test, 210-8
  - Tukey-Kramer test, 210-8
- Multiple regression
  - robust, 305-24
- Multiple regression, 305-1
  - assumptions, 305-6
- Multiple regression
  - all possible, 312-1
- Multiple regression
  - binary response, 320-8
- Multiple regression with serial correlation, 306-1
- Multiplicative seasonality
  - exponential smoothing, 467-2
- Multiplicity factor
  - t-test, 205-19
- Multivariate analysis of variance, 415-1
- Multivariate normal
  - factor analysis, 420-7
  - principal components analysis, 425-11
- Multivariate polynomial ratio fit, 376-1
- Multivariate variable selection, 310-1
- Multiway frequency analysis
  - loglinear models, 530-1
- Mutation rate
  - hybrid appraisal models, 487-9
  - Weibull fitting, 550-15

N

- Nam and Blackwelder test
  - correlated proportions, 520-5
- Nam test
  - correlated proportions, 520-7
- Nam's score
  - correlated proportions, 520-2
- Navigator, 107-1
- NC CRIMINAL dataset, 320-64, 320-68
- NcBetaProb transformation, 119-9
- NcBetaValue transformation, 119-10
- NcCsProb transformation, 119-10
- NcCsValue transformation, 119-10
- NcFprob transformation, 119-10
- NcFvalue transformation, 119-10
- NCSS
  - quitting, 101-4
- NcTprob transformation, 119-10
- NcTvalue transformation, 119-10
- Nearest neighbor
  - double dendrograms, 450-2
  - hierarchical clustering, 445-3
- Negative binomial distribution
  - probablility calculator, 135-5
- Negative binomial transformation, 119-10
- NegBinomProb transformation, 119-10
- Neighborhood
  - appraisal ratios, 485-7
- Nelson-Aalen estimates
  - Weibull fitting, 550-1
- Nelson-Aalen estimator, 555-7
  - Weibull fitting, 550-33
- Nelson-Aalen hazard
  - Kaplan-Meier, 555-1
  - Weibull fitting, 550-4
- Nested factor
  - GLM, 212-4
- Nested factors
  - design generator, 268-1
- New database, 103-1
- New spreadsheet, 103-1
- New template, 105-1
- Newman-Keuls test
  - one-way ANOVA, 210-8
- Newton-Raphson
  - Weibull fitting, 550-11
- Nominal variables
  - fuzzy clustering, 448-4
  - hierarchical clustering, 445-7
  - medoid partitioning, 447-2
- Non-central Beta transformation, 119-10
- Non-central Chi-square transformation, 119-10
- noncentral-F distribution transformation, 119-10
- Noncentral-t distribution transformation, 119-10
- Nondetects analysis, 240-1
  - confidence limits, 240-7
  - flipping constant, 240-2
  - Gehan test, 240-3

  - Kaplan-Meier product-limit, 240-14
  - log-rank test, 240-3
  - Peto-Peto test, 240-3
  - Tarone-Ware test, 240-3
- NONDETECTS dataset, 240-4
- Nondetects regression, 345-1
  - confidence limits, 345-11
  - Cox-Snell residual, 345-13
  - R-squared, 345-11
  - standardized residual, 345-13
- Noninferiority
  - 2-sample binary diagnostic, 537-10
  - clustered binary diagnostic, 538-9
  - paired binary diagnostic, 536-8
  - ROC curves, 545-31
- Noninferiority test
  - correlated proportions, 520-8
  - two proportions, 515-17
- Noninferiority tests
  - two proportions, 515-37
- Nonlinear regression, 315-1
  - appraisal, 487-1
  - functions, 315-4
  - starting values, 315-1
  - user written models, 385-1
- Nonparametric tests
  - t-test, 205-17
- Nonstationary models
  - Box Jenkins, 470-3
- Normal
  - curve fitting, 351-10
  - growth curves, 360-9
- Normal distribution
  - probablility calculator, 135-5
  - simulation, 122-9, 122-20
  - Weibull fitting, 550-4
- Normal line
  - histograms, 143-12
- Normal probability plot
  - descriptive statistics, 200-26
- Normality, 200-4
  - descriptive statistics, 200-21
  - ROC curves, 545-12
  - t-test, 205-15
- Normality test alpha, 118-3
- Normality tests
  - Anderson-Darling test, 200-22
  - D’Agostino kurtosis, 200-24
  - D’Agostino omnibus, 200-25
  - D’Agostino skewness, 200-23
  - descriptive statistics, 200-21
  - Kolmogorov-Smirnov, 200-23
  - Lilliefors' critical values, 200-23
  - linear regression, 300-48
  - Martinez-Iglewicz, 200-22
  - multiple regression, 305-52
  - Shapiro-Wilk test, 200-22
  - skewness test, 200-24
  - tolerance intervals, 585-11
- NormalProb transformation, 119-10
- NormalValue transformation, 119-10
- NormScore transformation, 119-16
- Notes
  - omitting them in linear regression, 300-26
- NP-chart, 251-1
- Number exposed
  - life-table analysis, 570-2
- Number of correlations
  - canonical correlation, 400-5
- Number of points
  - linear regression, 300-33
- Numeric data, 102-1
- Numeric functions, 119-6

O

- Objective function
  - linear programming, 480-1
- Observational study matching, 123-1
- Observational study stratification, 124-1
- Odds ratio
  - 1-sample binary diagnostic test, 535-4
  - 2-sample binary diagnostic, 537-9
  - confidence interval of, 515-23
  - correlated proportions, 520-5
  - meta-analysis of correlated proportions, 457-2
  - meta-analysis of proportions, 456-2
  - two proportions, 515-1, 515-3
- Odds ratios
  - Mantel-Haenszel test, 525-1
- ODOR dataset, 330-3, 330-11
- Omission report
  - multiple regression, 305-54
- One proportion, 510-1
- One-sample tests, 205-1
- One-sample t-test, 205-1
- One-way analysis of variance, 210-1
- One-way ANOVA
  - Bonferroni, 210-4
  - Duncan’s test, 210-5
  - Dunn’s test, 210-7
  - Dunnett’s test, 210-6
  - Fisher's LSD, 210-6
  - Hsu’s test, 210-6
  - Kruskal-Wallis Z test, 210-7
  - multiple comparisons, 210-3
  - Newman-Keuls test, 210-8
  - orthogonal contrasts, 210-11
  - orthogonal polynomials, 210-11
  - planned comparisons, 210-10
  - Scheffe’s test, 210-8
  - Tukey-Kramer test, 210-8
- Open database, 103-1
- Open log file, 106-2
- Open output file, 106-2
- Open spreadsheet, 103-1
- Open template, 105-1
- Opening a database
  - tutorial, 101-3
- Optimal matching, 123-1, 123-2
- Optimal solution section
  - linear programming, 480-5
- Optimal value
  - linear programming, 480-5
- Or
  - if-then transformation, 120-2
- Ordinal variables
  - fuzzy clustering, 448-4
  - hierarchical clustering, 445-6
  - medoid partitioning, 447-2
- Original cost
  - linear programming, 480-5
- Orthogonal arrays, 266-1
- Orthogonal contrasts
  - one-way ANOVA, 210-11
- Orthogonal polynomial
  - ANOVA balanced, 211-6
  - GLM, 212-5
  - repeated measures, 214-11
- Orthogonal polynomials
  - one-way ANOVA, 210-11
- Orthogonal regression
  - linear regression, 300-9, 300-41
- Orthogonal sets of Latin squares, 263-2
- Outlier detection
  - linear regression, 300-64
  - multiple regression, 305-83
- Outlier report
  - linear regression, 300-66
- Outliers
  - Cox regression, 565-14
  - linear regression, 300-15
  - multiple regression, 305-1, 305-24, 305-78
  - stepwise regression, 311-3
  - t-test, 205-22
- Outliers, 200-3
- Output, 106-1
  - log of, 9-6
  - printing, 9-4
  - ruler, 106-4
  - saving, 9-5
- Output document, 106-1
- Output window, 1-6, 9-1
- Overdispersion
  - Poisson regression, 325-3, 325-12
- Overlay
  - scatter plot, 161-3

P

Page setup, 103-2
PAIN dataset, 220-51
Paired data
    clustered binary diagnostic, 538-11
Paired t-test
    1-Sample T2, 405-1
Paired t-tests, 205-1
Pair-wise removal
    correlation matrix, 401-3
Paradox exporting, 116-1
Paradox importing, 115-1
Parallel slopes
    multiple regression, 305-86
Parameterization
    curve fitting, 350-5
Pareto chart, 253-1
Pareto charts, 250-41
Parsimony
    ratio of polynomials, 370-2
Partial association
    loglinear models, 530-5
Partial autocorrelation, 472-1
Partial autocorrelation function
    Box Jenkins, 470-4
Partial correlation
    multiple regression, 305-56
Partial residual plots, 305-71
Partial variables
    canonical correlation, 400-4
    correlation matrix, 401-3
Partial-regression coefficients, 305-47
Partition coefficient
    fuzzy clustering, 448-3
Paste, 103-4
Paste output, 106-3
Pasting data, 7-2
PCA, 425-1
PCA2 dataset, 118-4, 420-5, 420-11, 425-9, 425-15
P-chart, 251-1
Pearson chi-square
    loglinear models, 530-4, 530-13
Pearson correlation
    linear regression, 300-45
Pearson correlations
    matrix of, 401-1
Pearson residuals
    logistic regression, 320-13
    Poisson regression, 325-5, 325-31
Pearson test
    Poisson regression, 325-3
Pearson’s contingency coefficient
    cross tabulation, 501-14
Percentile plots, 140-5
Percentile Plots, 153-1
Percentile type
    descriptive statistics, 200-6
Percentiles, 200-2
Percentiles of absolute residuals
    multiple regression, 305-78
Period effect
    cross-over analysis using t-tests, 235-4
Period plot
    cross-over analysis using t-tests, 235-24
Periodogram
    spectral analysis, 468-1
Perspective
    3D scatter plot, 170-6
    3D surface plot, 171-6
    bar charts, 141-12
PET dataset, 538-11
Peto-Peto test
    Kaplan-Meier, 555-12
    nondetects analysis, 240-3
Phase
    spectral analysis, 468-1
Phi
    cross tabulation, 501-14
    factor analysis, 420-13
    Poisson regression, 325-3, 325-12, 325-27
    principal components analysis, 425-17
Phis
    theoretical ARMA, 475-2
Pie charts, 140-2, 142-1
PIE dataset, 142-6
Piecewise polynomial models, 365-1
Pillai's trace
    MANOVA, 415-3
Plackett-Burman design, 265-1
Planned comparisons
    one-way ANOVA, 210-10
PLANT dataset, 212-27
Plot size
    linear regression, 300-29
Plots
    3D scatter plots, 140-10, 170-1
    3D surface plots, 140-10, 171-1
    area charts, 140-1, 141-1
    bar charts, 140-1, 141-1
    box plots, 140-5, 152-1
    contour plots, 140-11, 172-1
    density trace, 143-1
    dot plots, 140-4, 150-1
    error-bar charts, 140-6, 155-1
    function plots, 160-1
    grid plots, 140-11, 173-1
    histograms, 140-2, 143-1
    histograms - comparative, 140-4, 151-1
    line charts, 140-1, 141-1
    percentile plots, 140-5, 153-1
    pie charts, 140-2
    probability plots, 140-3, 144-1
    scatter plot matrix, 140-8, 162-1
    scatter plot matrix (curve fitting), 163-1
    scatter plot matrix for curve fitting, 140-9
    scatter plots, 140-7, 161-1
    single-variable charts, 140-1
    surface charts, 140-1, 141-1
    surface plots, 140-10, 171-1
    three-variable charts, 140-10
    two-variable charts, 140-4, 140-7
    violin plots, 140-6, 154-1
POISREG dataset, 325-37
Poisson distribution
    probablility calculator, 135-5
    simulation, 122-9
Poisson regression, 325-1
PoissonProb transformation, 119-11
POLITIC dataset, 13-1, 14-1
Polynomial
    logistic regression, 320-23
    multiple regression, 305-31
    multivariate ratio fit, 376-1
    Poisson regression, 325-11
Polynomial fit
    scatter plot, 161-13
Polynomial model
    response surface regression, 330-1
Polynomial models, 365-1
Polynomial ratio fit, 375-1
Polynomial ratios
    model search (many X variables), 371-1
Polynomial regression model, 330-1
Polynomials
    ratio of, 370-1, 375-1
Pooled terms, 213-2
POR exporting, 116-1
Portmanteau test
    ARIMA, 471-12
    automatic ARMA, 474-12
    Box Jenkins, 470-10
Power
    multiple regression, 305-47
Power spectral density
    spectral analysis, 468-3
Power spectrum
    theoretical ARMA, 475-8
PRD
    appraisal ratios, 485-8
Precision-to-tolerance
    R & R, 254-20
Precision-to-tolerance ratio
    R & R, 254-3
Predicted value
    Poisson regression, 325-32
Predicted values
    linear regression, 300-27, 300-52
    multiple regression, 305-61
Prediction interval

multiple regression, 305-61
Prediction limits
linear regression, 300-33, 300-53, 300-59
multiple regression, 305-61
Pre-post
multiple regression, 305-87
PREPOST dataset, 305-87
PRESS
linear regression, 300-21, 300-51
multiple regression, 305-21, 305-51
PRESS R2
multiple regression, 305-52
Press R-squared
multiple regression, 305-21
PRESS R-squared
linear regression, 300-22
Prevalence
ROC curves, 545-5
Price related differential
appraisal ratios, 485-8
hybrid appraisal models, 487-17
Principal axis method
factor analysis, 420-1
Principal components
linear regression, 300-9
Principal components analysis, 425-1
Principal components regression, 340-1
Print
output, 106-3
Printer setup, 103-2
Printing
data, 2-7, 103-3
output, 9-4
output reports, 4-5
Printing data, 117-1
Prior probabilities
discriminant analysis, 440-5
Prob level, 415-13
linear regression, 300-47
Prob to enter
stepwise regression, 311-4
Prob to remove
stepwise regression, 311-4
Probability Calculator, 135-1
Beta distribution, 135-1
Binomial distribution, 135-2
Bivariate normal distribution, 135-2
Chi-square distribution, 135-2
Correlation coefficient distribution, 135-3
F distribution, 135-3
Gamma distribution, 135-4
Hotelling's T2 distribution, 135-4
Hypergeometric distribution, 135-4
Negative binomial distribution, 135-5
Normal distribution, 135-5
Poisson distribution, 135-5
Student's t distribution, 135-6
Studentized range distribution, 135-6
Weibull distribution, 135-6
Probability ellipse
linear regression, 300-8, 300-33
Probability functions
transformations, 119-8
Probability plot
descriptive statistics, 200-26
linear regression, 300-57
multiple regression, 305-67
t-test, 205-20
Weibull, 144-17
Probability plot style file, 144-19
Probability plots, 140-3
asymmetry, 144-3
quantile scaling, 144-7
Probability Plots, 144-1
Probit analysis, 575-1
Probit plot
probit analysis, 575-10
Procedure, 105-1
running, 101-3
Procedure window, 1-5, 8-1
Product-limit survival distribution
beta distribution fitting, 551-14
gamma distribution fitting, 552-16
Kaplan-Meier, 555-32
Weibull fitting, 550-33
Product-moment correlation
correlation matrix, 401-3
Profiles
correspondence analysis, 430-1
Projection method
3D scatter plot, 170-8
3D surface plot, 171-7
bar charts, 141-14
PROPENSITY dataset, 123-5, 123-12, 124-4
Propensity score, 123-2
stratification, 124-1
Proportion trend test
Armitage, 501-5
Proportions
2-sample binary diagnostic, 537-1
clustered binary diagnostic, 538-1
confidence interval of ratio, 515-21
correlated, 520-1
Meta-analysis of correlated proportions, 457-1
meta-analysis of proportions, 456-1
one, 510-1
paired binary diagnostic, 536-1
two, 515-1
Proportions test
1-sample binary diagnostic test, 535-1
Proximity matrix
multidimensional scaling, 435-1
Proximity measures
multidimensional scaling, 435-4
Pseudo R-squared
multidimensional scaling, 435-12
Poisson regression, 325-4
Pure error
linear regression, 300-16

Q

QATEST dataset, 250-14, 250-27, 250-33, 250-35, 250-37, 251-3, 251-11, 253-2, 253-7, 253-9
Quadratic fit
curve fitting, 351-2
Qualitative factors
D-optimal designs, 267-6, 267-25
Quality
correspondence analysis, 430-13
Quantile scaling
probability plots, 144-7
Quantile test, 205-17
Quantiles
Kaplan-Meier, 555-30
Quartiles
descriptive statistics, 200-21
Quartimax rotation
factor analysis, 420-4
principal components analysis, 425-8
Quatro exporting, 116-1
Quick launch window, 107-1, 107-2
Quick start, 100-1
Quitting NCSS, 101-4

R

R & R study, 254-1
Radial plot
meta analysis of hazard ratios, 458-18
meta-analysis of correlated proportions, 457-21
meta-analysis of means, 455-18
meta-analysis of proportions, 456-21
Random coefficients example
mixed models, 220-103
Random coefficients models
mixed models, 220-5

Random effects model
- meta-analysis of correlated proportions, 457-5
- meta-analysis of hazard ratios, 458-5
- meta-analysis of means, 455-4
- meta-analysis of proportions, 456-5

Random effects models, 220-1
- mixed models, 220-4

Random factor
- ANOVA balanced, 211-5
- GLM, 212-4
- repeated measures, 214-8

Random numbers, 122-1
- uniform, 15-1

Randomization
- Latin square designs, 263-2

Randomization test
- curve fitting, 351-16
- linear regression, 300-24
- log-rank, 555-1
- T2, 410-7

Randomization tests
- 1-Sample T2, 405-1, 405-8
- T2, 410-1

Randomized block design
- repeated measures, 214-6

RandomNormal transformation, 119-11

Random-number functions
- transformations, 119-11

Range
- descriptive statistics, 200-17
- interquartile, 200-17

Range chart, 250-1

Rank transformation, 119-16

Rate ratio
- Poisson regression, 325-30

Ratio of polynomials
- model search (many X variables), 371-1
- model search (one X variable), 370-1

Ratio of polynomials fit, 375-1
- many variables, 376-1

Ratio of two proportions
- two proportions, 515-6

Ratio plot
- decomposition forecasting, 469-12

Ratio section
- appraisal ratios, 485-7

Ratio study
- appraisal ratios, 485-1

Ratio variables
- fuzzy clustering, 448-4
- hierarchical clustering, 445-6
- medoid partitioning, 447-2

Rayleigh test
- circular data, 230-4

Rbar-squared
- linear regression, 300-8
- multiple regression, 305-15

RCBD data example
- mixed models, 220-94

RCBD dataset, 220-94

REACTION dataset, 214-6, 214-29

Readout
- parametric survival regression, 566-3
- Weibull fitting, 550-11

READOUT105 dataset, 550-47

Rearrangement functions
- transformations, 119-12

Recalc all, 103-9, 119-4

Recalc current, 103-8, 119-4

Reciprocal model
- curve fitting, 351-4
- growth curves, 360-2

Recode functions transformations, 119-14

Recode transformation, 3-4, 119-15

Recoding, 11-1

Reduced cost
- linear programming, 480-5

Redundancy indices
- canonical correlation, 400-4

Reference group
- logistic regression, 320-19

Reference value
- logistic regression, 320-21
- multiple regression, 305-3, 305-29
- Poisson regression, 325-9
- Xbar R, 250-23

Reflection C.I. method
- multiple regression, 305-41

Reflection method
- linear regression, 300-30
- two proportions, 515-28

REGCLUS dataset, 449-2, 449-5

Regression
- all possible, 312-1
- appraisal model, 487-1
- backward selection, 311-2
- binary response, 320-1, 320-8
- clustering, 449-1
- Cox, 565-1
- diagnostics, 305-63
- exponential, 566-1
- extreme value, 566-1
- forward selection, 311-1
- growth curves, 360-1
- hybrid appraisal model, 487-1
- linear, 300-1
- logistic, 320-1, 566-1
- log-logistic, 566-1
- lognormal, 566-1
- model search (many X variables), 371-1
- multiple, 312-8
- nondetects, 345-1
- nonlinear, 315-1
- normal, 566-1
- orthogonal regression, 300-9
- Poisson, 325-1
- polynomial ratio, 375-1
- polynomial ratio (search), 370-1
- principal components, 340-1
- proportional hazards, 565-1
- response surface regression, 330-1
- ridge, 335-1
- stepwise, 311-1
- sum of functions models, 380-1
- user written, 385-1
- variable selection, 311-1
- Weibull, 566-1

Regression analysis, 6-1
- multiple regression, 305-1

Regression clustering, 449-1

Regression coefficients
- Cox regression, 565-32

Regression coefficients report
- multiple regression, 305-48

Regression equation report
- multiple regression, 305-46

Relative risk
- meta-analysis of correlated proportions, 457-2
- meta-analysis of proportions, 456-2
- two proportions, 515-1

Reliability
- beta distribution fitting, 551-1, 551-15
- gamma distribution fitting, 552-1
- item analysis, 505-1
- Kaplan-Meier, 555-1
- kappa, 501-15
- Weibull fitting, 550-1

Reliability analysis
- Weibull fitting, 550-1

Reliability function
- beta distribution fitting, 551-2
- gamma distribution fitting, 552-2
- Weibull fitting, 550-2

Remove last sheet, 103-2

Remove transformation, 119-18

Removed lambda
- discriminant analysis, 440-12

Repeat transformation, 119-18

Repeatability
- R & R, 254-1, 254-14

Repeated measures, 214-1
- 1-Sample T2, 405-6
- mixed models, 220-1

Repeated measures data example
- mixed models, 220-51

Repeated measures design
- generating, 268-7

Repeated-measures design

  GLM, 212-23
Replace, 103-6
Replace in output, 106-4
Replace transformation, 119-18
Replication
  two level designs, 260-4
Reporting data, 117-1
Reports
  selecting in linear regression, 300-26
Reproducibility
  R & R, 254-1, 254-14
RESALE dataset, 117-4, 151-14, 155-1, 155-7, 201-1, 201-11, 201-12, 201-14, 201-15, 201-17, 201-19, 201-21, 305-81, 500-1, 500-9, 500-10, 500-12, 500-14, 501-1, 501-8, 501-11, 501-17
Resampling tab
  linear regression, 300-29
Residual
  diagnostics, 305-63
  linear regression, 300-2, 300-18
  multiple regression, 305-17
Residual diagnostics
  linear regression, 300-15
  multiple regression, 305-15
  Poisson regression, 325-33
Residual life
  life-table analysis, 570-22
  Weibull fitting, 550-40
Residual plots
  linear regression, 300-53
  multiple regression, 305-67, 305-70
  partial residuals, 305-71
Residual report
  linear regression, 300-61
  multiple regression, 305-62
Residuals
  Cox regression, 565-13
  Cox regression, 565-39
  logistic regression, 320-11
  multiple regression, 305-1
  Poisson regression, 325-4, 325-31
Residuals-deviance
  Cox regression, 565-14
Residuals-Martingale
  Cox regression, 565-13
Residuals-scaled Schoenfeld
  Cox regression, 565-15
Residuals-Schoenfeld
  Cox regression, 565-14
Response surface regression, 330-1
Response-surface designs, 264-1
Restart method
  Xbar R, 250-23
Restricted maximum likelihood
  mixed models, 220-18
Richards model
  curve fitting, 351-8
  growth curves, 360-7
Ridge regression, 335-1
Ridge trace
  ridge regression, 335-4, 335-18
RIDGEREG dataset, 335-7, 335-15, 340-3, 340-11
Right censored
  parametric survival regression, 566-2
  Weibull fitting, 550-11
Right transformation, 119-19
Right-hand sides
  linear programming, 480-1
Risk ratio
  correlated proportions, 520-4
  Cox regression, 565-33, 565-35
  meta-analysis of correlated proportions, 457-2
  meta-analysis of proportions, 456-2
Risk set
  Cox regression, 565-16
  Kaplan-Meier, 555-3
RMSF dataset, 545-3
RNDBLOCK dataset, 211-4, 211-11, 212-3, 212-12
Robins odds ratio C. L.
  Mantel-Haenszel test, 525-11
Robust estimation
  principal components analysis, 425-5
Robust iterations
  Xbar R, 250-18
Robust loess
  linear regression, 300-14
Robust method
  multiple regression, 305-39
Robust regression
  multiple regression, 305-24, 305-31
Robust regression reports
  multiple regression, 305-77
Robust regression tutorial
  multiple regression, 305-76
Robust sigma multiplier
  Xbar R, 250-18
Robust tab
  multiple regression, 305-39
Robust weight
  factor analysis, 420-7
  principal components analysis, 425-11
Robust weights
  multiple regression, 305-78
ROC curves, 545-1
  comparing, 545-9
ROC dataset, 545-19
Root MSE
  all possible regressions, 312-8
Rose plot
  circular data, 230-16
Rose plots, 230-1
Rotation
  3D scatter plot, 170-7
  3D surface plot, 171-7
  bar charts, 141-13
  factor analysis, 420-7
  principal components analysis, 425-11
Round transformation, 119-7
Row heights, 103-15
Row profiles
  correspondence analysis, 430-1
Rows, 251-4, 251-5
Row-wise removal
  correlation matrix, 401-3
Roy's largest root
  MANOVA, 415-4
RRSTUDY dataset, 254-1, 254-10
RRSTUDY1 dataset, 254-24
R-squared
  adjusted, 300-46
  adjusted, 305-45
  all possible regressions, 312-8
  Cox regression, 565-11
  definition, 305-44
  linear regression, 300-7, 300-46
  logistic regression, 320-10
  multiple regression, 305-14
  Poisson regression, 325-4, 325-24
R-squared increment
  stepwise regression, 311-8
R-squared report
  multiple regression, 305-53
R-squared vs variable count plot, 310-8
RStudent
  linear regression, 300-20, 300-62
  multiple regression, 305-19, 305-63
RStudent plot
  multiple regression, 305-69
Rstudent residuals
  scatter plot of, 300-55
RTF, 106-3
  tutorial, 101-4
RTF output format, 106-1
Ruler
  output, 106-4
Run summary report
  multiple regression, 305-44
Running a procedure
  tutorial, 101-3
Running a regression analysis, 6-1
Running a two-sample t-test, 5-1
Running descriptive statistics, 4-1
Runs tests
  attribute charts, 251-3
  Xbar R, 250-9

S

S0 database, 102-1
S0/S0Z comparison, 102-4
S0Z/S0 comparison, 102-4
Sale date variable
    appraisal ratios, 485-4
    comparables, 486-7
Sale price variables
    appraisal ratios, 485-2
SALES dataset, 467-9, 469-9
Sales price
    multiple regression, 305-81
SALESRATIO dataset, 485-1, 485-6, 486-4
SAMPLE dataset, 101-3, 161-20, 162-5, 171-9, 172-7, 200-4, 200-10, 205-12, 206-12, 210-16, 310-3, 310-6, 311-3, 311-6, 312-2, 312-6, 400-8, 401-2, 401-5, 585-8
SAS exporting, 116-1
SAS importing, 115-1
Saturated model
    loglinear models, 530-3
Save, 103-3
Save as, 103-3
Save output, 106-3
Saved colors, 180-3
Saving
    data, 2-6
        tutorial, 101-2
    output, 9-5
    template, 8-5
Saving a template, 105-2
Saving results
    multiple regression, 305-42
SC
    medoid partitioning, 447-5
Scaled Schoenfeld residuals
    Cox regression, 565-15, 565-42
Scaling
    multidimensional, 435-1
Scaling factors
    D-optimal designs, 267-2
Scaling method
    fuzzy clustering, 448-5
    hierarchical clustering, 445-8
Scatter plot
    loess smooth, 161-14
    lowess smooth, 161-14
    median smooth, 161-15
    overlay, 161-3
    polynomial fit, 161-13
    spline, 161-15
    sunflower plot, 161-18
Scatter plot matrix, 140-8, 162-1
Scatter plot matrix (curve fitting), 163-1
Scatter plot matrix for curve fitting, 140-9
Scatter plot style file, 161-22
Scatter plots, 140-7, 161-1
    3D, 140-10, 170-1
Scheffe’s test
    one-way ANOVA, 210-8
Schoenfeld residuals
    Cox regression, 565-14, 565-41
Schuirmann’s test
    cross-over analysis using t-tests, 235-7
Scientific notation, 102-4
Score, 320-45
Score coefficients
    factor analysis, 420-17
    principal components analysis, 425-2
Scores plots
    canonical correlation, 400-12
Scree graph
    factor analysis, 420-3
Scree plot
    factor analysis, 420-15
    principal components analysis, 425-18
Screening data, 118-1, 200-3
Screening designs, 265-1
Searches
    ratio of polynomials, 370-1, 371-1
Seasonal adjustment
    exponential smoothing, 467-1
Seasonal autoregressive parameters
    ARIMA, 471-3
Seasonal decomposition forecasting, 469-1
Seasonal differencing
    ARIMA, 471-2
Seasonal moving average parameters
    ARIMA, 471-3
Seasonal time series
    Box Jenkins, 470-4
Second format, 102-8
Select all output, 106-4
Selecting procedures, 1-7
Selection method
    stepwise regression, 311-4
Selection procedure
    forward, 311-1
Sensitivity
    1-sample binary diagnostic test, 535-2
    2-sample binary diagnostic, 537-2
    clustered binary diagnostic, 538-8
    paired binary diagnostic, 536-2
    ROC curves, 545-1, 545-24
Sequence plot
    multiple regression, 305-69
Sequence transformation, 119-6
Sequential models report
    multiple regression, 305-56
Ser transformation, 119-6
Serial correlation
    linear regression, 300-4
    residuals, 305-53
Serial correlation plot
    multiple regression, 305-68
Serial numbers, 1-3, 100-1
Serial-correlation
    linear regression, 300-50
SERIESA dataset, 470-8, 474-7
Shapiro-Wilk
    linear regression, 300-18
    multiple regression, 305-17
Shapiro-Wilk test
    descriptive statistics, 200-22
    linear regression, 300-49
Shinozaki and Kari model
    curve fitting, 351-4
    growth curves, 360-2
Short transformation, 119-7
Sigma
    Xbar R, 250-19
Sigma multiplier
    Xbar R, 250-17
Sign test, 205-17
Sign transformation, 119-8
SIGN(z)
    piecewise polynomial models, 365-6
Signal-to-noise ratio
    R & R, 254-3
Silhouette
    fuzzy clustering, 448-9
    medoid partitioning, 447-13
Silhouettes
    medoid partitioning, 447-5
Similarities
    multidimensional scaling, 435-4
Simple average
    double dendrograms, 450-2
    hierarchical clustering, 445-3
Simplex algorithm
    linear programming, 480-1
Simulation, 122-1
    Beta distribution, 122-3
    Binomial distribution, 122-5
    Cauchy distribution, 122-5
    Constant distribution, 122-6
    contaminated normal, 122-21
    data, 15-1
    Exponential distribution, 122-6
    F distribution, 122-7
    Gamma distribution, 122-7
    Likert-scale, 122-8, 122-22
    Multinomial distribution, 122-8
    Normal distribution, 122-9, 122-20
    Poisson distribution, 122-9
    skewed distribution, 122-10

Student's T distribution, 122-10
syntax, 122-13
T distribution, 122-10
Tukey's lambda distribution, 122-10
Uniform distribution, 122-11
Weibull distribution, 122-12
Simultaneous C.I.'s
T2, 405-9, 410-10
Sin transformation, 119-17
Single linkage
double dendrograms, 450-2
hierarchical clustering, 445-3
Single-to-noise ratio
R & R, 254-19
Single-variable charts, 140-1
Sinh transformation, 119-17
Skewed distribution
simulation, 122-10
Skewness, 200-2
descriptive statistics, 200-17
t-test, 205-15
Skewness test
descriptive statistics, 200-24
Slices
pie charts, 142-1
Slope
linear regression, 300-39
Slopes
testing for equal
multiple regression, 305-86
SMOKING dataset, 525-2, 525-5
Smooth transformation, 119-16
Smoothing constant
exponential smoothing, 465-1, 466-2
Smoothing constants
exponential smoothing, 467-2
Smoothing interval
item response analysis, 506-4
Solo exporting, 116-1
Solo exporting, 116-1
Solo importing, 115-1
Sort, 103-6
Sort transformation, 119-12
Spath
medoid partitioning, 447-4
SPC fundamentals
Xbar R, 250-38
Spearman correlation
linear regression, 300-45
Spearman rank
correlation matrix, 401-3
Spearman rank correlation
linear regression, 300-12
Specificity
1-sample binary diagnostic test, 535-2
2-sample binary diagnostic, 537-2
clustered binary diagnostic, 538-8
paired binary diagnostic, 536-2
ROC curves, 545-1, 545-24
Spectral analysis, 468-1
Spectral density
spectral analysis, 468-3
Spectrum
spectral analysis, 468-1
Sphericity test
factor analysis, 420-14
Splice transformation, 119-12
Spline
scatter plot, 161-15
Split plot analysis
mixed models, 220-1
Split plot data example
mixed models, 220-98
Spread, 140-5
Spreadsheet
limits, 102-1
overview, 102-1
Spreadsheet/database comparison, 102-4
SPSS importing, 115-1
Sqrt transformation, 119-8
Standard deviation, 200-16
confidence limits, 207-2
descriptive statistics, 200-16
ratio, 207-2
unbiased, 200-17
Standard error, 200-13
linear regression, 300-40
Poisson regression, 325-26
Standardization
PC regression, 340-1
ridge regression, 335-3
Standardize transformation, 119-16
Standardized coefficients
linear regression, 300-40
multiple regression, 305-49
Standardized difference, 123-15
Standardized residual
linear regression, 300-19, 300-61, 300-64
multiple regression, 305-18, 305-63
nondetects regression, 345-13
Start time variable
Weibull fitting, 550-12
Starting NCSS, 1-2, 2-1, 100-1, 101-2
Starting values
curve fitting, 350-3
nonlinear regression, 315-1
Stata file exporting, 116-1
Statistical functions transformations, 119-15
Std error
of kurtosis, 200-18
of skewness, 200-18
of standard deviation, 200-16
of variance, 200-15
of X-mean, 200-20
Std Error
of Coefficient of Variation, 200-18
Stddev transformation, 119-16
StdRangeProb transformation, 119-11
StdRangeValue transformation, 119-11
Stem-leaf
depth, 200-27
leaf, 200-28
stem, 200-27
unit, 200-28
Stem-leaf plot
descriptive statistics, 200-27
Stephens test
circular data, 230-7
Stepwise regression, 311-1
Cox regression, 565-11
logistic regression, 320-17
multiple regression, 305-23
Poisson regression, 325-6
Storing results
linear regression, 300-35
multiple regression, 305-42
Stratification based on propensity scores, 124-1
Stratification of a database, 124-1
Stress
multidimensional scaling, 435-3
Stress A
parametric survival regression, 566-6
Stress B
parametric survival regression, 566-6
Stress plot
parametric survival regression, 566-19
Stress variable
parametric survival regression, 566-6
Student's t distribution
probablility calculator, 135-6
Studentized deviance residuals
Poisson regression, 325-5
Studentized Pearson residuals
Poisson regression, 325-5
Studentized range
one-way ANOVA, 210-5
Studentized range distribution
probablility calculator, 135-6
Studentized residuals
Poisson regression, 325-34
Studentized-range distribution
transformation, 119-11
Student's T distribution
simulation, 122-10
Style file
grid plot, 173-8

Style file
box plot, 152-13
histogram, 143-16
probability plot, 144-19
scatter plot, 161-22
Style files
multiple regression, 305-38
Subset of a database, 14-1
Subset selection
Cox regression, 565-11, 565-48
logistic regression, 320-17
multiple regression, 305-23, 305-32
Poisson regression, 325-6, 325-37
Subset selection report
multiple regression, 305-80
Subset selection tutorial
multiple regression, 305-79
Sum of exponentials
curve fitting, 351-9
growth curves, 360-8
Sum of functions models, 380-1
Sum of squares
multiple regression, 305-49, 305-55
Sum transformation, 119-16
Sunflower plot
scatter plot, 161-18
SUNSPOT dataset, 468-9, 472-7
Support services, 100-2
Surface charts, 140-1, 141-1
Surface plot
depth, 171-7
elevation, 171-6
perspective, 171-6
projection method, 171-7
rotation, 171-7
Surface plots, 140-10, 171-1
Survival
cumulative, 565-4
Survival analysis
Kaplan-Meier, 555-1
life-table analysis, 570-1
time calculator, 580-1
Weibull fitting, 550-1
Survival curves
Kaplan-Meier, 555-1
SURVIVAL dataset, 555-14, 555-37, 575-1, 575-5
Survival distribution
Cox regression, 565-2
Survival function
Kaplan-Meier, 555-2
Weibull fitting, 550-2
Survival plot
Kaplan-Meier, 555-35
Survival quantiles
Kaplan-Meier, 555-6, 555-30
SUTTON 22 dataset, 456-6, 456-14
SUTTON30 dataset, 455-6, 455-13
Symbol settings window, 181-1
Symmetric-binary variables
fuzzy clustering, 448-4
hierarchical clustering, 445-6
medoid partitioning, 447-2
Symmetry, 200-2, 206-25
Symphony exporting, 116-1
Syntax
macros, 130-2
SYS exporting, 116-1
Systat exporting, 116-1
Systat importing, 115-1
System requirements, 1-1

T

T distribution
simulation, 122-10
T2 alpha
data screening, 118-3
T2 Dataset, 405-3, 405-5, 405-10
T2 value, 410-7
Tables
descriptive, 201-1
Taguchi designs, 266-1
Tan transformation, 119-17
Tanh transformation, 119-17
Target specification, 250-20
Tarone-Ware test
Kaplan-Meier, 555-12
nondetects analysis, 240-3
Template, 105-1
default, 105-1
new, 105-1
open, 105-1
save, 105-2
saving, 8-5
Terms
multiple regression, 305-35
Text data, 102-1
Text functions transformations, 119-17
Text settings window, 182-1
Theoretical ARMA, 475-1
Thetas
theoretical ARMA, 475-2
Three-variable charts, 140-10
Threshold limit
Xbar R, 250-23
Tick label settings window, 186-1
Tick settings window, 185-1
Tickmarks, 185-1
Ties method
Cox regression, 565-17
Tile horizontally, 106-5
Tile vertically, 106-5
Time calculator, 580-1
Time format, 102-8
Time remaining
life-table analysis, 570-4
Time variable
Cox regression, 565-16
life-table analysis, 570-6
parametric survival regression, 566-4
TIMECALC dataset, 580-3
TNH(Z)
piecewise polynomial models, 365-6
Tolerance
multiple regression, 305-57
PC regression, 340-13
ridge regression, 335-17
Tolerance intervals, 585-1
Toolbar
customizing, 107-3
Topic search
goto window, 106-4
TOST
two-sample, 207-1
Tprob transformation, 119-11
TPT exporting, 116-1
Transformation
recoding, 3-4
Transformation operators, 119-4
Transformations, 3-1, 102-6, 119-1
Abs, 119-7
Arc sine, 119-17
Arc tangent, 119-17
ArCosh, 119-17
Arsine, 119-17
ArSinh, 119-17
ArTan, 119-17
ArTanh, 119-17
Average, 119-15
BetaProb, 119-8
BetaValue, 119-8
BinomProb, 119-8
BinomValue, 119-8
BinormProb transformation, 119-8
Collate, 119-12
conditional, 120-1
Contains, 119-17
CorrProb, 119-8
CorrValue, 119-8
Cos, 119-17
Cosh, 119-17
Cosine, 119-17
Count, 119-15
CsProb, 119-9
CsValue, 119-9
Cum, 119-7
date functions, 119-6
Day, 119-6
Exp, 119-7
ExpoProb, 119-9
ExpoValue, 119-9
Extract, 119-18
file function, 119-15
fill functions, 119-6

Fprob, 119-9
Fraction, 119-7
Fvalue, 119-9
GammaProb, 119-9
GammaValue, 119-9
HypergeoProb, 119-9
if-then, 120-1
indicator variables, 119-19
Int, 119-7
Join, 119-18
Julian date, 119-6
Lagk, 119-16
Lcase, 119-18
Ledk, 119-16
Left, 119-18
Length, 119-18
Ln(X), 119-7
Log, 119-7
LogGamma, 119-9
logic operators, 119-5
Logit, 119-7
Lookup, 119-14
mathematical functions, 119-7
Mavk, 119-16
Max, 119-16
Min, 119-16
Mod, 119-7
Month, 119-6
NcBetaProb, 119-9
NcBetaValue, 119-10
NcCsProb, 119-10
NcCsValue, 119-10
NcFprob, 119-10
NcFvalue, 119-10
NcTprob, 119-10
NcTvalue, 119-10
Negative binomial, 119-10
NegBinomProb, 119-10
Non-central Beta, 119-10
Non-central Chi-square, 119-10
noncentral-F distribution, 119-10
noncentral-t distribution
    transformation, 119-10
NormalProb, 119-10
NormalValue, 119-10
NormScore, 119-16
numeric functions, 119-6
PoissonProb, 119-11
probability functions, 119-8
RandomNormal, 119-11
random-number functions, 119-11
Rank, 119-16
rearrangement functions, 119-12
Recode, 119-15
recode functions, 119-14
recoding, 11-1
Remove, 119-18
Repeat, 119-18
Replace, 119-18
Right, 119-19
Round, 119-7
Sequence, 119-6
Ser, 119-6
Short, 119-7
Sign, 119-8
simulation, 15-1
Sin, 119-17
Sinh, 119-17
Smooth, 119-16
Sort, 119-12
Splice, 119-12
Sqrt, 119-8
Standardize, 119-16
statistical functions, 119-15
Stddev, 119-16
StdRangeProb, 119-11
StdRangeValue, 119-11
Studentized-range distribution, 119-11
Sum, 119-16
Tan, 119-17
Tanh, 119-17
text functions, 119-17
Tprob, 119-11
trigonometric functions, 119-17
Tvalue, 119-11
Ucase, 119-19
UnCollate, 119-13
Uniform, 119-11
Uniques, 119-13
UnSplice, 119-14
WeibullProb, 119-11
WeibullValue, 119-11
Year, 119-6
Transition type
    piecewise polynomial models, 365-6
Tricube weights
    linear regression, 300-13
Trigamma
    beta distribution fitting, 551-14
Trigonometric functions
    transformations, 119-17
Trim-mean
    descriptive statistics, 200-19
Trimmed
    descriptive statistics, 200-19
Trim-std dev
    descriptive statistics, 200-19
Tschuprow's T
    cross tabulation, 501-14
T-test
    1-Sample T2, 405-1
    assumptions, 205-22
    average difference plot, 205-20
    bootstrapping, 205-3
    histogram, 205-20
    kurtosis, 205-15
    multiplicity factor, 205-19
    nonparametric tests, 205-17
    normality, 205-15
    outliers, 205-22
    probability plot, 205-20
    skewness, 205-15
T-test of difference
    two proportions, 515-8
T-tests
    meta-analysis of means, 455-1
    one sample, 205-1
    paired, 205-1
    two-sample, 206-1
    two-sample (means/SDs), 207-1
Tukey's biweight
    multiple regression, 305-27
Tukey-Kramer test
    one-way ANOVA, 210-8
Tukey's lambda distribution
    simulation, 122-10
TUTOR dataset, 220-98
Tutorial
    general, 101-1
    linear regression, 300-37
Tvalue transformation, 119-11
Two correlated proportions, 520-1
Two independent proportions, 515-1
Two proportions, 515-1
Two sample t-test (from means/SDs), 207-1
Two-level designs, 260-1
Two-level factorial designs, 260-1
TWOSAMPLE dataset, 220-69, 220-72
Two-sample t-test, 5-1, 206-1
    assumptions, 206-18, 206-27
    bootstrapping, 206-3
    degrees of freedom, 206-13
TWOSAMPLE2 dataset, 220-70, 220-73
TWOSAMPLECOV dataset, 220-76
Two-variable charts, 140-4, 140-7
Two-way tables
    cross tabulation, 501-1
TXT exporting, 116-1
TXT importing, 115-1

U

Ucase transformation, 119-19
U-chart, 251-2
Unbiased std dev
    descriptive statistics, 200-17
UnCollate transformation, 119-13
Unconditional tests
    two proportions, 515-5
Undo, 103-4
Unequal variance t-test, 206-2
Uniform distribution
    simulation, 122-11
Uniform kernel
    Kaplan-Meier, 555-8

Weibull fitting, 550-34
Uniform transformation, 119-11
Uniformity test
circular data, 230-3
Uniques transformation, 119-13
Unknown censor
Cox regression, 565-18
Kaplan-Meier, 555-17
life-table analysis, 570-6
UnSplice transformation, 119-14
Unweighted means F-tests, 211-1
User written models, 385-1
UWM F-tests, 211-1
properties of, 211-1

V

Validation
Cox regression, 565-55
life-table analysis, 570-24
Validity
item analysis, 505-1
Value labels, 13-1, 102-10
Variable
data type, 102-10
format, 102-6
labels, 102-6
names, 101-1, 102-5
numbers, 102-5
transformations, 102-6
Variable format, 102-6
Variable info, 102-5
tutorial, 101-2
Variable info file, 102-1
Variable info sheet, 102-1
Variable info tab, 2-4
Variable labeling, 2-4
Variable labels, 102-6
Variable matching, 123-3
Variable name, 2-4
Variable names, 102-5
rules for, 2-5
Variable numbers, 102-5
Variable selection, 310-1
Cox regression, 565-11
logistic regression, 320-17
multiple regression, 305-23
Poisson regression, 325-6
principal components analysis, 425-8
Variables
naming, 101-2
Variables charts, 250-1
Variance
descriptive statistics, 200-15
linear regression, 300-5
multiple regression, 305-13
Variance components
R & R, 254-3, 254-11
Variance inflation factor
multiple regression, 305-8, 305-57
PC regression, 340-12
ridge regression, 335-16
Variance inflation factor plot
ridge regression, 335-19
Variance inflation factors
ridge regression, 335-2
Variance ratio test, 206-19
Variance test
equal, 206-19
linear regression, 300-50
Variances
equality of, 206-20
testing equality of multiple, 210-18
Variates
canonical correlation, 400-1
Varimax rotation
factor analysis, 420-4
principal components analysis, 425-7
VIF
multiple regression, 305-8
ridge regression, 335-2
Violin plot
density trace, 154-1
Violin plots, 140-6, 154-1
Von Mises distribution
circular data, 230-5

W

W mean
appraisal ratios, 485-8
Wald method
correlated proportions, 520-4
Wald statistic
Poisson regression, 325-26
Wald test
Cox regression, 565-11, 565-33
logistic regression, 320-9
Walter's confidence intervals
two proportions, 515-22
Ward's minimum variance
double dendrograms, 450-3
hierarchical clustering, 445-4
Watson & Williams test
circular data, 230-7
Watson test
circular data, 230-4
Watson-Williams F test
circular data, 230-10
WEIBULL dataset, 550-12, 550-27, 550-44, 552-3, 552-12, 555-27
Weibull distribution
probablility calculator, 135-6
simulation, 122-12
Weibull fitting, 550-6
Weibull fitting, 550-1
Weibull model
curve fitting, 351-7
growth curves, 360-6
Weibull probability plot, 144-17
Weibull regression, 566-1
WEIBULL2 dataset, 144-17
WeibullProb transformation, 119-11
WeibullValue transformation, 119-11
Weight variable
linear regression, 300-25
multiple regression, 305-28
WEIGHTLOSS dataset, 220-85
WESTGARD dataset, 252-9
Westgard rules, 252-1
Westlake's confidence interval, 235-6
Whiskers
box plot, 152-5
Wilcoxon rank-sum test, 206-1, 206-20
Wilcoxon signed-rank test, 205-18
Wilcoxon-Mann-Whitney test
cross-over analysis using t-tests, 235-8
Wilks' lambda
canonical correlation, 400-10
discriminant analysis, 440-2
MANOVA, 415-2
Wilson score limits
one proportion, 510-2
Wilson's score
correlated proportions, 520-3
two proportions, 515-19
Window
data, 7-1
output, 9-1
Windows
navigating, 1-4
Winters forecasting
exponential smoothing, 467-1
Within factor
repeated measures, 214-9
Within subject
repeated measures, 214-2
WK exporting, 116-1
WKQ exporting, 116-1
Woolf's odds ratio analysis
Mantel-Haenszel test, 525-11
Word processor, 9-1
Working-Hotelling C.I. band
linear regression, 300-6
Working-Hotelling limits
linear regression, 300-60
WR1 exporting, 116-1
WRK exporting, 116-1

X

Xbar chart, 250-1
Xbar R chart, 250-1
XLS exporting, 116-1

Y

Year format, 102-8
Year transformation, 119-6
Yule-Walker
 automatic ARMA, 474-1

Z

Zero time replacement
 beta distribution fitting, 551-3
 cumulative incidence, 560-4
 gamma distribution fitting, 552-4
 parametric survival regression, 566-4
 Weibull fitting, 550-13
ZHOU 175 dataset, 545-33
ZINC dataset, 345-15