Chapter 300

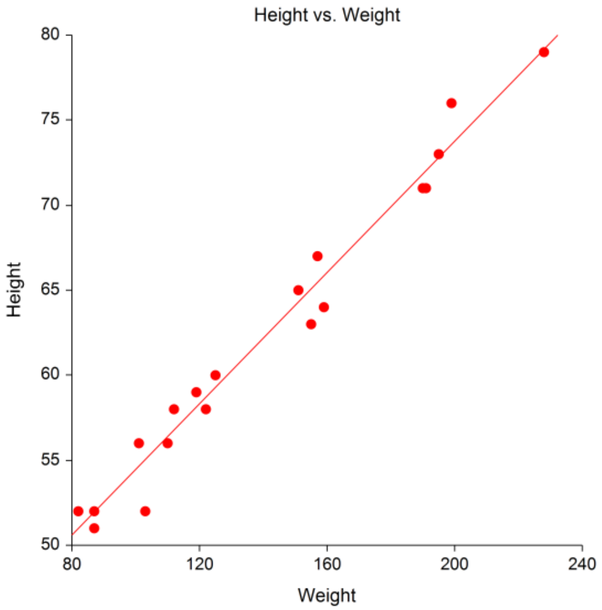
Linear Regression and Correlation

Introduction

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

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

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



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

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

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

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

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

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

Correlation

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

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

Possible Uses of Linear Regression Analysis

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

Description

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

Coefficient Estimation

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

Prediction

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

Control

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

Assumptions

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

Linearity

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

Constant Variance

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

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

Special Causes

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

Normality

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

Independence

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

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

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

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

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

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

Technical Details

Regression Analysis

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

Define the following vectors and matrices.

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

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

Least Squares

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

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

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

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

The predicted values of the dependent variable are given by

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

The residuals are calculated using

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

Estimated Variances

An estimate of the variance of the residuals is computed using

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

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

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

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

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

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

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

Hypothesis Tests of the Intercept and Slope

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

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

and

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

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

Confidence Intervals of the Intercept and Slope

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

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

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

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

Confidence Interval of Y for Given X

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

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

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

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

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

Working-Hotelling Confidence Band for the Mean of Y

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

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

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

Confidence Interval of X for Given Y

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

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

Then, calculate the interval using

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

where

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

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

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

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

R-Squared (Percent of Variation Explained)

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

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

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

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

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

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

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

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

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

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

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

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

Rbar-Squared (Adjusted R-Squared)

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

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

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

Probability Ellipse

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

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

Orthogonal Regression Line

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

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

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

where

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

The estimate of the intercept is then computed using

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

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

The Correlation Coefficient

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

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

The corresponding formulas for the calculation of the correlation coefficient are

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

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

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

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

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

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

Facts about the Correlation Coefficient

The correlation coefficient has the following characteristics.

1. The range of *r* is between -1 and 1, inclusive.
2. If *r* = 1, the observations fall on a straight line with positive slope.
3. If *r* = -1, the observations fall on a straight line with negative slope.
4. If *r* = 0, there is no linear relationship between the two variables.
5. *r* is a measure of the linear (straight-line) association between two variables.

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

Hypothesis Tests for the Correlation

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

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

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

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

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

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

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

is closely approximated by a normal distribution with mean

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

and variance

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

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

$$\begin{aligned} z &= \frac{F(r) - F(\rho_0)}{\sqrt{\frac{1}{N-3}}} \\ &= \frac{\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{2\sqrt{\frac{1}{N-3}}} \end{aligned}$$

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

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

Confidence Intervals for the Correlation

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

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

Spearman Rank Correlation Coefficient

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

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

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

Smoothing with Loess

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

The loess algorithm works as follows.

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

Mathematical Details of Loess

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

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

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

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

The tricube weight function is defined as

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

The weight for each observation is defined as

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

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

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

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

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

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

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

Robust Loess

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

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

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

New weights are defined as

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

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

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

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

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

Testing Assumptions Using Residual Diagnostics

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

Notation – Use of (j) and p

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

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

1 – No Outliers

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

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

2 – Linear Regression Function - No Curvature

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

Loess Curve

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

Lack of Fit Test

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

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

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

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

3 – Constant Variance

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

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

Modified Levene Test

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

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

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

where

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

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

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

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

4 – Independent Errors

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

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

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

Durbin-Watson Test

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

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

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

5 – Normality of Residuals

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

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

Influential Observations

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

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

Definitions Used in Residual Diagnostics

Residual

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

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

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

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

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

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

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

Hence, the variance of e_j is given by

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

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

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

Hat Diagonal

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

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

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

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

Standardized Residual

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

The formula for this residual is

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

s(j) or MSEi

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

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

RStudent

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

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

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

DFFITS

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

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

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

Cook's D

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

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

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

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

This approximation is exact when no weight variable is used.

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

CovRatio

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

The general formula for the CovRatio is

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

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

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

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

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

DFBETAS

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

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

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

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

Press Value

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

The formula for PRESS is

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

Press R-Squared

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

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

Sum |Press residuals|

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

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

Bootstrapping

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

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

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

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

When applied to linear regression, there are two types of bootstrapping that can be used. See Neter, Kutner, Nachtsheim, Wasserman (1996) page 430.

Modified Residuals

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

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

where

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

Bootstrap the Observations

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

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

Bootstrap the Residuals

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

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

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

Bootstrap Prediction Intervals

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

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

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

Randomization Test

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

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

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

Data Structure

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

LinReg1 dataset (subset)

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

Missing Values

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

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y: Dependent Variable(s)

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

X: Independent Variable

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

Frequency Variable

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

Weight Variable

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

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

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

Model Specification

Remove Intercept

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

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

Resampling

Calculate Bootstrap C.I.'s

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

Run randomization tests

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

Alpha Levels

Alpha for C.I.'s and Tests

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

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

Alpha for Assumptions

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

We recommend 0.20.

Reports Tab

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

Select Report / Plot Group

Select a Group of Reports and Plots

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

Report Options

Show Notes

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

Show All Rows

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

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

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

Select Reports – Summaries

Run Summary ... Summary Matrices

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

Select Reports – Estimation

Regression Estimation

Indicate whether to display this report.

Select Reports – ANOVA

ANOVA

Indicate whether to display this report.

Select Reports – Assumptions

Assumptions

Indicate whether to display this report.

Levene Groups

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

Durbin-Watson

Indicate whether to display this report.

PRESS

Indicate whether to display this report.

Select Reports – Prediction

Predict Y at these X Values

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

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

Examples:

10

10 20 30 40 50

0:100(10)

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

Predicted Y – C.L.

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

Predicted Y – P.L.

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

Select Reports – Row-by-Row Lists

Original Data ... Predicted X Individuals

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

Select Reports – Regression Diagnostics

Residuals ... Outlier-Influence Chart

Indicate whether to display these reports.

Format Tab

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

Report Options

Precision

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

Variable Names

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

Report Options – Decimal Places

Probability ... Matrix Decimals

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

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

Plots Tab

These options specify which plots are produced as well as the plot format.

Select Plots

Y vs X Plot ... Probability Plot

Indicate whether to display these plots. Click the plot format button to change the plot settings.

Plot Options

Y vs X Plot Size and All Other Plot Sizes

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

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

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

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

Resampling Tab

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

Bootstrap Calculation Options

The following options control the calculation of bootstrap confidence intervals.

Bootstrap Calculation Options – Sampling

Samples (N)

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

We recommend setting this value to at least 3000.

Sampling Method

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

- **Observations**

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

- **Residuals**

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

  We recommend setting this value to at least 3000.

Retries

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

Bootstrap Calculation Options – Estimation

Percentile Type

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

C.I. Method

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

- **Percentile**

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

- **Reflection**

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

Bootstrap Confidence Coefficients

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

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

Examples:

0.90 0.95 0.99

0.90:.99(0.01)

0.90.

Randomization Test Options

Monte Carlo Samples

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

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

Storage Tab

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

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

Data Storage Options

Storage Option

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

- **Do not store data**

  No data are stored even if they are checked.

- **Store in empty columns only**

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

- **Store in designated columns**

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

Store First Variable In

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

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

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

Data Storage Options – Select Items to Store

Predicted Y ... LOESS Values

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

Example 1 – Running a Linear Regression Analysis

This section presents an example of how to run a linear regression analysis of the data in the LinReg1 dataset. In this example, we will run a regression of *Height* on *Weight*. Predicted values of Height are wanted at Weight values equal to 90, 100, 150, 200, and 250.

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

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Linear Regression and Correlation window.

1 **Open the LinReg1 dataset.**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **LinReg1.NCSS**.
- Click **Open**.

2 **Open the Linear Regression and Correlation window.**

- Using the Analysis menu or the Procedure Navigator, find and select the **Linear Regression and Correlation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 **Specify the variables.**

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

4 **Specify the randomization test options.**

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

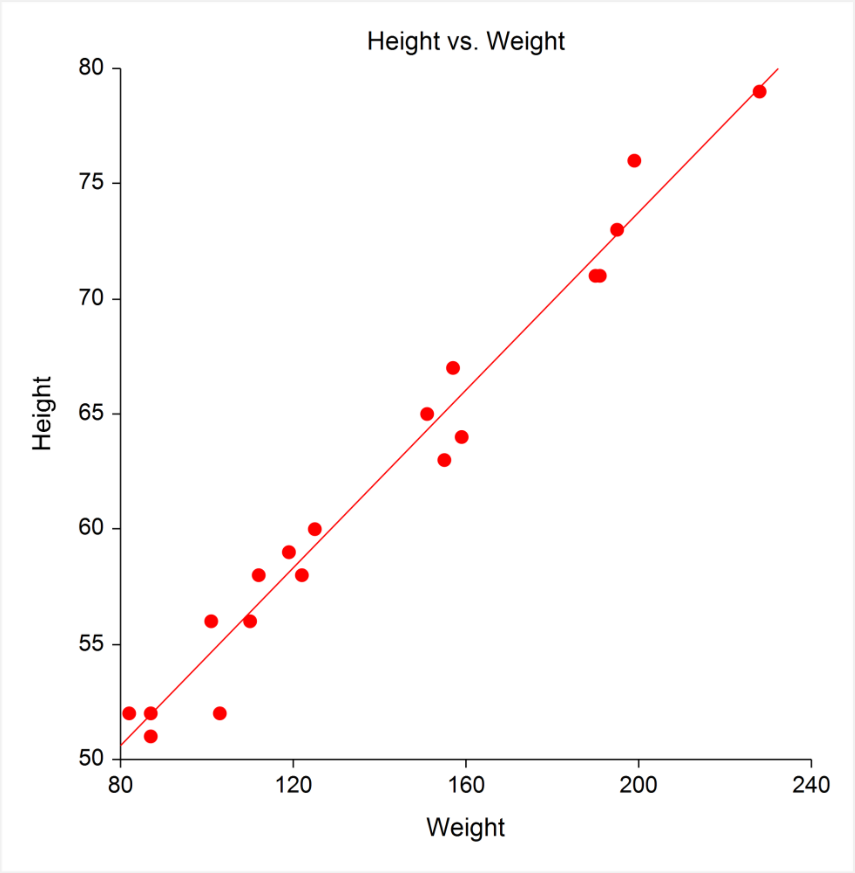
5 **Specify the reports.**

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

6 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Linear Regression Plot Section



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

Run Summary Section

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

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

Coefficient of Variation

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

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

Summary Statement

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

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

Descriptive Statistics Section

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

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

Regression Estimation Section

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

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

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

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

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

Regression Coefficients

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

Lower - Upper 95% Confidence Limits

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

The formulas for the lower and upper confidence limits are

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

Standard Error

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

Standardized Coefficient

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

The formula for the standardized regression coefficient is:

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

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

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

T-Value

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

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

Prob Level

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

Prob Level (Randomization Test)

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

Reject H0 (Alpha = 0.05)

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

Power (Alpha = 0.05)

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

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

Regression of Y on X

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

Regression of X on Y

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

Orthogonal Regression of Y and X

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

Estimated Model

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

Bootstrap Section

```
---  Estimation Results   ------  | --- Bootstrap Confidence Limits----
Parameter              Estimate  | Conf. Level Lower           Upper
Intercept
Original Value          35.1337  | 0.9000    33.5138         36.8691
Bootstrap Mean          35.1391  | 0.9500    33.1520         37.2812
Bias (BM - OV)           0.0055  | 0.9900    32.6492         38.1285
Bias Corrected          35.1282
Standard Error           1.0178
Slope
Original Value           0.1932  | 0.9000     0.1815          0.2047
Bootstrap Mean           0.1931  | 0.9500     0.1785          0.2069
Bias (BM - OV)           0.0000  | 0.9900     0.1729          0.2118
Bias Corrected           0.1932
Standard Error           0.0071
Correlation
Original Value           0.9868  | 0.9000     0.9799          0.9973
Bootstrap Mean           0.9865  | 0.9500     0.9789          1.0000
Bias (BM - OV)          -0.0003  | 0.9900     0.9772          1.0000
Bias Corrected           0.9871
Standard Error           0.0056
R-Squared
Original Value           0.9738  | 0.9000     0.9601          0.9943
Bootstrap Mean           0.9733  | 0.9500     0.9582          0.9996
Bias (BM - OV)          -0.0005  | 0.9900     0.9548          1.0000
Bias Corrected           0.9743
Standard Error           0.0109
Standard Error of Estimate
Original Value           1.4036  | 0.9000     1.1710          1.8446
Bootstrap Mean           1.3241  | 0.9500     1.1225          1.9071
Bias (BM - OV)          -0.0795  | 0.9900     1.0355          2.0552
Bias Corrected           1.4832
Standard Error           0.2046
Orthogonal Intercept
Original Value          35.1076  | 0.9000    33.4855         36.8576
Bootstrap Mean          35.1123  | 0.9500    33.1251         37.2581
Bias (BM - OV)           0.0047  | 0.9900    32.6179         38.1223
Bias Corrected          35.1028
Standard Error           1.0231
```

```
Orthogonal Slope
Original Value          0.1934   | 0.9000      0.1816          0.2048
Bootstrap Mean          0.1933   | 0.9500      0.1786          0.2071
Bias (BM - OV)          0.0000   | 0.9900      0.1731          0.2120
Bias Corrected          0.1934
Standard Error          0.0071
Predicted Mean and Confidence Limits of Height when Weight = 90.0000
Original Value         52.5188   | 0.9000     51.8172         53.2993
Bootstrap Mean         52.5220   | 0.9500     51.6895         53.4913
Bias (BM - OV)          0.0032   | 0.9900     51.4648         53.8741
Bias Corrected         52.5157
Standard Error          0.4549

(Report continues for the other values of Weight)

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.
```

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

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

Original Value

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

Bootstrap Mean

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

Bias (BM - OV)

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

Bias Corrected

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

Standard Error

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

Conf. Level

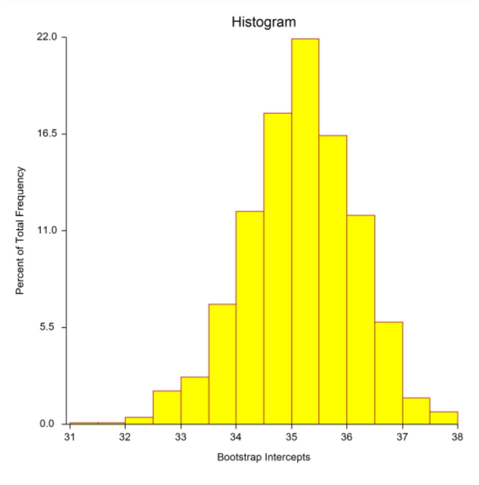
This is the confidence coefficient of the bootstrap confidence interval given to the right.

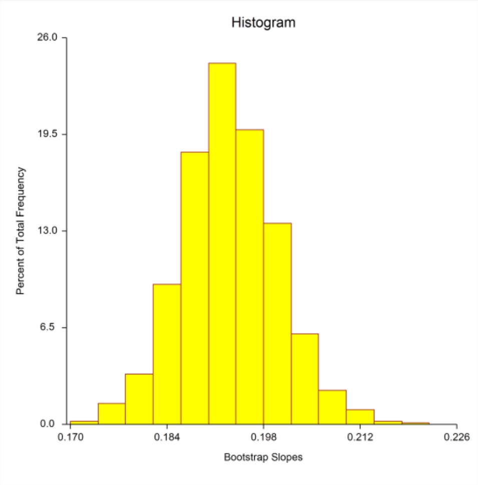
Bootstrap Confidence Limits - Lower and Upper

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

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

Bootstrap Histograms Section





(five more histograms are shown)

Each histogram shows the distribution of the corresponding parameter estimate.

Correlation and R-Squared Section

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

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

Pearson Correlation Coefficient

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

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

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

Prob Level (Randomization Test)

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

Spearman Rank Correlation Coefficient

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

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

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

R-Squared

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

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

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

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

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

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

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

Adjusted R-Squared

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

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

Analysis of Variance Section

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

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

Source

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

DF

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

Sum of Squares

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

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

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

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

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

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

Mean Square

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

F-Ratio

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

Prob Level

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

Power(5%)

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

S = Root Mean Square Error

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

Summary Matrices

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

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

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

Tests of Assumptions Section

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

Constant Residual Variance?
Modified Levene Test                0.0946   0.761964   Yes

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

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

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

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

Residuals follow Normal Distribution?

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

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

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

Shapiro-Wilk W Test

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

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

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

Anderson-Darling Test

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

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

D’Agostino Skewness

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

D’Agostino Kurtosis

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

D’Agostino Omnibus

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

Constant Residual Variance?

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

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

Modified Levene Test

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

Relationship is a Straight Line?

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

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

Lack of Linear Fit Test

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

Serial Correlation and Durbin-Watson Sections

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

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

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

Lag

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

Serial Correlation

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

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

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

Durbin-Watson Value

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

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

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

PRESS Section

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

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

Sum of Squared Residuals

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

Sum of |Press residuals|

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

Press R-Squared

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

Predicted Values and Confidence Limits Section

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

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

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

X

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

Predicted Y (Yhat|X)

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

Standard Error of Yhat

This is the estimated standard deviation of the predicted value.

Lower 95% Confidence Limit of Y|X

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

Upper 95% Confidence Limit of Y|X

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

Predicted Values and Prediction Limits Section

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

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

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

X

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

Predicted Y (Yhat|X)

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

Standard Error of Yhat

This is the estimated standard deviation of the predicted value.

Lower 95% Prediction Limit of Y|X

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

Upper 95% Prediction Limit of Y|X

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

Residual Plots

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

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

Point Cloud

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

Wedge

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

Bowtie

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

Sloping Band

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

Curved Band

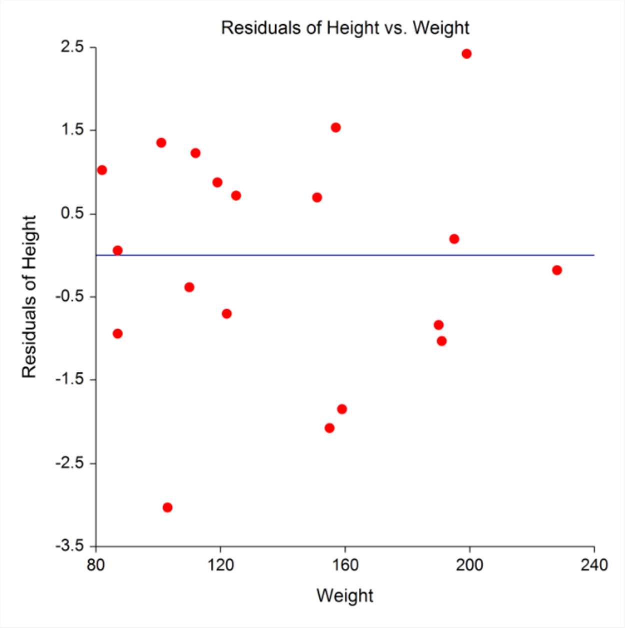
This kind of residual plot may be indicative of a nonlinear relationship between *Y* and the independent variable that was not accounted for. The solution might be to use a transformation on *Y* to create a linear relationship with *X*. Another possibility might be to add quadratic or cubic terms of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

Curved Band with Increasing or Decreasing Variability

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

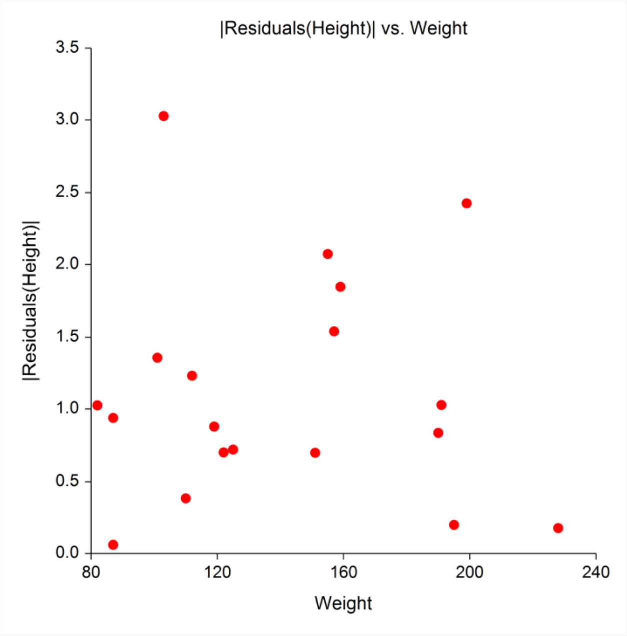
Residuals Plots

Residuals vs X Plot



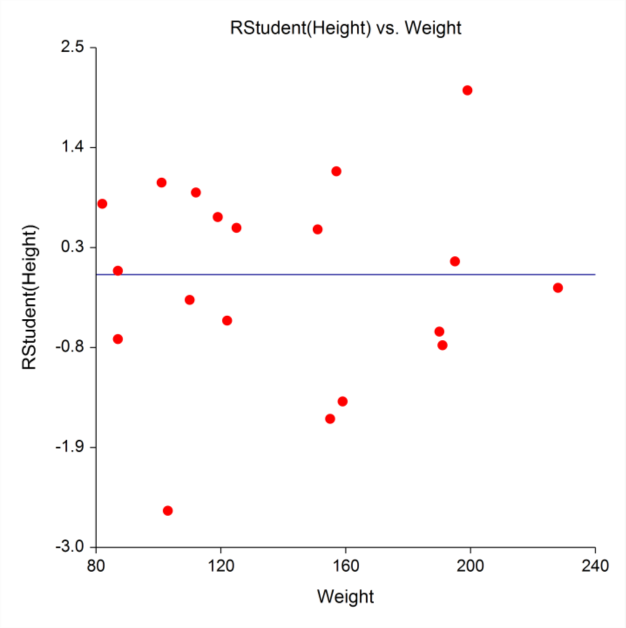
This plot is useful for showing nonlinear patterns and outliers. The preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.

|Residual| vs X Plot



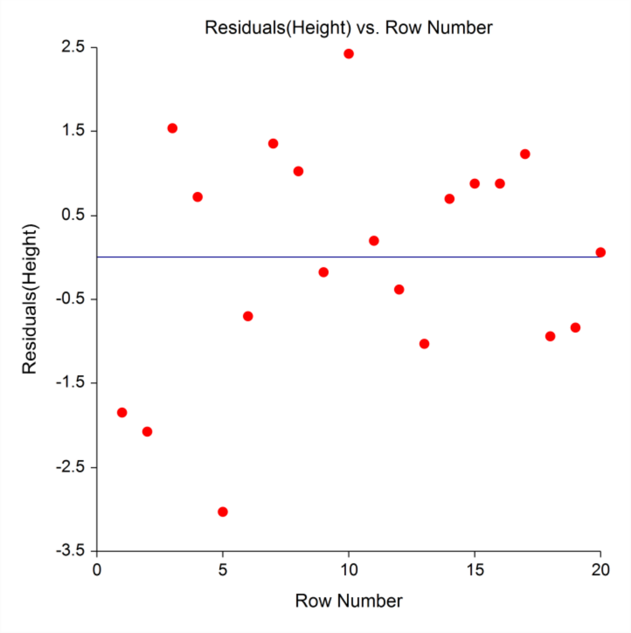
This plot is useful for showing nonconstant variance in the residuals. The preferred pattern is a rectangular shape or point cloud. The most common type of nonconstant variance occurs when the variance is proportion to *X*. This is shown by a funnel shape. Remedies for nonconstant variances were discussed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

Rstudent vs X Plot



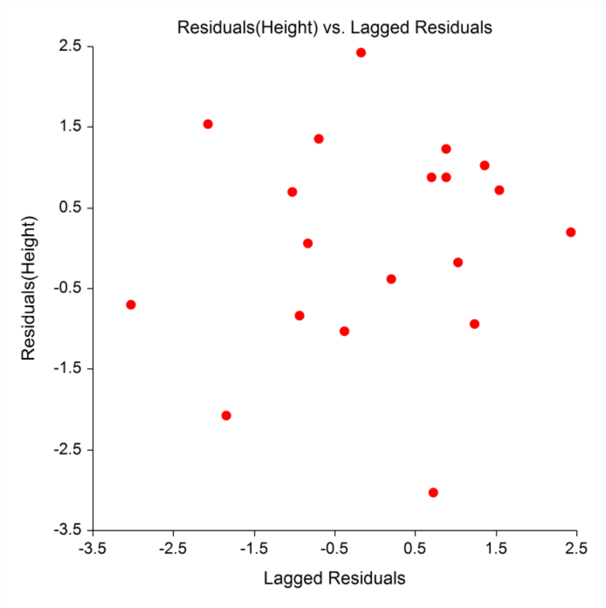
This is a scatter plot of the RStudent residuals versus the independent variable. The preferred pattern is a rectangular shape or point cloud. This plot is helpful in identifying any outliers.

Sequence Plot: Residuals vs Row



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

Serial Correlation Plot: Residuals vs Lagged Residuals



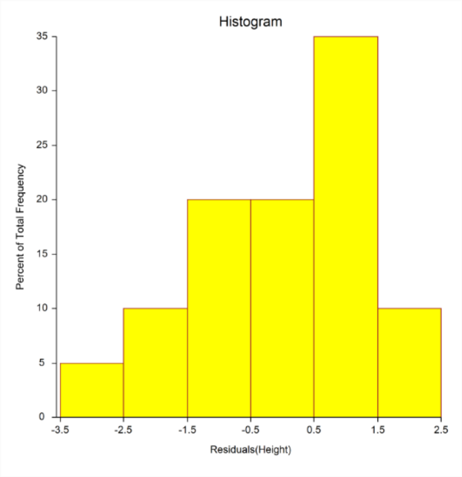
This is a scatter plot of the *i*th residual versus the *i*th-1 residual. It is only useful for time series data where the order of the rows on the database is important.

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

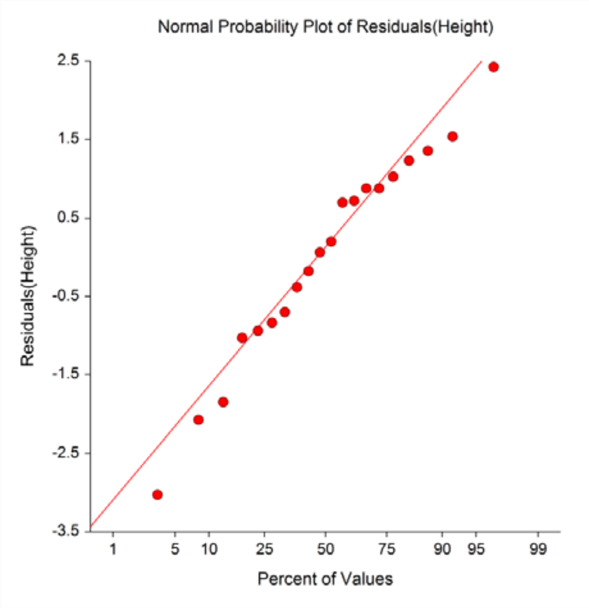
Positive autocorrelation or serial correlation means that the residual in time period *t* tends to have the same sign as the residual in time period (*t* - 1). On the other hand, a strong negative autocorrelation means that the residual in time period *t* tends to have the opposite sign as the residual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

Be sure to check the Durbin-Watson statistic.

Distributional Plots of Residuals



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



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

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

Original Data Section

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

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

Predicted Values of Means Section

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

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

X

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

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

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

Standard Error of Yhat

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

Lower 95% Confidence Limit of Y|X

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

Upper 95% Confidence Limit of Y|X

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

Predicted Values and Prediction Limits Section

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

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

X

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

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

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

Standard Error of Yhat

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

Lower 95% Prediction Limit of Y|X

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

Upper 95% Prediction Limit of Y|X

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

Working-Hotelling Simultaneous Confidence Band

```
                                 Predicted    Standard    Lower 95%   Upper 95%
            Weight     Height       Height       Error   Conf. Band  Conf. Band
Row            (X)        (Y)     (Yhat|X)     of Yhat  of Y Mean|X of Y Mean|X
1         159.0000    64.0000      65.8475      0.3457      64.8036     66.8914
2         155.0000    63.0000      65.0748      0.3343      64.0654     66.0842
3         157.0000    67.0000      65.4611      0.3397      64.4353     66.4869
4         125.0000    60.0000      59.2797      0.3323      58.2764     60.2831
5         103.0000    52.0000      55.0300      0.4162      53.7732     56.2868
6         122.0000    58.0000      58.7002      0.3403      57.6727     59.7278
.                .          .            .           .            .           .
.                .          .            .           .            .           .
.                .          .            .           .            .           .
```

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

X

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

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

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

Standard Error of Yhat

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

Lower 95% Confidence Band of Y|X

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

Upper 95% Confidence Band of Y|X

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

Residual Section

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

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

X

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

Y

This is the actual value of *Y*.

Predicted Y (Yhat|X)

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

Residual

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

Standardized Residual

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

The formula for this residual is

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

Percent Absolute Error

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

Residual Diagnostics Section

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

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

X

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

Residual

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

RStudent

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

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

Hat Diagonal

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

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

Cook's D

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

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

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

MSEi

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

Leave One Row Out Section

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

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

RStudent

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

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

Dffits

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

Cook's D

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

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

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

CovRatio

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

DFBETAS(0) and DFBETAS(1)

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

Outlier Detection Chart

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

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

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

X

This is the value of *X*.

Residual

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

Standardized Residual

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

RStudent

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

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

Influence Detection Chart

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

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

X

This is the value of *X*.

Dffits

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

Cook's D

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

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

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

DFBETAS(1)

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

Outlier & Influence Detection Chart

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

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

X

This is the value of *X*.

RStudent (Outlier)

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

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

Cook's D (Influence)

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

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

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

Hat Diagonal (Leverage)

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

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

Inverse Prediction of X Means

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

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

Y

This is the actual value of *Y*.

X

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

Predicted X (Xhat|Y)

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

Lower 95% Confidence Limit of X Mean|Y

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

Upper 95% Confidence Limit of X Mean|Y

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

Inverse Prediction of X Individuals

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

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

Y

This is the actual value of *Y*.

X

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

Predicted X (Xhat|Y)

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

Lower 95% Prediction Limit of X|Y

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

Upper 95% Prediction Limit of X|Y

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