Chapter 305

# Multiple Regression

## Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

# Regression Models

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

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

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

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

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

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

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

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

## Representing a Curvilinear Relationship

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

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

might be expanded to

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

$$= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5$$

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

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

$$X_1 \quad X_2 \quad X_1 * X_1 \quad X_1 * X_2 \quad X_2 * X_2$$

## Representing Categorical Variables

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

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

There are many ways in which these new variables may be generated. You can use the Contrasts data tool in **NCSS** (Data Window Menu: Data > Create Contrast Variables) to automatically create many types of contrasts and binary indicator variables. We will present a few examples here.

## Indicator Variables

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

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

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

## Contrast Variables

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

Continuing with our example, one set of contrast variables is

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

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

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

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

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

## Representing Interactions of Numeric Variables

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

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

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

## Representing Interactions of Numeric and Categorical Variables

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

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

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

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

## Representing Interactions Two or More Categorical Variables

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

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

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

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

# Possible Uses of Regression Analysis

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

## Description

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

## Coefficient Estimation

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

## Prediction

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

## Control

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

## Variable Selection or Screening

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

# Assumptions

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

## Linearity

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

## Constant Variance

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

## Special Causes

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

## Normality

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

## Independence

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

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

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

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

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

## Multicollinearity

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

### Effects of Multicollinearity

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

## Sources of Multicollinearity

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

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

## Detection of Collinearity

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

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

## Correction of Collinearity

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

### Centering and Scaling Issues in Collinearity

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

# Multiple Regression Checklist

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

## Step 1 – Data Preparation

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

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

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

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

## Step 2 – Variable Selection

Variable selection seeks to reduce the number of IV's to a manageable few. There are several variable selection methods in regression: Subset Selection, Stepwise Regression, All Possible Regressions, or Multivariate Variable Selection. Each of these variable selection methods has advantages and disadvantages. We suggest that you begin with the Subset Select procedure since it allows you to look at interactions, powers, and categorical variables.

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

# Step 3 – Setup and Run the Regression

## Introduction

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

## Enter Variables

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

## Choose Report Options

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

## Specify Alpha

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

## Select All Plots

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

# Step 4 – Check Model Adequacy

## Introduction

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

## Check 1. Linearity

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

## Check 2. Normality

- Look at the *Normal Probability Plot*. If all of the residuals fall within the confidence bands for the *Normal Probability Plot*, the normality assumption is likely met. One or two residuals outside the confidence bands may be an indicator of outliers, not non-normality.
- Look at the *Normal Assumptions Section*. The formal normal goodness of fit tests are given in the *Normal Assumptions Section*. If the decision is accepted for the *Normality (Omnibus)* test, there is no evidence that the residuals are not normal.
- If normality does not exist, take the appropriate action and return to Step 2. Appropriate action includes removing outliers and/or using the logarithm of the dependent variable.

## Check 3. Non-constant Variance

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

## Check 4. Independence or Serial Correlation

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

## Check 5. Outliers

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

## Check 6. Multicollinearity

- Look at the Multicollinearity Section. If any variable has a variance inflation factor greater than 10, collinearity could be a problem.
- Look at the Eigenvalues of Centered Correlations Section. Condition numbers greater than 1000 indicate severe collinearity. Condition numbers between 100 and 1000 imply moderate to strong collinearity.
- Look at the Correlation Matrix Section. Strong pairwise correlation here may give some insight as to the variables causing the collinearity.
- If multicollinearity does exist in the model, it could be due to an outlier (return to Check 5 and then Step 2) or due to strong interdependencies between independent variables. In the latter case, return to Step 2 and try a different variable selection procedure.

## Check 7. Predictability

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

## Step 5 – Record Your Results

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

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

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

# Multiple Regression Technical Details

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

Define the following vectors and matrices:

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

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

## Least Squares

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

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

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

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

The predicted values of the dependent variable are given by

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

The residuals are calculated using

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

## Estimated Variances

An estimate of the variance of the residuals is computed using

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

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

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

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

$$s^2_{\hat{Y}_m|X_0} = s^2(1, X_0)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\begin{pmatrix} 1 \\ X_0 \end{pmatrix}$$

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

$$s^2_{\hat{Y}_I|X_0} = s^2 + s^2_{\hat{Y}_m|X_0}$$

## Hypothesis Tests of the Intercept and Slopes

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

$$t_{b_i} = \frac{b_i - B_i}{s_{b_i}}$$

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

## Confidence Intervals of the Intercept and Slope

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

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

## Confidence Interval of Y for Given X

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

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

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

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

## $R^2$ (Percent of Variation Explained)

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

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

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

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

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

1. *Additional independent variables*. It is possible to increase $R^2$ by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This usually happens when the sample size is small.
2. *Range of the independent variables*. $R^2$ is influenced by the range of the independent variables. $R^2$ increases as the range of the *X*'s increases and decreases as the range of the *X*'s decreases.
3. *Slope magnitudes*. $R^2$ does not measure the magnitude of the slopes.
4. *Linearity*. $R^2$ does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *X* and *Y* was a perfect sphere. Although there is a perfect relationship between the variables, the $R^2$ value would be zero.
5. *Predictability*. A large $R^2$ does not necessarily mean high predictability, nor does a low $R^2$ necessarily mean poor predictability.

6. *No-intercept model*. The definition of $R^2$ assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of $R^2$ changes dramatically. The fact that your $R^2$ value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying definition of $R^2$.
7. *Sample size*. $R^2$ is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

## Rbar$^2$ (Adjusted R$^2$)

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

$$\bar{R}_2 = 1 - \frac{(N-1)(1-R^2)}{N-p-1}$$

# Testing Assumptions Using Residual Diagnostics

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

## Notation – Use of (j) and p

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

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

## 1 – No Outliers

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

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

## 2 – Linear Regression Function - No Curvature

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

## 3 – Constant Variance

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

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

## 4 – Independent Errors

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

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

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

### Durbin-Watson Test

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

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

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

## 5 – Normality of Residuals

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

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

## Influential Observations

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

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

## Definitions Used in Residual Diagnostics

### Residual

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

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

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

By assumption, the variance of the $\varepsilon_j$ is $\sigma^2$. However, the variance of the $e_j$ is not $\sigma^2$. In vector notation, the covariance matrix of **e** is given by

$$\mathbf{V(e)} = \sigma^2 \left( \mathbf{I} - \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}} \right)$$

$$= \sigma^2(\mathbf{I} - \mathbf{H})$$

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

$$\hat{Y} = \mathbf{Xb}$$

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

$$= \mathbf{HY}$$

Hence, the variance of $e_j$ is given by

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

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

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

## Hat Diagonal

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

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

**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 ensure that they are correct. You should not delete an observation just because it has a high-influence. However, when you interpret the regression equation, you should bear in mind that the results may be due to a few, high-leverage observations.

## Standardized Residual

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

The formula for this residual is

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

## s(j) or MSEi

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

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

$$= \frac{(N-p)s^2 - \frac{w_j e_j^2}{1-h_{jj}}}{N-p-1}$$

## RStudent

Rstudent is similar to the studentized residual. The difference is the *s*(*j*) is used rather than *s* in the denominator. The quantity *s*(*j*) is calculated using the same formula as *s*, except that observation *j* is omitted. The hope is that be excluding this observation, a better estimate of $\sigma^2$ 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

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

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

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

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

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

## DFBETAS

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

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

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

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.

$$R^2_{predict} = 1 - \frac{PRESS}{SS_{tot}}$$

## Sum |Press residuals|

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

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

# Bootstrapping

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

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

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

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

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

## Modified Residuals

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

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

where

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

## Bootstrap the Observations

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

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

## Bootstrap Prediction Intervals

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

$$\hat{y}_+ = \hat{y} - \sum x_i(b_i^* - b_i) + e_+^*$$

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.

# Data Structure

The data are entered in two or more columns. An example of data appropriate for this procedure is shown below. These data are from a study of the relationship of several variables with a person's I.Q. Fifteen people were studied. Each person's IQ was recorded along with scores on five different personality tests. The data are contained in the IQ dataset. We suggest that you open this database now so that you can follow along with the example.

**IQ Dataset**

| Test1 | Test2 | Test3 | Test4 | Test5 | IQ |
|---|---|---|---|---|---|
| 83 | 34 | 65 | 63 | 64 | 106 |
| 73 | 19 | 73 | 48 | 82 | 92 |
| 54 | 81 | 82 | 65 | 73 | 102 |
| 96 | 72 | 91 | 88 | 94 | 121 |
| 84 | 53 | 72 | 68 | 82 | 102 |
| 86 | 72 | 63 | 79 | 57 | 105 |
| 76 | 62 | 64 | 69 | 64 | 97 |
| 54 | 49 | 43 | 52 | 84 | 92 |
| 37 | 43 | 92 | 39 | 72 | 94 |
| 42 | 54 | 96 | 48 | 83 | 112 |
| 71 | 63 | 52 | 69 | 42 | 130 |
| 63 | 74 | 74 | 71 | 91 | 115 |
| 69 | 81 | 82 | 75 | 54 | 98 |
| 81 | 89 | 64 | 85 | 62 | 96 |
| 50 | 75 | 72 | 64 | 45 | 103 |

# Missing Values

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

# Example 1 – Multiple Regression (All Reports)

This section presents an example of how to run a multiple regression analysis of the data presented earlier in this chapter. The data are in the IQ dataset. This example will run a regression of *IQ* on *Test1* through *Test5*. This regression program outputs over thirty different reports and plots, many of which contain duplicate information. For the purposes of annotating the output, all output is displayed. Normally, you would only select a few these reports.

## Setup

To run this example, complete the following steps:

**1 Open the IQ example dataset**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **IQ** and click **OK**.

**2 Specify the Multiple Regression procedure options**

- Find and open the **Multiple Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab

Y ..... **IQ**
Numeric X's ..... **Test1-Test5**

Reports Tab

Alphas, Confidence Level, and Power

Compute Power ..... **Checked**

Select Reports

All Available Reports..... **Checked** (click the *Check All* button)

Plots Tab

Select Plots

All Available Plots ..... **Checked** (click the *Check All* button)

**3 Run the procedure**

- Click the **Run** button to perform the calculations and generate the output.

## Run Summary

**Run Summary**

| Item | Value | Rows | Value |
|---|---|---|---|
| Dependent Variable (Y) | IQ | Rows Processed | 17 |
| Number of Independent Variables (X) | 5 | Rows Used in Estimation | 15 |
| Weight Variable | None | Rows with X's Missing | 0 |
| $R^2$ | 0.3991 | Rows with Y Missing | 2 |
| Adjusted $R^2$ | 0.0652 | | |
| Coefficient of Variation | 0.1021 | | |
| Mean Square Error (MSE) | 113.4648 | | |
| Square Root of MSE | 10.65198 | | |
| Average \|Percent Error\| | 6.218 | | |
| Completion Status | Normal Completion | | |

This report summarizes the multiple regression results. It presents the variables used, the number of rows used, and the basic results.

### $R^2$

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

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

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

1. *Additional independent variables*. It is possible to increase $R^2$ by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This case happens when your sample size is small.
2. *Range of the independent variables*. $R^2$ is influenced by the range of each independent variable. $R^2$ increases as the range of the *X's* increases and decreases as the range of the *X's* decreases.
3. *Slope magnitudes*. $R^2$ does not measure the magnitude of the slopes.
4. *Linearity*. $R^2$ does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *x* and *Y* was a perfect circle. The $R^2$ value of this relationship would be zero.
5. *Predictability*. A large $R^2$ does not necessarily mean high predictability, nor does a low $R^2$ necessarily mean poor predictability.
6. *No-intercept model*. The definition of $R^2$ assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of $R^2$ changes dramatically. The fact that your $R^2$ value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying meaning of $R^2$.
7. *Sample size*. $R^2$ is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

### Adjusted $R^2$

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

### Coefficient of Variation

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

$$CV = \frac{\sqrt{MSE}}{\bar{y}}$$

### Average |Percent Error|

This is the average of the absolute percent errors. It is another measure of the goodness of fit of the regression model to the data. It is calculated using the formula

$$AAPE = \frac{100 \sum_{j=1}^{N} \left| \frac{y_j - \hat{y}_j}{y_j} \right|}{N}$$

Note that when the dependent variable is zero, its predicted value is used in the denominator.

## Descriptive Statistics

**Descriptive Statistics**

| Variable | Count | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Test1 | 15 | 67.93333 | 17.39239 | 37 | 96 |
| Test2 | 15 | 61.4 | 19.39735 | 19 | 89 |
| Test3 | 15 | 72.33334 | 14.73415 | 43 | 96 |
| Test4 | 15 | 65.53333 | 13.95332 | 39 | 88 |
| Test5 | 15 | 69.93333 | 16.15314 | 42 | 94 |
| IQ | 15 | 104.3333 | 11.0173 | 92 | 130 |

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected.

## Correlation Matrix

**Correlation Matrix**

| | Test1 | Test2 | Test3 | Test4 | Test5 | IQ |
|---|---|---|---|---|---|---|
| **Test1** | 1.0000 | 0.1000 | -0.2608 | 0.7539 | 0.0140 | 0.2256 |
| **Test2** | 0.1000 | 1.0000 | 0.0572 | 0.7196 | -0.2814 | 0.2407 |
| **Test3** | -0.2608 | 0.0572 | 1.0000 | -0.1409 | 0.3473 | 0.0741 |
| **Test4** | 0.7539 | 0.7196 | -0.1409 | 1.0000 | -0.1729 | 0.3714 |
| **Test5** | 0.0140 | -0.2814 | 0.3473 | -0.1729 | 1.0000 | -0.0581 |
| **IQ** | 0.2256 | 0.2407 | 0.0741 | 0.3714 | -0.0581 | 1.0000 |

Pearson correlations are given for all variables. Outliers, nonnormality, nonconstant variance, and nonlinearities can all impact these correlations. Note that these correlations may differ from pair-wise correlations generated by the correlation matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with one another may cause collinearity problems.

## Regression Coefficient T-Tests

**Regression Coefficient T-Tests**

| Independent Variable | Regression Coefficient b(i) | Standard Error Sb(i) | Standardized Coefficient | T-Test of H0: β(i) = 0 | | | |
|---|---|---|---|---|---|---|---|
| | | | | T-Statistic | P-Value | Reject H0 at α = 0.05? | Power* |
| Intercept | 85.24039 | 23.69514 | 0.0000 | 3.597 | 0.0058 | Yes | 0.8915 |
| Test1 | -1.933571 | 1.029096 | -3.0524 | -1.879 | 0.0930 | No | 0.3896 |
| Test2 | -1.659881 | 0.872896 | -2.9224 | -1.902 | 0.0897 | No | 0.3974 |
| Test3 | 0.1049543 | 0.219902 | 0.1404 | 0.477 | 0.6445 | No | 0.0713 |
| Test4 | 3.778377 | 1.834497 | 4.7853 | 2.060 | 0.0695 | No | 0.4522 |
| Test5 | -0.04057754 | 0.2012205 | -0.0595 | -0.202 | 0.8447 | No | 0.0538 |

* Power was calculated using the observed T-Statistic as the population effect size with a significance level of α = 0.05.

This section reports the values and significance tests of the regression coefficients. Before using this report, check that the assumptions are reasonable. For instance, collinearity can cause the t-tests to give false results and the regression coefficients to be of the wrong magnitude or sign.

### Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the *Report Options* tab. This should create a better-looking report when the names are extra-long.

### Regression Coefficient b(i)

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in that particular X when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \ldots, b_p$.

### Standard Error Sb(i)

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

### Standardized Coefficient

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

When the independent variables have vastly different scales of measurement, this value provides a way of making comparisons among variables. The formula for the standardized regression coefficient is:

$$b_{j,std} = b_j \left( \frac{s_{X_j}}{s_Y} \right)$$

where $s_Y$ and $s_{X_j}$ are the standard deviations for the dependent variable and the $j^{th}$ independent variable.

### T-Statistic

This is the t-test value for testing the hypothesis that $\beta_j = 0$ versus the alternative that $\beta_j \neq 0$ after removing the influence of all other *X's*. This *t*-value has *n*-*p*-1 degrees of freedom.

To test for a value other than zero, use the formula below. There is an easier way to test hypothesized values using confidence limits. See the discussion below under Confidence Limits. The formula for the *t*-test is

$$t_j = \frac{b_j - \beta_j^*}{s_{b_j}}$$

### P-Value

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

### Reject H0 at α = 0.05?

This is the conclusion reached about the null hypothesis. It will be either reject *H0* at the 5% level of significance or not.

Note that the level of significance is specified in the Tests Alpha box on the *Reports* tab panel.

### Power

Power is the probability of rejecting the null hypothesis that $\beta_j = 0$ when $\beta_j = \beta_j^* \neq 0$. The power is calculated for the case when $\beta_j^* = b_j$, $\sigma^2 = s^2$, and alpha is as specified in the Tests Alpha option.

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

## Regression Coefficient Confidence Intervals

**Regression Coefficient Confidence Intervals**

| Independent Variable | Regression Coefficient b(i) | Standard Error Sb(i) | 95% Confidence Interval Limits for β(i) | |
|---|---|---|---|---|
| | | | Lower | Upper |
| Intercept | 85.24039 | 23.69514 | 31.63827 | 138.8425 |
| Test1 | -1.933571 | 1.029096 | -4.261548 | 0.3944054 |
| Test2 | -1.659881 | 0.872896 | -3.634509 | 0.3147467 |
| Test3 | 0.1049543 | 0.219902 | -0.3924985 | 0.6024072 |
| Test4 | 3.778377 | 1.834497 | -0.3715436 | 7.928297 |
| Test5 | -0.04057754 | 0.2012205 | -0.49577 | 0.4146149 |

Note: The T-Value used to calculate the confidence interval limits was 2.262.

### Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the *Report Options* tab. This should create a better-looking report when the names are extra-long.

### Regression Coefficient b(i)

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in *x* when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \ldots, b_p$.

### Standard Error Sb(i)

The standard error of the regression coefficient, $s_{b_j}$, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

### 95% Confidence Interval Limits for β(i) (Lower and Upper)

These are the lower and upper values of a $100(1-\alpha)\%$ interval estimate for $\beta_j$ based on a *t*-distribution with *n*-*p*-1 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed.

These confidence limits may be used for significance testing values of $\beta_j$ other than zero. If a specific value is not within this interval, it is significantly different from that value. Note that these confidence limits are set up as if you are interested in each regression coefficient separately.

The formulas for the lower and upper confidence limits are:

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

### Note: The T-Value ...

This is the value of $t_{1-\alpha/2,n-p-1}$ used to construct the confidence limits.

## Estimated Equation

**Estimated Equation**

IQ =
85.2403846967439 - 1.93357123818932 * Test1 - 1.65988116961152 * Test2 + 0.104954325385776 * Test3 + 3.77837667941384 * Test4 - 0.0405775409260279 * Test5

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.

## Analysis of Variance Summary

**Analysis of Variance Summary**

| Source | DF | $R^2$ Lost If Term(s) Removed | Sum of Squares | Mean Square | F-Ratio | P-Value | Power* |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | | 163281.7 | 163281.7 | | | |
| Model | 5 | 0.3991 | 678.1504 | 135.6301 | 1.195 | 0.3835 | 0.2565 |
| Error | 9 | 0.6009 | 1021.183 | 113.4648 | | | |
| Total (Adjusted) | 14 | | 1699.333 | 121.381 | | | |

* Power was calculated using the observed F-Ratio as the population effect size with a significance level of Alpha = 0.05.

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

### Source

This represents a partition of the variation in Y.

### DF

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

### $R^2$ Lost if Term(s) Removed

This is the amount that $R^2$ is reduced when this term is removed from the regression model.

### Sum of Squares

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

$$SS_{Intercept} = n\bar{y}^2$$

$$SS_{Model} = \sum\left(\hat{y}_j - \bar{y}\right)^2$$

$$SS_{Error} = \sum\left(y_j - \hat{y}_j\right)^2$$

$$SS_{Total} = \sum\left(y_j - \bar{y}\right)^2$$

### 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.

### F-Ratio

This is the *F*-statistic for testing the null hypothesis that all $\beta_j = 0$. This *F*-statistic has *p* degrees of freedom for the numerator variance and *n*-*p*-1 degrees of freedom for the denominator variance.

### P-Value

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 $\alpha$, then the null hypothesis is accepted.

### Power

Power is the probability of rejecting the null hypothesis that all the regression coefficients are zero when at least one is not. Power is calculated using the observed F-Ratio as the population effect size with the specified significance level.

# Analysis of Variance Detail

**Analysis of Variance Detail**

| Source | DF | $R^2$ Lost If Term(s) Removed | Sum of Squares | Mean Square | F-Ratio | P-Value | Power* |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | | 163281.7 | 163281.7 | | | |
| Model | 5 | 0.3991 | 678.1504 | 135.6301 | 1.195 | 0.3835 | 0.2565 |
| Test1 | 1 | 0.2357 | 400.562 | 400.562 | 3.530 | 0.0930 | 0.3896 |
| Test2 | 1 | 0.2414 | 410.2892 | 410.2892 | 3.616 | 0.0897 | 0.3974 |
| Test3 | 1 | 0.0152 | 25.8466 | 25.8466 | 0.228 | 0.6445 | 0.0713 |
| Test4 | 1 | 0.2832 | 481.3241 | 481.3241 | 4.242 | 0.0695 | 0.4522 |
| Test5 | 1 | 0.0027 | 4.614109 | 4.614109 | 0.041 | 0.8447 | 0.0538 |
| Error | 9 | 0.6009 | 1021.183 | 113.4648 | | | |
| Total (Adjusted) | 14 | | 1699.333 | 121.381 | | | |

* Power was calculated using the observed F-Ratio as the population effect size with a significance level of Alpha = 0.05.

This analysis of variance table provides a line for each term in the model. It is especially useful when you have categorical independent variables.

### Source

This is the term from the design model.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the *Report Options* tab. This should create a better-looking report when the names are extra-long.

### DF

This is the number of degrees of freedom that the model is degrees of freedom is reduced when this term is removed from the model. This is the numerator degrees of freedom of the *F-test*.

### $R^2$ Lost if Term(s) Removed

This is the amount that $R^2$ is reduced when this term is removed from the regression model.

### Sum of Squares

This is the amount that the model sum of squares that are reduced when this term is removed from the model.

### Mean Square

The mean square is the sum of squares divided by the degrees of freedom.

### F-Ratio

This is the *F*-statistic for testing the null hypothesis that all $\beta_j$ associated with this term are zero. This *F-statistic* has *DF* and *n-p*-1 degrees of freedom.

### P-Value

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 $\alpha$, then the null hypothesis is accepted.

### Power

Power is the probability of rejecting the null hypothesis that all the regression coefficients are zero when at least one is not. Power is calculated using the observed F-Ratio as the population effect size with the specified significance level.

## Residual Normality Tests

**Residual Normality Tests**

| | Test of H0: Residuals Normally Distributed | | |
|---|---|---|---|
| Test Name | Test Statistic Value | P-Value | Reject H0 at α = 0.2? |
| Shapiro-Wilk | 0.908 | 0.1243 | Yes |
| Anderson-Darling | 0.458 | 0.2639 | No |
| D'Agostino Skewness | 2.033 | 0.0421 | Yes |
| D'Agostino Kurtosis | 1.580 | 0.1141 | Yes |
| D'Agostino Omnibus | 6.629 | 0.0364 | Yes |

This report gives the results of applying several normality tests to the residuals. The Shapiro-Wilk test is probably the most popular, so it is given first. These tests are discussed in detail in the Normality Test section of the Descriptive Statistics procedure.

## Serial-Correlation of Residuals and the Durbin-Watson Test for Serial Correlation

**Serial Correlation of Residuals**

| Lag | Serial Correlation | Lag | Serial Correlation | Lag | Serial Correlation |
|---|---|---|---|---|---|
| 1 | 0.4529 | 9 | -0.2769 | 17 | 0.0000 |
| 2 | -0.2507 | 10 | -0.2287 | 18 | 0.0000 |
| 3 | -0.5518 | 11 | -0.0197 | 19 | 0.0000 |
| 4 | -0.3999 | 12 | 0.0669 | 20 | 0.0000 |
| 5 | 0.0780 | 13 | 0.0000 | 21 | 0.0000 |
| 6 | 0.2956 | 14 | 0.0000 | 22 | 0.0000 |
| 7 | 0.1985 | 15 | 0.0000 | 23 | 0.0000 |
| 8 | -0.0016 | 16 | 0.0000 | 24 | 0.0000 |

Above serial correlations are significant if their absolute values are greater than 0.5164.

**Durbin-Watson Test for Serial Correlation**

| | Test of H0: ρ(1) = 0 | | |
|---|---|---|---|
| Test Type | Test Statistic Value | P-Value | Reject H0 at α = 0.2? |
| Positive Serial Correlation Test | 1.001 | 0.0072 | Yes |
| Negative Serial Correlation Test | 1.001 | 0.9549 | No |

This section reports the autocorrelation structure of the residuals. Of course, this report is only useful if the data represent a time series.

## Lag and Serial Correlation

The lag, *k,* is the number of periods (rows) back. The correlation here is the sample autocorrelation coefficient of lag *k*. It is computed as:

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

To test the null hypothesis that $\rho_k = 0$ at a 5% level of significance with a large-sample normal approximation, reject when the absolute value of the autocorrelation coefficient, $|r_k|$, is greater than two over the square root of *N.*

## Test Statistic Value

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

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

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

# PRESS Statistics

**PRESS Statistics**

| Parameter | From PRESS Residuals | From Regular Residuals |
|---|---|---|
| Sum of Squared Residuals | 2839.941 | 1021.183 |
| Sum of \|Residuals\| | 169.6438 | 99.12155 |
| $R^2$ | 0.0000 | 0.3991 |

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

## Sum of Squared Residuals

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

$$\sum \left(y_j - \hat{y}_{j,-j}\right)^2$$

## Sum of |Press Residuals|

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

$$\sum \left|y_j - \hat{y}_{j,-j}\right|$$

## Press $R^2$

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

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

# $R^2$ Report

**R² Report**

| Independent Variable (IV) | Total R² for this IV and IV's Above | Increase in R² if this IV Included with IV's Above | Decrease in R² if this IV was Removed | R² if this IV was Fit Alone | Partial R² if Adjusted for All Other IV's |
|---|---|---|---|---|---|
| Test1 | 0.0509 | 0.0509 | 0.2357 | 0.0509 | 0.2817 |
| Test2 | 0.0990 | 0.0480 | 0.2414 | 0.0579 | 0.2866 |
| Test3 | 0.1131 | 0.0142 | 0.0152 | 0.0055 | 0.0247 |
| Test4 | 0.3964 | 0.2832 | 0.2832 | 0.1379 | 0.3203 |
| Test5 | 0.3991 | 0.0027 | 0.0027 | 0.0034 | 0.0045 |

$R^2$ reflects the percent of variation in *Y* explained by the independent variables in the model. A value of $R^2$ near zero indicates a complete lack of fit between *Y* and the *Xs,* while a value near one indicates a perfect fit.

In this section, various types of $R^2$ values are given to provide insight into the variation in the dependent variable explained either by the independent variables added in order (i.e., sequential) or by the independent variables added last. This information is valuable in an analysis of which variables are most important.

## Independent Variable

This is the name of the independent variable reported on in this row.

## Total $R^2$ for This I.V. and Those Above

This is the $R^2$ value that would result from fitting a regression with this independent variable and those listed above it. The IV's below it are ignored.

## $R^2$ Increase When This IV Added to Those Above

This is the amount that this IV adds to $R^2$ when it is added to a regression model that includes those IV's listed above it in the report.

## $R^2$ Decrease When This IV is Removed

This is the amount that $R^2$ would be reduced if this IV were removed from the model. Large values here indicate important independent variables, while small values indicate insignificant variables.

One of the main problems in interpreting these values is that each assumes all other variables are already in the equation. This means that if two variables both represent the same underlying information, they will each seem to be insignificant after considering the other. If you remove both, you will lose the information that either one could have brought to the model.

## $R^2$ When This IV Is Fit Alone

This is the $R^2$ that would be obtained if the dependent variable were only regressed against this one independent variable. Of course, a large $R^2$ value here indicates an important independent variable that can stand alone.

### Partial $R^2$ Adjusted For All Other IV's

The is the square of the partial correlation coefficient. The partial $R^2$ reflects the percent of variation in the dependent variable explained by one independent variable controlling for the effects of the rest of the independent variables. Large values for this partial $R^2$ indicate important independent variables.

## Variable Omission Report

**Variable Omission Report**

| Independent Variable (IV) | When IV Omitted R² | When IV Omitted MSE | When IV Omitted Mallow's Cp | Test of H0: β(i) = 0 P-Value | R² of Regression of This IV on Other IV's |
|---|---|---|---|---|---|
| Full Model | 0.3991 | 113.4648 | | | |
| Test1 | 0.1634 | 142.1745 | 7.530 | 0.0930 | 0.9747 |
| Test2 | 0.1576 | 143.1472 | 7.616 | 0.0897 | 0.9717 |
| Test3 | 0.3839 | 104.703 | 4.228 | 0.6445 | 0.2280 |
| Test4 | 0.1158 | 150.2507 | 8.242 | 0.0695 | 0.9876 |
| Test5 | 0.3964 | 102.5797 | 4.041 | 0.8447 | 0.2329 |

One way of assessing the importance of an independent variable is to examine the impact on various goodness-of-fit statistics of removing it from the model. This section provides this.

### Independent Variable (IV)

This is the name of the predictor variable reported on in this row. Note that the *Full Model* row gives the statistics when no variables are omitted.

### $R^2$ (When IV Omitted)

This is the $R^2$ for the multiple regression model when this independent variable is omitted, and the remaining independent variables are retained. If this $R^2$ is close to the $R^2$ for the full model, this variable is not very important. On the other hand, if this $R^2$ is much smaller than that of the full model, this independent variable is important.

### MSE (When IV Omitted)

This is the mean square error for the multiple regression model when this IV is omitted and the remaining IV's are retained. If this MSE is close to the MSE for the full model, this variable may not be very important. On the other hand, if this MSE is much larger than that of the full model, this IV is important.

### Mallow's Cp (When IV Omitted)

Another criterion for variable selection and importance is Mallow's *Cp* statistic. The optimum model will have a *Cp* value close to *p*+1, where *p* is the number of independent variables. A *Cp* greater than (*p*+1) indicates that the regression model is over specified (contains too many variables and stands a chance of having collinearity problems). On the other hand, a model with a *Cp* less than (*p*+1) indicates that the regression model is underspecified (at least one important independent variable has been omitted). The formula for the *Cp* statistic is as follows, where *k* is the maximum number of independent variables available

$$C_p = (n - p - 1)\left[\frac{MSE_p}{MSE_k}\right] - [n - 2(p + 1)]$$

### Test of H0: β(i) = 0 P-Value

This is the two-tail *p*-value for testing the significance of the regression coefficient. Most likely, you would deem IV's with small *p*-values as important. However, you must be careful here. Collinearity can cause extra-large *p*-values, so you must check for its presence.

### $R^2$ Of Regression of This IV on Other IV's

This is the $R^2$ value that would result if this independent variable were regressed on the remaining independent variables. A high value indicates a redundancy between this IV and the other IV's. IV's with a high value here (above 0.90) are candidates for omission from the model.

## Sum of Squares and Correlations

**Sum of Squares and Correlations**

| Independent Variable | Sum of Squares | | | Correlation | |
|---|---|---|---|---|---|
| | Sequential | Incremental | Last | Simple | Partial |
| Test1 | 86.5252 | 86.5252 | 400.562 | 0.2256 | -0.5308 |
| Test2 | 168.1614 | 81.6362 | 410.2892 | 0.2407 | -0.5354 |
| Test3 | 192.2748 | 24.11342 | 25.8466 | 0.0741 | 0.1571 |
| Test4 | 673.5363 | 481.2615 | 481.3241 | 0.3714 | 0.5660 |
| Test5 | 678.1504 | 4.614109 | 4.614109 | -0.0581 | -0.0671 |

This section provides the sum of squares and correlations equivalent to the *$R^2$ Section*.

### Independent Variable

This is the name of the IV reported on in this row.

### Sequential Sum of Squares

The is the sum of squares value that would result from fitting a regression with this independent variable and those above it. The IV's below it are ignored.

### Incremental Sum of Squares

This is the amount that this predictor adds to the sum of squares value when it is added to a regression model that includes those predictors listed above it.

### Last Sum of Squares

This is the amount that the model sum of squares would be reduced if this variable were removed from the model.

### Simple Correlation

This is the Pearson correlation coefficient between the dependent variable and the specified independent variable.

### Partial Correlation

The partial correlation coefficient is a measure of the strength of the linear relationship between $Y$ and $X_j$ after adjusting for the remaining ($p$-1) variables.

## Sequential Models Report

**Sequential Models Report**

| Independent Variable | Included | | | Omitted | | |
|---|---|---|---|---|---|---|
| | $R^2$ | F-Ratio | P-Value | $R^2$ | F-Ratio | P-Value |
| Test1 | 0.0509 | 0.697 | 0.4187 | 0.3482 | 1.304 | 0.3390 |
| Test2 | 0.0990 | 0.659 | 0.5351 | 0.3001 | 1.498 | 0.2801 |
| Test3 | 0.1131 | 0.468 | 0.7107 | 0.2859 | 2.141 | 0.1735 |
| Test4 | 0.3964 | 1.641 | 0.2390 | 0.0027 | 0.041 | 0.8447 |
| Test5 | 0.3991 | 1.195 | 0.3835 | 0.0000 | | |

Notes:
1. INCLUDED variables are those listed from current row up (includes current row).
2. OMITTED variables are those listed below (but not including) this row.

This section examines the step-by-step effect of adding variables to the regression model.

### Independent Variable

This is the name of the predictor variable reported on in this row.

### Included $R^2$

This is the $R^2$ that would be obtained if only those IV's on this line and above were in the regression model.

### Included F-Ratio

This is an $F$-ratio for testing the hypothesis that the regression coefficients ($\beta$'s) for the IV's listed on this row and above are zero.

### Included P-Value

This is the $p$-value for the associated $F$-ratio.

### Omitted $R^2$

This is the $R^2$ for the full model minus the *Included* $R^2$. This is the amount of $R^2$ explained by the independent variables listed below the current row. Large values indicate that there is much more to come with later independent variables. On the other hand, small values indicate that remaining independent variables contribute little to the regression model.

### Omitted F-Ratio

This is an *F*-ratio for testing the hypothesis that the regression coefficients (*β's*) for the variables listed below this row are all zero. The alternative is that at least one coefficient is nonzero.

### Omitted P-Value

This is the *p*-value for the associated *F*-ratio.

## Multicollinearity Report

**Multicollinearity Report**

| Independent Variable (IV) | Variance Inflation Factor | R² Versus Other IV's | Tolerance | Diagonal of X'X Inverse |
|---|---|---|---|---|
| Test1 | 39.5273 | 0.9747 | 0.0253 | 0.009333631 |
| Test2 | 35.3734 | 0.9717 | 0.0283 | 0.006715277 |
| Test3 | 1.2953 | 0.2280 | 0.7720 | 0.0004261841 |
| Test4 | 80.8456 | 0.9876 | 0.0124 | 0.02966012 |
| Test5 | 1.3035 | 0.2329 | 0.7671 | 0.0003568483 |

This report provides information useful in assessing the amount of multicollinearity in your data.

### Variance Inflation Factor

The variance inflation factor (*VIF*) is a measure of multicollinearity. It is the reciprocal of $1 - R_X^2$, where $R_X^2$ is the $R^2$ obtained when this variable is regressed on the remaining IV's. A *VIF* of 10 or more for large data sets indicates a collinearity problem since the $R_X^2$ with the remaining *IV*'s is 90 percent. For small data sets, even *VIF's* of 5 or more can signify collinearity. Variables with a high *VIF* are candidates for exclusion from the model.

$$VIF_j = \frac{1}{1 - R_j^2}$$

### $R^2$ Versus Other IV's

$R_X^2$ is the $R^2$ obtained when this variable is regressed on the remaining independent variables. A high $R_X^2$ indicates a lot of overlap in explaining the variation among the remaining independent variables.

### Tolerance

Tolerance is just $1 - R_X^2$, the denominator of the variance inflation factor.

### Diagonal of X'X Inverse

The *X'X* inverse is an important matrix in regression. This is the $j^{th}$ row and $j^{th}$ column element of this matrix.

## Eigenvalues of Centered Correlations

**Eigenvalues of Centered Correlations**

| No. | Eigenvalue | Incremental Percent | Cumulative Percent | Condition Number |
|---|---|---|---|---|
| 1 | 2.2150 | 44.29912 | 44.29912 | 1 |
| 2 | 1.2277 | 24.55417 | 68.85329 | 1.804138 |
| 3 | 1.1062 | 22.12431 | 90.9776 | 2.002283 |
| 4 | 0.4446 | 8.892469 | 99.87006 | 4.981644 |
| 5 | 0.0065 | 0.1299325 | 100 | 340.9394 |

Some Condition Numbers greater than 100. Multicollinearity is a MILD problem.

This section gives an eigenvalue analysis of the independent variables when they have been centered and scaled.

### Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of IV's. Eigenvalues near zero indicate a high degree of is collinearity in the data.

### Incremental Percent

Incremental percent is the percent this eigenvalue is of the total. In an ideal situation, these percentages would be equal. Percents near zero indicate collinearity in the data.

### Cumulative Percent

This is the running total of the Incremental Percent.

### Condition Number

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variances. Condition numbers greater than 1000 indicate a severe collinearity problem while condition numbers between 100 and 1000 indicate a mild collinearity problem.

# Eigenvector Percent of Regression-Coefficent-Variance using Centered Correlations

**Eigenvector Percent of Regression-Coefficient-Variance using Centered Correlations**

| No. | Eigenvalue | Test1 | Test2 | Test3 | Test4 | Test5 |
|---|---|---|---|---|---|---|
| 1 | 2.2150 | 0.2705 | 0.2850 | 1.8773 | 0.2331 | 2.3798 |
| 2 | 1.2277 | 0.0330 | 0.1208 | 31.1222 | 0.0579 | 23.6898 |
| 3 | 1.1062 | 0.8089 | 0.8397 | 7.6430 | 0.0015 | 14.3442 |
| 4 | 0.4446 | 0.8059 | 1.0889 | 59.3291 | 0.0002 | 59.5804 |
| 5 | 0.0065 | 98.0817 | 97.6657 | 0.0284 | 99.7072 | 0.0058 |

This report displays how the eigenvectors associated with each eigenvalue are related to the independent variables.

### No.

The number of the eigenvalue.

### Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is collinearity in your data.

### Values

The rest of this report gives a breakdown of what percentage each eigenvector is of the total variation for the regression coefficient. Hence, the percentages sum to 100 down a column.

A small eigenvalue (large condition number) along with a subset of two or more independent variables having high variance percentages indicates a dependency involving the independent variables in that subset. This dependency has damaged or contaminated the precision of the regression coefficients estimated in the subset. Two or more percentages of at least 50% for an eigenvector or eigenvalue suggest a problem. For certain, when there are two or more variance percentages greater than 90%, there is definitely a collinearity problem.

Again, take the following steps when using this table.

1. Find rows with condition numbers greater than 100 (find these in the *Eigenvalues of Centered Correlations* report).
2. Scan across each row found in step 1 for two or more percentages greater than 50. If two such percentages are found, the corresponding variables are being influenced by collinearity problems. You should remove one and re-run your analysis.

# Eigenvalues of Uncentered Correlations Section

**Eigenvalues of Uncentered Correlations**

| No. | Eigenvalue | Incremental Percent | Cumulative Percent | Condition Number |
|---|---|---|---|---|
| 1 | 5.7963 | 96.60564 | 96.60564 | 1 |
| 2 | 0.1041 | 1.734821 | 98.34045 | 55.68621 |
| 3 | 0.0670 | 1.116414 | 99.45687 | 86.53209 |
| 4 | 0.0214 | 0.3567024 | 99.81357 | 270.8298 |
| 5 | 0.0109 | 0.1809921 | 99.99456 | 533.756 |
| 6 | 0.0003 | 0.005437351 | 100 | 17767.04 |

Some Condition Numbers greater than 1000. Multicollinearity is a SEVERE problem.

This report gives an eigenvalue analysis of the independent variables when they have been scaled but not centered (the intercept is included in the collinearity analysis). The eigenvalues for this situation are generally not the same as those in the previous eigenvalue analysis. Also, the condition numbers are much higher.

## Eigenvalue

The eigenvalues of the scaled, but not centered, matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is collinearity in your data.

## Incremental Percent

Incremental percent is the percent this eigenvalue is of the total. In an ideal situation, these percentages would be equal. Percents near zero mean that there is collinearity in your data.

## Cumulative Percent

This is the running total of the *Incremental Percent.*

## Condition Number

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. Since the eigenvalues are really variances, the condition number is a ratio of variances. There has not been any formalization of rules on condition numbers for uncentered matrices. You might use the criteria mentioned earlier for mild collinearity and severe collinearity. Since the collinearity will always be worse with the intercept in the model, it is advisable to have more relaxed criteria for mild and severe collinearity, say 500 and 5000, respectively.

# Eigenvector Percent of Regression-Coefficent-Variance using Uncentered Correlations

**Eigenvector Percent of Regression-Coefficient-Variance using Uncentered Correlations**

| No. | Eigenvalue | Test1 | Test2 | Test3 | Test4 | Test5 | Intercept |
|---|---|---|---|---|---|---|---|
| 1 | 5.7963 | 0.0042 | 0.0068 | 0.0826 | 0.0015 | 0.1033 | 0.0397 |
| 2 | 0.1041 | 0.0308 | 0.8177 | 3.8156 | 0.0610 | 11.8930 | 0.2599 |
| 3 | 0.0670 | 1.1375 | 0.9627 | 7.4272 | 0.0261 | 0.0897 | 0.0106 |
| 4 | 0.0214 | 0.2675 | 0.9263 | 51.4298 | 0.0006 | 79.7835 | 1.6692 |
| 5 | 0.0109 | 0.4157 | 0.0499 | 37.2046 | 0.0931 | 8.1292 | 97.0221 |
| 6 | 0.0003 | 98.1444 | 97.2367 | 0.0402 | 99.8177 | 0.0013 | 0.9986 |

This report displays how the eigenvectors associated with each eigenvalue are related to the independent variables.

### No.

The number of the eigenvalue.

### Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero mean that there is collinearity in your data.

### Values

The rest of this report gives a breakdown of what percentage each eigenvector is of the total variation for the regression coefficient. Hence, the percentages sum to 100 down a column.

A small eigenvalue (large condition number) along with a subset of two or more independent variables having high variance percentages indicates a dependency involving the independent variables in that subset. This dependency has damaged or contaminated the precision of the regression coefficients estimated in the subset. Two or more percentages of at least 50% for an eigenvector or eigenvalue suggest a problem. For certain, when there are two or more variance percentages greater than 90%, there is definitely a collinearity problem.

# Predicted Values with Confidence Limits for Means

**Predicted Values with Confidence Interval Limits for Means**

| Row | IQ | | Standard Error of Predicted | 95% Confidence Interval Limits for the Mean | |
|---|---|---|---|---|---|
| | Actual | Predicted | | Lower | Upper |
| 1 | 106 | 110.58080 | 7.156587 | 94.39149 | 126.7701 |
| 2 | 92 | 98.24833 | 7.075774 | 82.24182 | 114.2548 |
| 3 | 102 | 97.61574 | 6.222865 | 83.53864 | 111.6928 |
| 4 | 121 | 118.33980 | 8.686609 | 98.68933 | 137.9903 |
| 5 | 102 | 96.00567 | 6.369277 | 81.59736 | 110.4140 |
| 6 | 105 | 102.23280 | 5.433011 | 89.94245 | 114.5231 |
| 7 | 97 | 100.20440 | 4.099899 | 90.92982 | 109.4791 |
| 8 | 92 | 97.07347 | 9.098985 | 76.49014 | 117.6568 |
| 9 | 94 | 96.41426 | 7.088691 | 80.37853 | 112.4500 |
| 10 | 112 | 102.46660 | 6.351596 | 88.09826 | 116.8349 |
| 11 | 130 | 107.84570 | 6.464094 | 93.22288 | 122.4685 |
| 12 | 115 | 112.93300 | 7.331204 | 96.34866 | 129.5173 |
| 13 | 98 | 107.16690 | 5.338698 | 95.08994 | 119.2439 |
| 14 | 96 | 106.25500 | 5.531748 | 93.74129 | 118.7687 |
| 15 | 103 | 111.61760 | 7.100195 | 95.55581 | 127.6793 |
| 16 | | 97.70544 | 7.031275 | 81.79959 | 113.6113 |
| 17 | | 100.19840 | 4.305268 | 90.45926 | 109.9376 |

Confidence intervals for the mean response of *Y* given specific levels for the IV's are provided here. It is important to note that violations of any regression assumptions will invalidate these interval estimates.

### Actual Y

This is the actual value of *Y*.

### Predicted Y

The predicted value of *Y*. It is predicted using the values of the IV's for this row. If the input data had all IV values but no value for *Y*, the predicted value is still provided.

### Standard Error of Predicted

This is the standard error of the mean response for the specified values of the IV's. Note that this value is not constant for all IV's values. In fact, it is a minimum at the average value of each IV.

### 95% Confidence Interval Limits for the Mean (Lower and Upper)

These are the lower and upper limits of a 95% confidence interval estimate of the mean of *Y* for this observation. Note that you set the confidence interval alpha on the *Reports* tab of the procedure input window.

# Predicted Values with Prediction Limits for Individuals

**Predicted Values with Prediction Interval Limits for Individuals**

| Row | IQ | | Standard Error of Predicted | 95% Prediction Interval Limits for an Individual | |
|---|---|---|---|---|---|
| | Actual | Predicted | | Lower | Upper |
| 1 | 106 | 110.58080 | 12.83283 | 81.55093 | 139.6107 |
| 2 | 92 | 98.24833 | 12.78794 | 69.32001 | 127.1767 |
| 3 | 102 | 97.61574 | 12.33648 | 69.70868 | 125.5228 |
| 4 | 121 | 118.33980 | 13.74489 | 87.24670 | 149.4329 |
| 5 | 102 | 96.00567 | 12.41098 | 67.93008 | 124.0813 |
| 6 | 105 | 102.23280 | 11.95752 | 75.18298 | 129.2826 |
| 7 | 97 | 100.20440 | 11.41376 | 74.38472 | 126.0242 |
| 8 | 92 | 97.07347 | 14.00915 | 65.38257 | 128.7644 |
| 9 | 94 | 96.41426 | 12.79509 | 67.46976 | 125.3588 |
| 10 | 112 | 102.46660 | 12.40192 | 74.41148 | 130.5217 |
| 11 | 130 | 107.84570 | 12.45991 | 79.65940 | 136.0319 |
| 12 | 115 | 112.93300 | 12.93102 | 83.68099 | 142.1850 |
| 13 | 98 | 107.16690 | 11.91497 | 80.21338 | 134.1204 |
| 14 | 96 | 106.25500 | 12.00271 | 79.10297 | 133.4070 |
| 15 | 103 | 111.61760 | 12.80147 | 82.65864 | 140.5765 |
| 16 | | 97.70544 | 12.76337 | 68.83269 | 126.5782 |
| 17 | | 100.19840 | 11.48913 | 74.20823 | 126.1887 |

A prediction interval for the individual response of *Y* given specific values of the IV's is provided here for each row.

### Actual Y

This is the actual value of *Y*.

### Predicted Y

The predicted value of *Y*. It is predicted using the levels of the IV's for this row. If the input data had all values of the IV's but no value for *Y*, a predicted value is provided.

### Standard Error of Predicted

This is the standard deviation of the mean response for the specified levels of the IV's. Note that this value is not constant for all IV's. In fact, it is a minimum at the average value of each IV.

### 95% Prediction Interval Limits for an Individual (Lower and Upper)

These are the lower and upper limits of a 95% prediction interval estimate of an individual *Y* for this observation. Note that you set the prediction interval alpha on the *Reports* tab of the procedure input window.

# Residuals

**Residuals**

| Row | IQ Actual | IQ Predicted | Residual | Absolute Percent Error | Sqrt(MSE) Without This Row |
|---|---|---|---|---|---|
| 1 | 106 | 110.58080 | -4.580812 | 4.321520 | 11.084530 |
| 2 | 92 | 98.24833 | -6.248330 | 6.791663 | 10.904760 |
| 3 | 102 | 97.61574 | 4.384259 | 4.298293 | 11.135540 |
| 4 | 121 | 118.33980 | 2.660196 | 2.198509 | 11.180660 |
| 5 | 102 | 96.00567 | 5.994334 | 5.876798 | 10.984390 |
| 6 | 105 | 102.23280 | 2.767226 | 2.635453 | 11.240730 |
| 7 | 97 | 100.20440 | -3.204443 | 3.303550 | 11.231260 |
| 8 | 92 | 97.07347 | -5.073471 | 5.514642 | 10.758520 |
| 9 | 94 | 96.41426 | -2.414264 | 2.568366 | 11.240110 |
| 10 | 112 | 102.46660 | 9.533430 | 8.511992 | 10.489000 |
| 11 | 130 | 107.84570 | 22.154330 | 17.041790 | 5.525609 |
| 12 | 115 | 112.93300 | 2.067002 | 1.797393 | 11.253140 |
| 13 | 98 | 107.16690 | -9.166913 | 9.353992 | 10.659280 |
| 14 | 96 | 106.25500 | -10.254980 | 10.682270 | 10.471290 |
| 15 | 103 | 111.61760 | -8.617564 | 8.366568 | 10.532950 |
| 16 | | 97.70544 | | | |
| 17 | | 100.19840 | | | |

This section reports on the sample residuals, or $e_i$'s.

### Actual Y

This is the actual value of *Y*.

### Predicted Y

The predicted value of *Y* using the values of the IV's given on this row.

### Residual

This is the error in the predicted value. It is equal to the *Actual* minus the *Predicted.*

### Absolute Percent Error

This is percentage that the absolute value of the *Residual* is of the *Actual* value. Scrutinize rows with the large percent errors.

### Sqrt(MSE) Without This Row

This is the value of the square root of the mean square error that is obtained if this row is deleted. A perusal of this statistic for all observations will highlight observations that have an inflationary impact on mean square error and could be outliers.

# Regression Diagnostics Section

**Regression Diagnostics**

| Row | Standardized Residual | RStudent | Hat Diagonal | Cook's D | DFFITS | CovRatio |
|---|---|---|---|---|---|---|
| 1 | -0.5806035 | -0.5579470 | 0.4514 | 0.0462 | -0.5061 | 2.9388350000 |
| 2 | -0.7847395 | -0.7665493 | 0.4413 | 0.0811 | -0.6812 | 2.3714230000 |
| 3 | 0.5071279 | 0.4851061 | 0.3413 | 0.0222 | 0.3492 | 2.5862670000 |
| 4 | 0.4314977 | 0.4110945 | 0.6650 | 0.0616 | 0.5792 | 5.3386670000 |
| 5 | 0.7020788 | 0.6808328 | 0.3575 | 0.0457 | 0.5079 | 2.2505870000 |
| 6 | 0.3020241 | 0.2862051 | 0.2601 | 0.0053 | 0.1697 | 2.5776560000 |
| 7 | -0.3259411 | -0.3091301 | 0.1481 | 0.0031 | -0.1289 | 2.2161910000 |
| 8 | -0.9160629 | -0.9069914 | 0.7297 | 0.3775 | -1.4901 | 4.1683770000 |
| 9 | -0.3036504 | -0.2877622 | 0.4429 | 0.0122 | -0.2566 | 3.4207480000 |
| 10 | 1.1148720 | 1.1321960 | 0.3556 | 0.1143 | 0.8410 | 1.2896020000 |
| 11 | 2.6167290 | 5.0443960 | 0.3683 | 0.6652 | 3.8514 | 0.0006009752 |
| 12 | 0.2674776 | 0.2531887 | 0.4737 | 0.0107 | 0.2402 | 3.6717440000 |
| 13 | -0.9945074 | -0.9938271 | 0.2512 | 0.0553 | -0.5756 | 1.3464700000 |
| 14 | -1.1265500 | -1.1459900 | 0.2697 | 0.0781 | -0.6964 | 1.1151360000 |
| 15 | -1.0852630 | -1.0975280 | 0.4443 | 0.1569 | -0.9814 | 1.5725230000 |
| 16 | | | 0.4357 | | | |
| 17 | | | 0.1634 | | | |

This report presents various statistics known as *regression diagnostics*. They let you conduct an influence analysis of the observations. The interpretation of these values is explained in modern regression books. Belsley, Kuh, and Welsch (1980) devote an entire book to the study of regression diagnostics.

These statistics flag observations that exert three types of influence on the regression.

1. *Outliers in the residual space*. The *Studentized Residual*, the *RStudent*, and the *CovRatio* will flag observations that are influential because of large residuals.
2. *Outliers in the X-space*. The *Hat Diagonal* flags observations that are influential because they are outliers in the *X*-space.
3. *Parameter estimates and fit*. The *Dffits* shows the influence on fitted values. It also measures the impact on the regression coefficients. *Cook's D* measures the overall impact that a single observation has on the regression coefficient estimates.

## Standardized Residual

The variances of the observed residuals are not equal, making comparisons among the residuals difficult. One solution is to standardize the residuals by dividing by their standard deviations. This will give a set of standardized residuals with constant variance. The formula for this residual is

$$r_j = \frac{e_j}{\sqrt{MSE\left(1 - h_{jj}\right)}}$$

## RStudent

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

If the regression assumptions of normality are valid, a single value of the RStudent has a *t* distribution with *n*-*p*-1 degrees of freedom.

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

## Hat Diagonal

The hat diagonal, $h_{jj}$, captures an observation's remoteness in the *X*-space. Some authors refer to the hat diagonal as a measure of *leverage* in the *X*-space. Hat diagonals greater than two times the number of coefficients in the model divided by the number of observations are said to have *high leverage* (i.e., $h_{ii} > 2p/n$).

## Cook's D

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

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

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

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

This approximation is exact when no weight variable is used.

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

## DFFITS

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

$$D_j = \left(\frac{r_j^2}{p}\right)\left(\frac{h_{jj}}{1 - h_{jj}}\right)$$

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

### CovRatio

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

The general formula for the CovRatio is

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

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

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

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

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

## DFBETAS Section

**DFBETAS Report**

| Row | Test1 | Test2 | Test3 | Test4 | Test5 | Intercept |
|---|---|---|---|---|---|---|
| 1 | 0.2160 | 0.3128 | -0.0390 | -0.2556 | 0.1723 | -0.1466 |
| 2 | -0.1123 | 0.0190 | -0.0830 | 0.0871 | 0.0045 | -0.1311 |
| 3 | 0.1822 | 0.2370 | 0.0291 | -0.2075 | 0.0674 | -0.0623 |
| 4 | -0.1792 | -0.2157 | 0.2157 | 0.2393 | 0.1963 | -0.4376 |
| 5 | 0.3932 | 0.3443 | 0.0108 | -0.3638 | 0.1240 | -0.1485 |
| 6 | 0.0969 | 0.0868 | -0.0110 | -0.0842 | -0.0534 | -0.0058 |
| 7 | -0.0771 | -0.0707 | 0.0286 | 0.0728 | 0.0202 | -0.0231 |
| 8 | 0.1301 | -0.0182 | 1.2984 | -0.0051 | -0.8487 | -0.7366 |
| 9 | -0.0334 | -0.0370 | -0.1136 | 0.0561 | 0.0525 | -0.0690 |
| 10 | -0.1257 | -0.0712 | 0.3963 | 0.0570 | 0.1128 | -0.0482 |
| 11 | -1.1326 | -1.2189 | -1.2510 | 1.1521 | -2.2675 | 2.6301 |
| 12 | -0.1456 | -0.1150 | -0.0686 | 0.1379 | 0.1606 | -0.0486 |
| 13 | -0.0758 | -0.0896 | -0.3057 | 0.0612 | 0.3288 | 0.0913 |
| 14 | -0.1772 | -0.2373 | 0.1757 | 0.1532 | -0.0325 | 0.1435 |
| 15 | 0.5669 | 0.4799 | -0.0701 | -0.5124 | 0.5187 | -0.4637 |
| 16 | | | | | | |
| 17 | | | | | | |

### DFBETAS

The DFBETAS is an influence diagnostic which gives the number of standard errors that an estimated regression coefficient changes if the $j^{th}$ observation is deleted. If one has *N* observations and *p* independent variables, there are *Np* of these diagnostics. Sometimes, Cook's D may not show any overall influence on the regression coefficients, but this diagnostic gives the analyst more insight into individual coefficients. The criteria of influence for this diagnostic are varied, but Belsley, Kuh, and Welsch (1980) recommend a cutoff of $2 / \sqrt{N}$. Other guidelines are ±1 or ±2. The formula for DFBETAS is

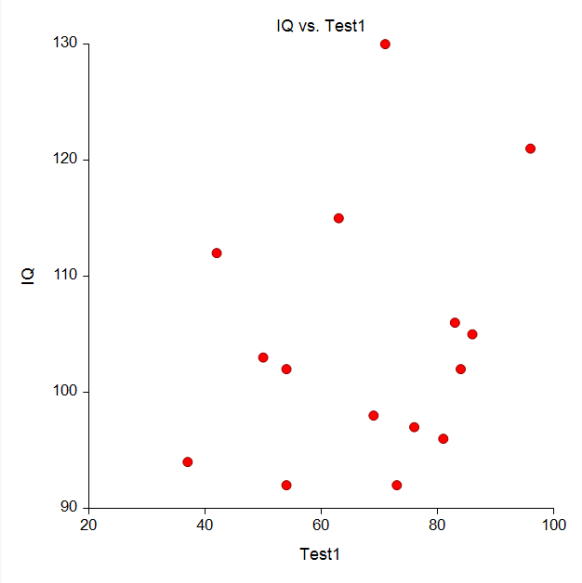
$$DFBetas_k = \frac{b_k - b_{k,-j}}{\sqrt{MSE_j c_{kk}}}$$

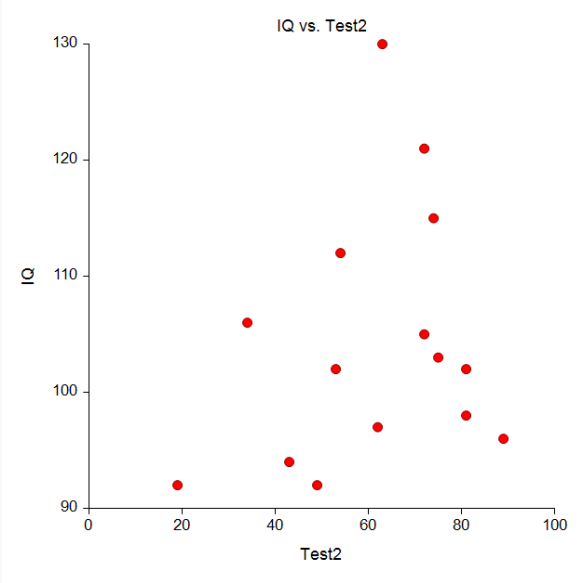
where $c_{kk}$ is the $k^{th}$ row and $k^{th}$ column element of the inverse matrix $(X'X)^{-1}$.

## Y vs X Plots

Actually, a regression analysis should always begin with a plot of *Y* versus each IV. These plots often show outliers, curvilinear relationships, and other anomalies.

**Y vs X Plots**




(3 more plots follow)

# Graphical Residual Analysis

The residuals can be graphically analyzed in numerous ways. Three types of residuals are graphically analyzed here: residuals, rstudent residuals, and partial residuals. For certain, the regression analyst should examine all of the basic residual graphs: the histogram, the density trace, the normal probability plot, the serial correlation plots, the scatter plot of the residuals versus the sequence of the observations, the scatter plot of the residuals versus the predicted value of the dependent variable, and the scatter plot of the residuals versus each of the independent variables.

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

## Point Cloud

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

## Wedge

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

## Bowtie

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

## Sloping Band

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

## Curved Band

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

## Curved Band with Increasing or Decreasing Variability

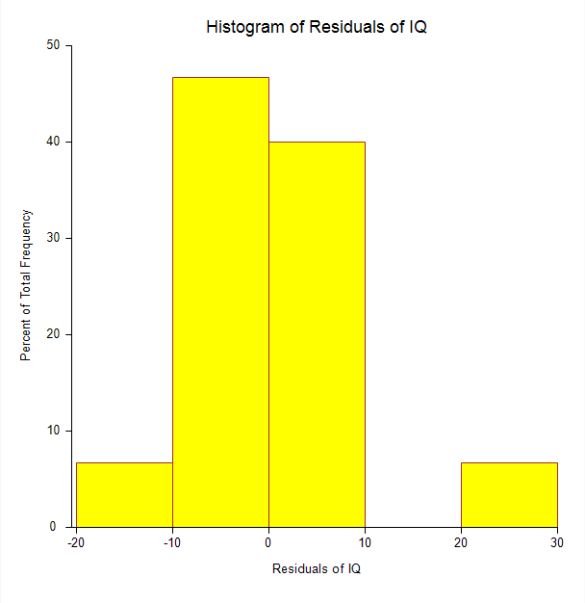
This residual plot is really a combination of the wedge and the curved band. It too must be avoided.

# Residual Distribution Plots

## Histogram

The purpose of the histogram is to evaluate whether the residuals are normally distributed. A dot plot can be added to the histogram that highlights the distribution of points in each bin of the histogram. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating normality of the residuals. The better choice would be the normal probability plot.

**Residual Distribution Plots**

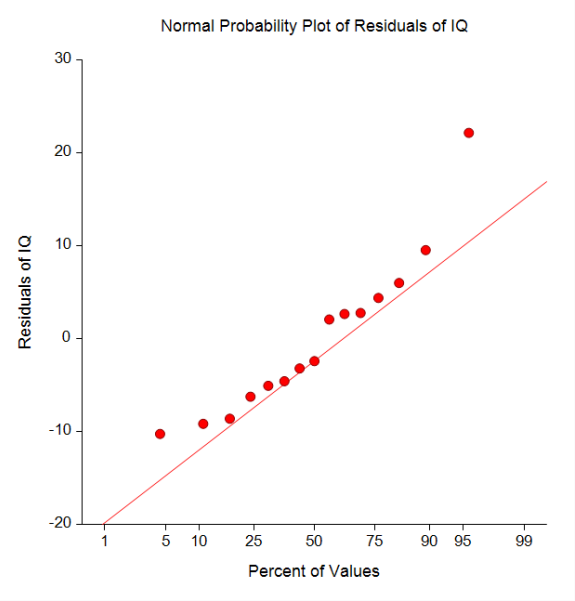


## Normal Probability Plot of Residuals

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.

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

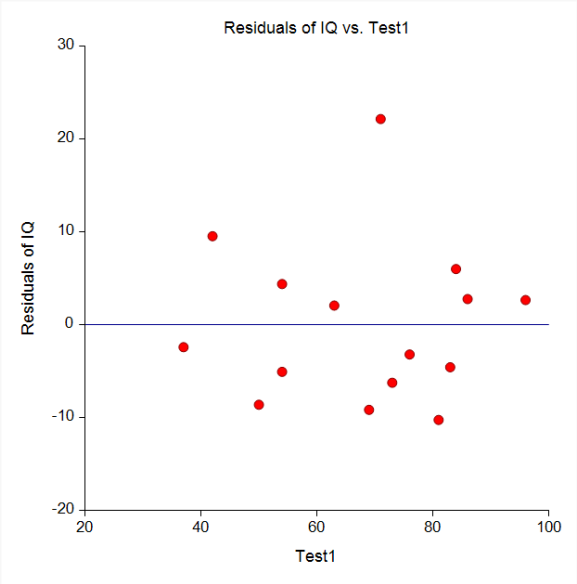
**Residual Distribution Plots**

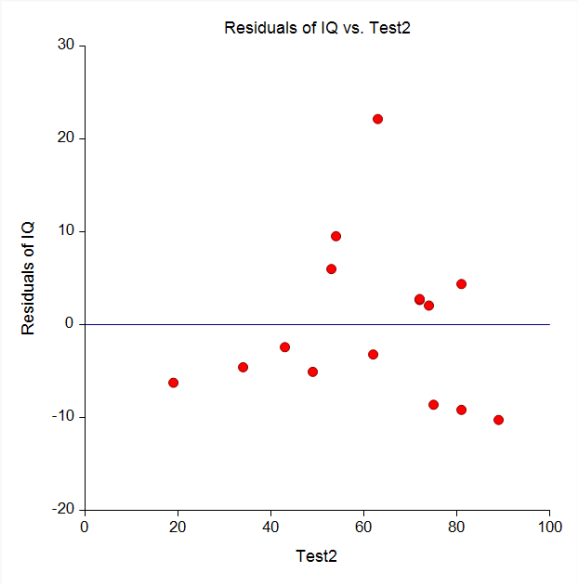


## Residuals vs X Plots

These are the scatter plots of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.

**Residuals vs X Plots**



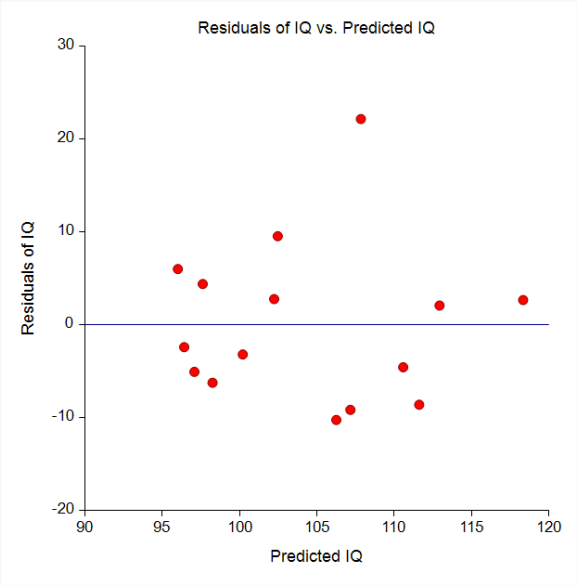




(3 more plots follow)

## Residuals vs Yhat (Predicted Y) Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of non-constant variance, a violation of a critical regression assumption. The sloping or curved band signifies inadequate specification of the model. The sloping band with increasing or decreasing variability suggests non-constant variance and inadequate specification of the model.

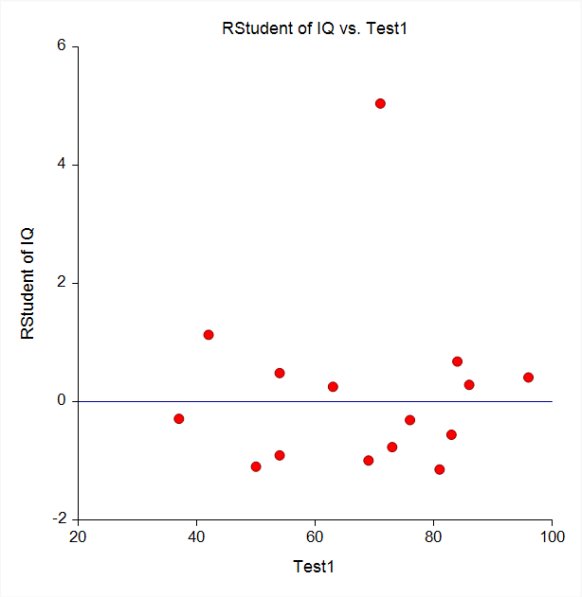
**Residuals vs Yhat (Predicted Y) Plot**

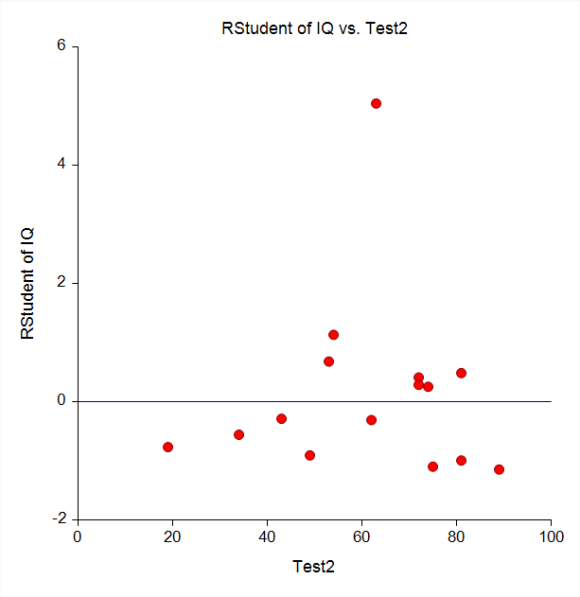


## RStudent vs X Plots

These are the scatter plots of the RStudent residuals versus each independent variable. The preferred pattern is a rectangular shape or point cloud. These plots are very helpful in visually identifying any outliers and nonlinear patterns.

**RStudent vs X Plots**



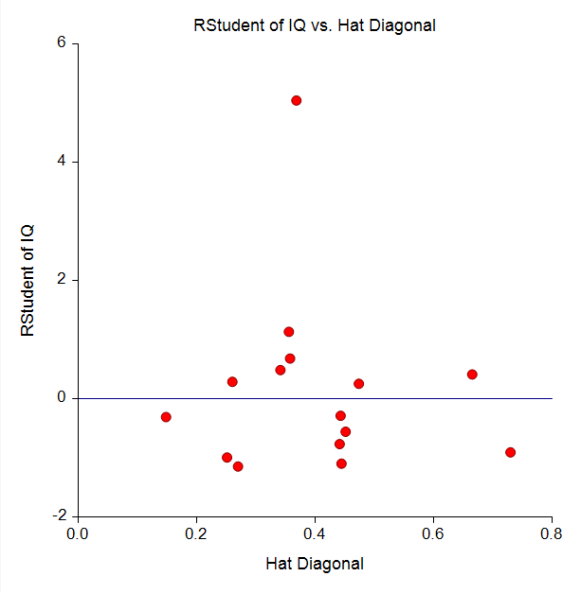

(3 more plots follow)

## RStudent vs Hat Diagonal Plot

In light of the earlier discussion in the Regression Diagnostics Section, Rstudent is one of the best single-case diagnostics for capturing large residuals, while the hat diagonal flags observations that are remote in the *X*-space. The purpose of this plot is to give a quick visual spotting of observations that are very different from the norm. It is best to rely on the actual regression diagnostics for any formal conclusions on influence. There are three influential realms you might be concerned with

1. Observations that are extreme along the rstudent (vertical) axis are outliers that need closer attention. They may have a major impact on the predictability of the model.
2. Observations that were extreme to the right (i.e., $h_{ii}>2p/n$) are outliers in the *X*-space. These kinds of observations could be data entry errors, so be sure the data is correct before proceeding.
3. Observations that are extreme on both axes are the most influential of all. Double-check these values.

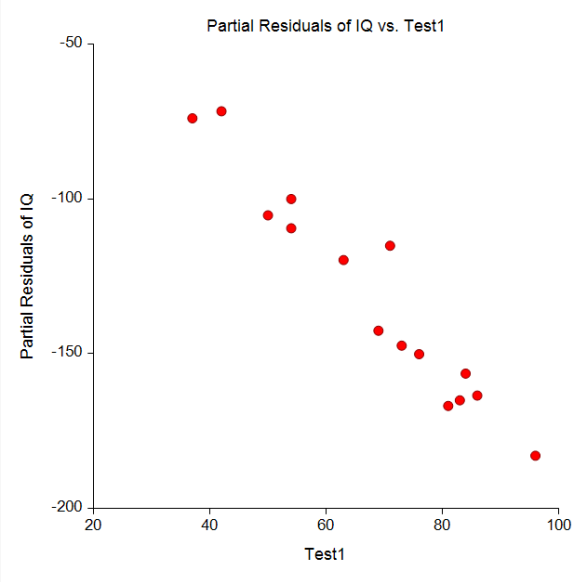
**RStudent vs Hat Diagonal Plot**

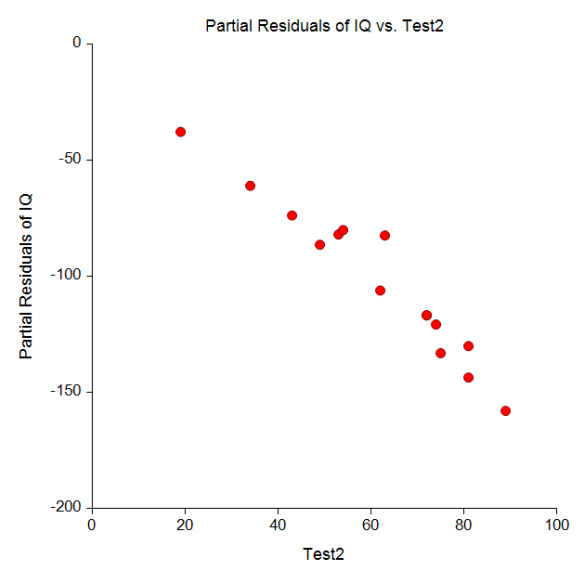


## Partial Residuals vs X's Plots

The scatter plot of the partial residuals against each independent variable allows you to examine the relationship between Y and each IV after the effects of the other IV's have been removed. These plots can be used to assess the extent and direction of linearity for each independent variable. In addition, they provide insight as to the correct transformation to apply and information on influential observations. One would like to see a linear pattern between the partial residuals and the independent variable.

**Partial Residuals vs X Plots**



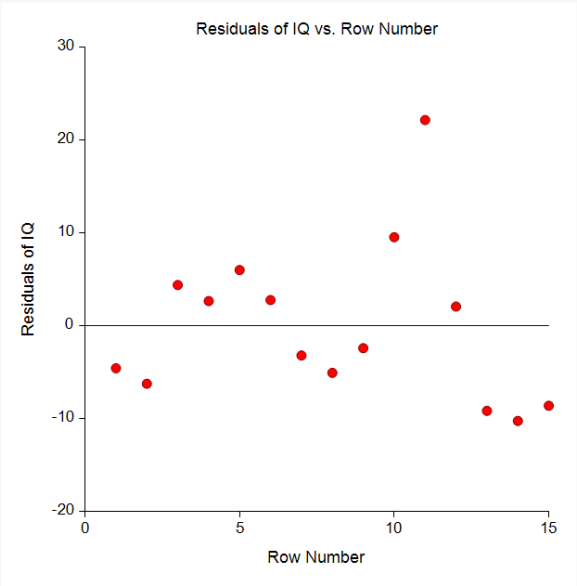




(3 more plots follow)

# Sequence Plot: Residuals vs Row Number

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.

**Sequence Plot: Residuals vs Row Number**

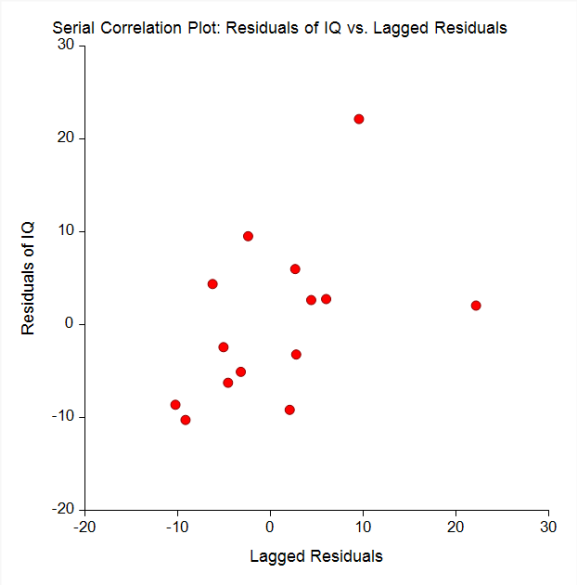


## Serial Correlation Plot: Residuals vs Lagged Residuals

This plot is only useful if your data represent a time series. This is a scatter plot of the $j^{th}$ residual versus the $j^{th}$-1 residual. The purpose of this plot is to check for first-order autocorrelation.

You would like to see a random pattern of these plotted residuals, i.e., a rectangular or uniform distribution. A strong positive or negative trend would indicate a need to redefine the model with some type of autocorrelation component. Positive autocorrelation or serial correlation means that the residual in time period *j* tends to have the same sign as the residual in time period (*j*-1). On the other hand, a strong negative autocorrelation means that the residual in time period *j* tends to have the opposite sign as the residual in time period (*j*-1). Be sure to check the Durbin-Watson statistic.

**Serial Correlation Plot: Residuals vs Lagged Residuals**



# Example 2 – Bootstrapping

This section presents an example of how to generate bootstrap confidence intervals with a multiple regression analysis. The tutorial will use the data are in the IQ dataset. This example will run a regression of IQ on Test1, Test2, and Test4.

## Setup

To run this example, complete the following steps:

**1 Open the IQ example dataset**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **IQ** and click **OK**.

**2 Specify the Multiple Regression procedure options**

- Find and open the **Multiple Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab

Y ..... **IQ**
Numeric X's ..... **Test1, Test2, Test4**

Reports Tab

Select Reports

Coefficient C.I.'s..... **Checked**
All Other Reports ..... **Unchecked**

Resampling

Calculate Bootstrap Confidence Intervals for..... **Checked**
Regression Estimates and Predicted Values
Random Seed..... **3278337** (for reproducibility)

Plots Tab

Bootstrap Distribution Plots

Bootstrap Histogram..... **Checked**

**3 Run the procedure**

- Click the **Run** button to perform the calculations and generate the output.

## Regression Coefficient Confidence Intervals

**Regression Coefficient Confidence Intervals**

| Independent Variable | Regression Coefficient b(i) | Standard Error Sb(i) | 95% Confidence Interval Limits for β(i) Lower | Upper |
|---|---|---|---|---|
| Intercept | 90.73267 | 12.82717 | 62.50026 | 118.9651 |
| Test1 | -1.965001 | 0.9406313 | -4.035316 | 0.1053149 |
| Test2 | -1.648506 | 0.7979558 | -3.404795 | 0.1077833 |
| Test4 | 3.789033 | 1.680081 | 0.09119987 | 7.486866 |

Note: The T-Value used to calculate the confidence interval limits was 2.201.

This report gives the confidence interval limits calculated under the assumption of normality. We have displayed it so that we can compare these to the bootstrap confidence intervals.

## Bootstrap Confidence Intervals

**Bootstrap Confidence Intervals for Regression Coefficient Estimates**

| Estimation Results | | Bootstrap Confidence Interval Limits | | |
|---|---|---|---|---|
| **Parameter** | **Estimate** | **Confidence Level** | **Lower** | **Upper** |
| **Intercept** | | | | |
| Original Value | 90.73267 | 90% | 68.82194 | 109.0556 |
| Bootstrap Mean | 91.8576 | 95% | 62.83783 | 112.3399 |
| Bias (BM - OV) | 1.124933 | 99% | 49.43907 | 124.4306 |
| Bias Corrected Value | 89.60773 | | | |
| Standard Error | 12.99079 | | | |
| **B(Test1)** | | | | |
| Original Value | -1.965001 | 90% | -3.063761 | -0.1985311 |
| Bootstrap Mean | -2.086367 | 95% | -3.313949 | 0.3396751 |
| Bias (BM - OV) | -0.121366 | 99% | -3.965512 | 1.432652 |
| Bias Corrected Value | -1.843635 | | | |
| Standard Error | 0.9170887 | | | |
| **B(Test2)** | | | | |
| Original Value | -1.648506 | 90% | -2.588346 | -0.01498824 |
| Bootstrap Mean | -1.786669 | 95% | -2.847273 | 0.3844686 |
| Bias (BM - OV) | -0.1381637 | 99% | -3.496344 | 1.812766 |
| Bias Corrected Value | -1.510342 | | | |
| Standard Error | 0.8231962 | | | |
| **B(Test4)** | | | | |
| Original Value | 3.789033 | 90% | 0.6039672 | 5.846 |
| Bootstrap Mean | 4.030579 | 95% | -0.3018321 | 6.370401 |
| Bias (BM - OV) | 0.2415457 | 99% | -2.654282 | 7.79888 |
| Bias Corrected Value | 3.547487 | | | |
| Standard Error | 1.681953 | | | |

Number of Bootstrap Samples = 3000, Sampling Method = Observation, Confidence Interval Method = Reflection, User-Entered Random Seed = 3278337.

**Bootstrap Confidence Intervals for Predicted Means**

| Estimation Results | | Bootstrap Confidence Interval Limits | | |
|---|---|---|---|---|
| **Parameter** | **Estimate** | **Confidence Level** | **Lower** | **Upper** |
| **Predicted Mean of IQ when Row = 16** | | | | |
| Original Value | 99.509 | 90% | 93.07262 | 105.0112 |
| Bootstrap Mean | 99.73473 | 95% | 91.1897 | 106.3544 |
| Bias (BM - OV) | 0.2257364 | 99% | 87.08662 | 109.9308 |
| Bias Corrected Value | 99.28326 | | | |
| Standard Error | 3.944446 | | | |
| **Predicted Mean of IQ when Row = 17** | | | | |
| Original Value | 101.2643 | 90% | 96.63694 | 105.4709 |
| Bootstrap Mean | 101.3102 | 95% | 95.72351 | 106.5511 |
| Bias (BM - OV) | 0.04589701 | 99% | 93.36579 | 108.5686 |
| Bias Corrected Value | 101.2184 | | | |
| Standard Error | 2.805879 | | | |

Number of Bootstrap Samples = 3000, Sampling Method = Observation, Confidence Interval Method = Reflection,
User-Entered Random Seed = 3278337.

**Bootstrap Confidence Intervals for Predicted Individuals**

| Estimation Results | | Bootstrap Confidence Interval Limits | | |
|---|---|---|---|---|
| **Parameter** | **Estimate** | **Confidence Level** | **Lower** | **Upper** |
| **Predicted IQ when Row = 16** | | | | |
| Original Value | 99.509 | 90% | 72.29645 | 122.2954 |
| Bootstrap Mean | 100.4496 | 95% | 66.69598 | 126.3246 |
| Bias (BM - OV) | 0.9405722 | 99% | 54.14439 | 136.7175 |
| Bias Corrected Value | 98.56843 | | | |
| Standard Error | 15.56534 | | | |
| **Predicted IQ when Row = 17** | | | | |
| Original Value | 101.2643 | 90% | 73.61021 | 123.7344 |
| Bootstrap Mean | 102.5657 | 95% | 68.27541 | 128.4504 |
| Bias (BM - OV) | 1.301438 | 99% | 54.5163 | 136.8572 |
| Bias Corrected Value | 99.96283 | | | |
| Standard Error | 15.61122 | | | |

Number of Bootstrap Samples = 3000, Sampling Method = Observation, Confidence Interval Method = Reflection,
User-Entered Random Seed = 3278337.

This report provides bootstrap intervals of the regression coefficients and predicted values for rows 16 and 17 which did not have an IQ (*Y*) value. Details of the bootstrap method were presented earlier in this chapter.

It is interesting to compare these confidence intervals with those provided in the Regression Coefficient report. The most striking difference is that the lower limit of the 95% bootstrap confidence interval for B(Test4) is now negative. When the lower limit is negative and the upper limit is positive, we know that a hypothesis test would not find the parameter significantly different from zero. Thus, while the regular confidence interval of B(Test4) indicates statistical significance (since both limits are positive), the bootstrap confidence interval does not.

### 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 Value

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.

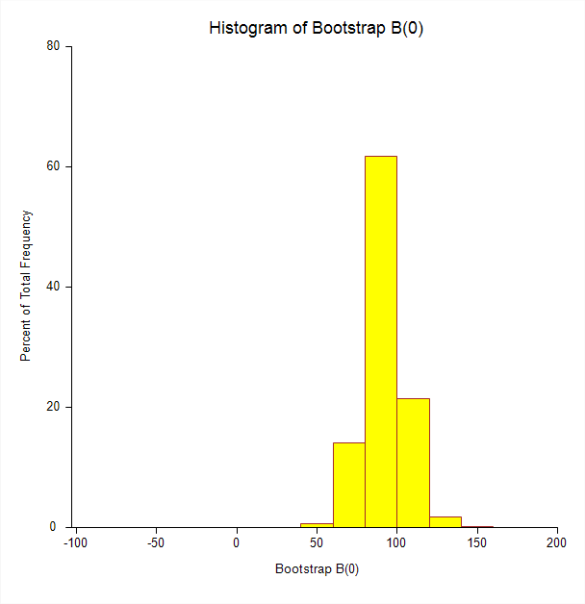
### Bootstrap Confidence Limits (Lower and Upper)

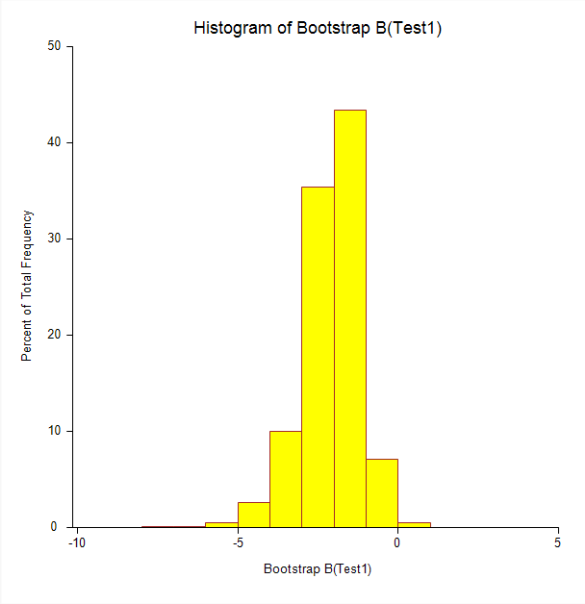
These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

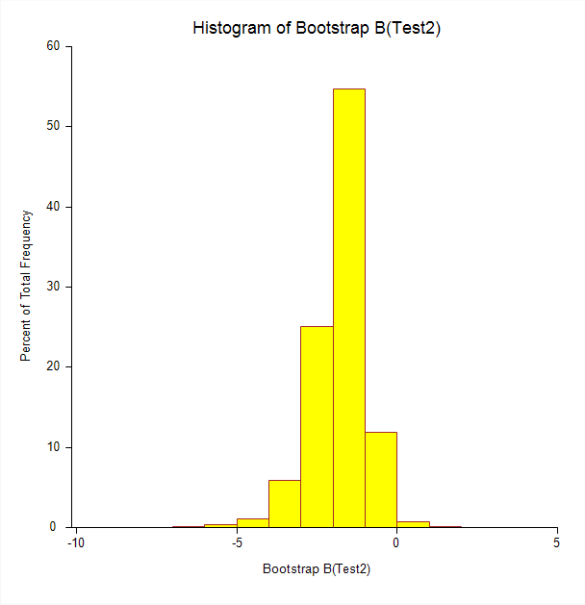
Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

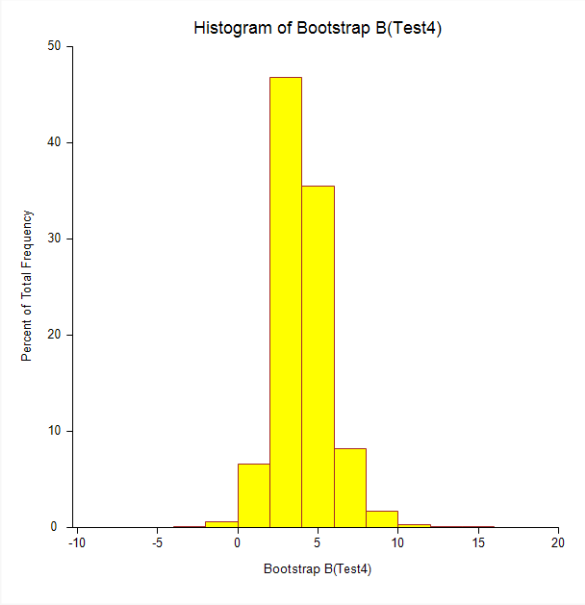
# Bootstrap Histograms

**Bootstrap Histograms of Regression Coefficient Estimates**

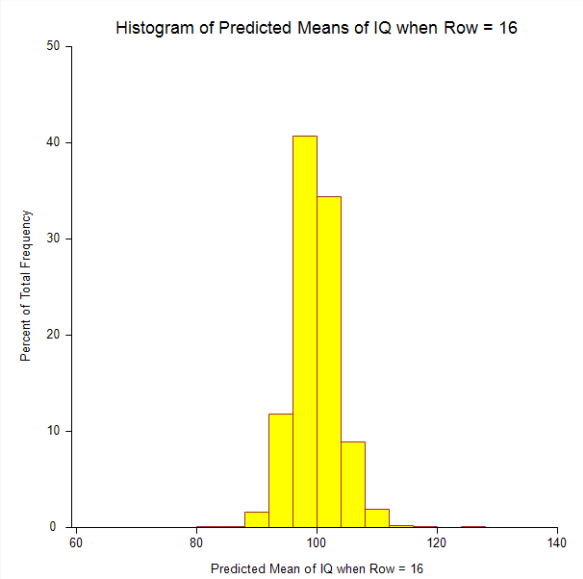


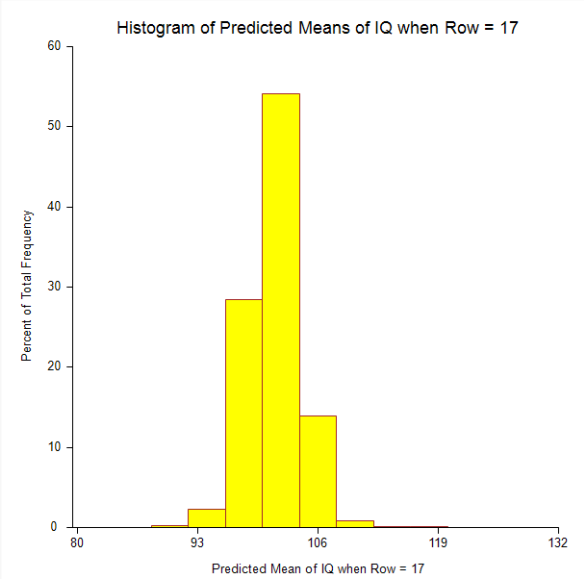



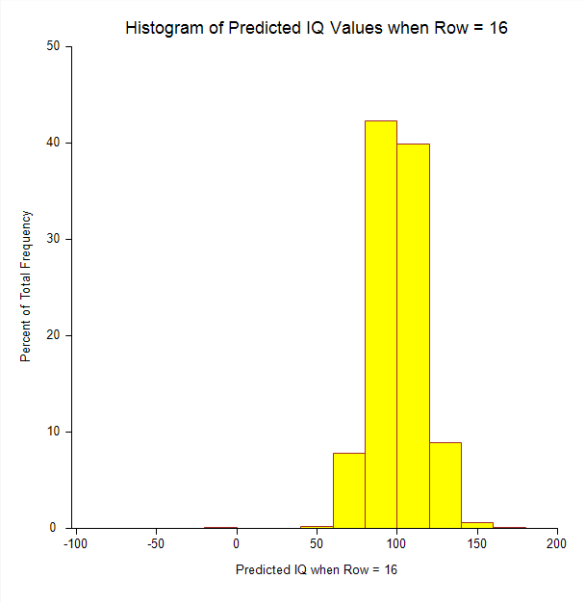



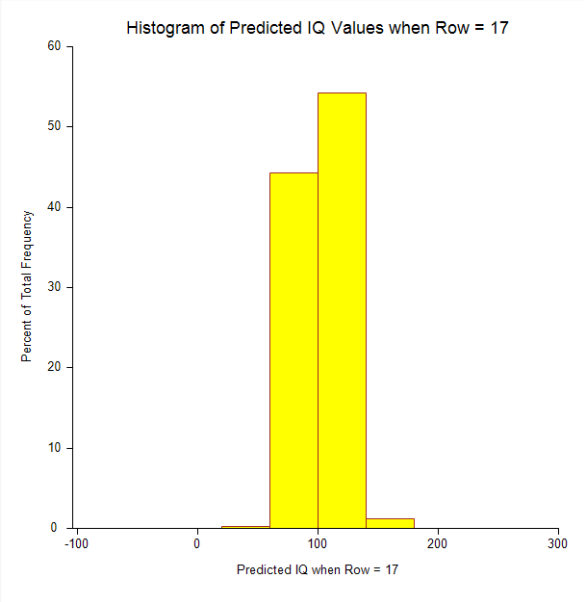
### Bootstrap Histograms of Predicted Means




### Bootstrap Histograms of Predicted Individuals




Each histogram shows the distribution of the corresponding parameter estimate.

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

# Example 3 – Checking the Parallel Slopes Assumption in Analysis of Covariance

An example of how to test the parallel slopes assumption is given in the General Linear Models chapter. Unfortunately, hand calculations and extensive data transformations are required to complete this test. This example will show you how to run this test without either transformations or hand calculations.

The ANCOVA dataset contains three variables: State, Age, and IQ. The researcher wants to test for IQ differences across the three states while controlling for each subject's age. An analysis of covariance should include a preliminary test of the assumption that the slopes between age and IQ are equal across the three states. Without parallel slopes, differences among mean state IQ's depend on age.

It turns out that a test for parallel slopes is a test for an Age by State interaction. All that needs to be done is to include this term in the model and the appropriate test will be generated.

## Setup

To run this example, complete the following steps:

**1 Open the ANCOVA example dataset**

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **ANCOVA** and click **OK**.

**2 Specify the Multiple Regression procedure options**

- Find and open the **Multiple Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

| Variables, Model Tab | |
|---|---|
| Y | **IQ** |
| Numeric X's | **Age** |
| Categorical X's | **State** |
| Terms | **Interaction Model** |

| Reports Tab | |
|---|---|
| ANOVA Detail | **Checked** |
| All Other Reports | **Unchecked** |

**3 Run the procedure**

- Click the **Run** button to perform the calculations and generate the output.

## Analysis of Variance Detail

**Analysis of Variance Detail**

| Source | DF | $R^2$ Lost If Term(s) Removed | Sum of Squares | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 1 | | 313345.2 | 313345.2 | | |
| Model | 5 | 0.2438 | 80.15984 | 16.03197 | 1.547 | 0.2128 |
| Age | 1 | 0.1219 | 40.07431 | 40.07431 | 3.868 | 0.0609 |
| State | 2 | 0.1417 | 46.57466 | 23.28733 | 2.248 | 0.1274 |
| Age*State | 2 | 0.1178 | 38.72052 | 19.36026 | 1.869 | 0.1761 |
| Error | 24 | 0.7562 | 248.6402 | 10.36001 | | |
| Total (Adjusted) | 29 | | 328.8 | 11.33793 | | |

The F-Value for the Age*State interaction term is 1.869. This matches the result that was obtained by hand calculations in the General Linear Model example. Since the probability level of 0.1761 is not significant, we cannot reject the assumption that the three slopes are equal.