Chapter 309

Multiple Regression (Old Version)

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 ε_j is the error (residual) of observation *j*.

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

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

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

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 ε_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 (β'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

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

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

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

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

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

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

Step 3 – Setup and Run the Regression

Introduction

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

Enter Variables

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

Choose Report Options

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

Specify Alpha

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

Select All Plots

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

Step 4 – Check Model Adequacy

Introduction

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

Check 1. Linearity

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

Check 2. Normality

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

Check 3. Nonconstant Variance

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

Check 4. Independence or Serial Correlation

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

Check 5. Outliers

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

Check 6. Multicollinearity

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

Check 7. Predictability

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

Step 5 – Record Your Results

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

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

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

Multiple Regression Technical Details

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

Define the following vectors and matrices:

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

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

Least Squares

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

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

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

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

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

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

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

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

Hypothesis Tests of the Intercept and Slopes

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

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

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

Confidence Intervals of the Intercept and Slope

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

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

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

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

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

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

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 σ^2 will be obtained. Some statisticians refer to these as the *studentized deleted residuals*.

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

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

DFFITS

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

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

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

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

Cook's D

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

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

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

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

This approximation is exact when no weight variable is used.

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

CovRatio

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

The general formula for the CovRatio is

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

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

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

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

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

DFBETAS

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

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

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

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.

Subset Selection

Subset selection refers to the task of finding a small subset of the available independent variables that does a good job of predicting the dependent variable. Exhaustive searches are possible for regressions with up to 15 IV's. However, when more than 15 IV's are available, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

An issue that comes up because of categorical IV's is what to do with the individual-degree of freedom variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms search on model terms rather than on the individual binary variables. Thus, the whole set of generated variables associated with a given term are considered together for inclusion in, or deletion from, the model. It's all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can save the generated set of variables in the first run and designate them as Numeric Variables.

Hierarchical Models

Another issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to consider only hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of *R*-Squared. Enter this term into the model.
3. Continue adding terms until a target value for *R*-Squared is achieved or until a preset limit on the maximum number of terms in the model is reached. Note that these terms can be limited to those keeping the model hierarchical.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations and terms so that other, more time consuming, methods are not feasible.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of *R*-Squared. If a switch can be found, it is made and the pool of terms is again searched to determine if another switch can be made. Note that this switching can be limited to those keeping the model hierarchical.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run. This two-step procedure works better than relying on some *F*-to-enter and *F*-to-remove tests whose properties are not well understood to begin with.

Robust Regression

Regular multiple regression is optimum when all of its assumptions are valid. When some of these assumptions are invalid, least squares regression can perform poorly. Thorough residual analysis can point to these assumption breakdowns and allow you to work around these limitations. However, this residual analysis is time consuming and requires a great deal of training.

Robust regression provides an alternative to least squares regression that works with less restrictive assumptions. Specifically, it provides much better regression coefficient estimates when outliers are present in the data. Outliers violate the assumption of normally distributed residuals in least squares regression. They tend to pull the least squares fit too much in their direction by receiving much more "weight" than they deserve. Typically, you would expect that the weight attached to each observation would be about 1/*N* in a dataset with *N* observations. However, these outlying observations may receive a weight of 10, 20, or even 50 %. This leads to serious distortions in the estimated regression coefficients.

Because of this distortion, these outliers are difficult to identify since their residuals are much smaller than they should be. When only one or two independent variables are used, these outlying points may be visually detected in various scatter plots. However, the complexity added by additional independent variables hides the outliers from view in these scatter plots. Robust regression down-weights the influence of outliers. This makes their residuals larger and easier to spot. Robust regression techniques are iterative procedures that seek to identify these outliers and minimize their impact on the coefficient estimates.

The amount of weighting assigned to each observation in robust regression is controlled by a special curve called an *influence function*. There are three influence functions available in **NCSS**.

Although robust regression can particularly benefit untrained users, careful consideration should be given to the results. Essentially, robust regression conducts its own residual analysis and down-weights or completely removes various observations. You should study the weights that are assigned to each observation, determine which have been largely eliminated, and decide if you want these observations in your analysis.

M-Estimators

Several families of robust estimators have been developed. The robust methods found in **NCSS** fall into the family of *M-estimators*. This estimator minimizes the sum of a function *ρ(·)* of the residuals. That is, these estimators are defined as the *β's* that minimize

$$\min_{\beta}\sum_{j=1}^{n}\rho\left(y_j - x_j'\beta\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(e_j\right)$$

M in *M*-estimators stands for maximum likelihood since the function *ρ(·)* is related to the likelihood function for a suitable choice of the distribution of the residuals. In fact, when the residuals follow the normal distribution, setting $\rho(u) = \frac{1}{2}u^2$ results in the usual method of least squares.

Unfortunately, *M*-estimators are not necessarily *scale invariant*. That is, these estimators may be influenced by the scale of the residuals. A scale-invariant estimator is found by solving

$$\min_{\beta}\sum_{j=1}^{N}\rho\left(\frac{y_j - x_j'\beta}{s}\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(\frac{e_j}{s}\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(u_j\right)$$

where *s* is a robust estimate of scale. The estimate of *s* is used in **NCSS** is

$$s = \frac{median\left|e_j - median(e_j)\right|}{0.6745}$$

This estimate of *s* yields an approximately unbiased estimator of the standard deviation of the residuals when *N* is large, and the error distribution is normal.

The function

$$\sum_{j=1}^{N} \rho\left(\frac{y_j - x_j'\beta}{s}\right)$$

is minimized by setting the first partial derivatives of *ρ(·)* with respect to each β_i to zero which forms a set of *p* + 1 nonlinear equations

$$\sum_{j=1}^{N} x_{ij}\,\psi\left(\frac{y_j - x_j'\beta}{s}\right) = 0, \quad i = 0, 1, \dots, p$$

where $\psi(u) = \rho'(u)$ is the *influence function*.

These equations are solved iteratively using an approximate technique called iteratively reweighted least squares (IRLS). At each step, new estimates of the regression coefficients are found using the matrix equation

$$\beta_{t+1} = (\mathbf{X}'\mathbf{W}_t\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_t\mathbf{Y}$$

where $\mathbf{W}_t$ is an *N-by-N* diagonal matrix of weights $w_{1t}, w_{2t}, \dots, w_{Nt}$ defined as

$$w_{jt} = \begin{cases} \dfrac{\psi\left[\left(y_j - x'\beta_{jt}\right)/s_t\right]}{\left(y_j - x'\beta_{jt}\right)/s_t} & \text{if } y_j \neq x'\beta_{jt} \\ 1 & \text{if } y_j = x'\beta_{jt} \end{cases}$$

The ordinary least squares regression coefficients are used at the first iteration to begin the iteration process. Iterations are continued until there is little or no change in the regression coefficients from one iteration to the next. Because of the masking nature of outliers, it is a good idea to run through at least five iterations to allow the outliers to be found.

Three functions are available in **NCSS.** These are Andrew's Sine, Huber's method, and Tukey's biweight. Huber's method is currently the most frequently recommended in the regression texts that we have seen. The specifics for each of these functions are as follows.

Andrew's Sine

$$\rho(u) = \begin{cases} c[1 - cos(u/c)] & \text{if } |u| < \pi c \\ 2c & \text{if } |u| \geq \pi c \end{cases}$$

$$\psi(u) = \begin{cases} \sin(u/c) & \text{if } |u| < \pi c \\ 0 & \text{if } |u| \geq \pi c \end{cases}$$

$$w(u) = \begin{cases} \dfrac{\sin(u/c)}{u/c} & \text{if } |u| < \pi c \\ 0 & \text{if } |u| \geq \pi c \end{cases}$$

$$c = 1.339$$

Huber's Method

$$\rho(u) = \begin{cases} u^2 & \text{if } |u| < c \\ |2u|c - c^2 & \text{if } |u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u & \text{if } |u| < c \\ c \text{ sign}(u) & \text{if } |u| \geq c \end{cases}$$

$$w(u) = \begin{cases} 1 & \text{if } |u| < c \\ c/|u| & \text{if } |u| \geq c \end{cases}$$

$$c = 1.345$$

Tukey's Biweight

$$\rho(u) = \begin{cases} \dfrac{c^2}{3}\left\{1 - \left[1 - \left(\dfrac{u}{c}\right)^2\right]^3\right\} & \text{if } |u| < c \\ 2c & \text{if } |u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u\left[1 - \left(\dfrac{u}{c}\right)^2\right]^2 & \text{if } |u| < c \\ 0 & \text{if } |u| \geq c \end{cases}$$

$$w(u) = \begin{cases} \left[1 - \left(\dfrac{u}{c}\right)^2\right]^2 & \text{if } |u| < c \\ 0 & \text{if } |u| \geq c \end{cases}$$

$$c = 4.685$$

This gives you a sketch of what robust regression is about. If you find yourself using the technique often, we suggest that you study one of the modern texts on regression analysis. All of these texts have chapters on robust regression. A good introductory discussion of robust regression is found in Hamilton (1991). A more thorough discussion is found in Montgomery and Peck (1992).

Data Structure

The data are entered in two or more 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 (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** 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 Tab

Y Dependent Variable(s) ....**IQ**
X's Numeric Independent Variables................**Test1-Test5**

Reports Tab

Select a Group of Reports and Plots ....**Display ALL reports & plots**

3 Run the procedure

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

Run Summary Section

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	IQ	Rows Processed	17
Number Ind. Variables	5	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	0
R2	0.3991	Rows with Weight Missing	0
Adj R2	0.0652	Rows with Y Missing	2
Coefficient of Variation	0.1021	Rows Used in Estimation	15
Mean Square Error	113.4648	Sum of Weights	15.000
Square Root of MSE	10.65198	Completion Status	Normal Completion
Ave Abs Pct Error	6.218		

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

R-Squared

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

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

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

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

Adjusted R-Squared

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

Coefficient of Variation

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

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

Ave Abs Pct Error

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

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

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

Descriptive Statistics Section

Descriptive Statistics Section

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 Section

Correlation Matrix Section

	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 Section

Regression Coefficient T-Tests Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0: β(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	85.2404	23.6951	3.597	0.0058	Yes	0.8915
Test1	-1.9336	1.0291	-1.879	0.0930	No	0.3896
Test2	-1.6599	0.8729	-1.902	0.0897	No	0.3974
Test3	0.1050	0.2199	0.477	0.6445	No	0.0713
Test4	3.7784	1.8345	2.060	0.0695	No	0.4522
Test5	-0.0406	0.2012	-0.202	0.8447	No	0.0538

Estimated Model

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

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

Independent Variable

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

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

Regression Coefficient

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

Standard Error

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

T-Value to test Ho: B(i)=0

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

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

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

Prob Level

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

Reject H0 at 5%?

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

Note that the level of significance is specified in the Alpha of C.I.'s and Tests box on the Format tab panel.

Power (5%)

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

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

Estimated Model

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

Note that a transformation must be less than 255 characters. Since these formulas are often greater than 255 characters in length, you must use the FILE(filename) transformation. To do so, copy the formula to a text file using Notepad, Windows Write, or Word to receive the model text. Be sure to save the file as an unformatted text (ASCII) file. The transformation is FILE(filename) where *filename* is the name of the text file, including directory information. When the transformation is executed, it will load the file and use the transformation stored there.

Regression Coefficient Confidence Intervals Section

Regression Coefficient Confidence Intervals Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper95% Conf. Limit of β(i)	Standardized Coefficient
Intercept	85.2404	23.6951	31.6383	138.8425	0.0000
Test1	-1.9336	1.0291	-4.2615	0.3944	-3.0524
Test2	-1.6599	0.8729	-3.6345	0.3147	-2.9224
Test3	0.1050	0.2199	-0.3925	0.6024	0.1404
Test4	3.7784	1.8345	-0.3715	7.9283	4.7853
Test5	-0.0406	0.2012	-0.4958	0.4146	-0.0595

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

Independent Variable

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

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

Regression Coefficient

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

Standard Error

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

Lower - Upper 95% C.L.

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

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

The formulas for the lower and upper confidence limits are:

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

Standardized Coefficient

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

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

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

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

Note: The T-Value ...

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

Analysis of Variance Section

Analysis of Variance Section

Source	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		163281.7	163281.7			
Model	5	0.3991	678.1504	135.6301	1.195	0.3835	0.2565
Error	9	0.6009	1021.183	113.4648			
Total(Adjusted)	14	1.0000	1699.333	121.381			

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

Source

This represents a partition of the variation in Y.

R2

This is the overall R^2 of this the regression model.

DF

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

Sum of Squares

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

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

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

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

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

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.

Prob Level

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

Power(5%)

Power is the probability of rejecting the null hypothesis that all the regression coefficients are zero when at least one is not.

Analysis of Variance Detail Section

Analysis of Variance Detail Section

Model Term	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		163281.7	163281.7			
Model	5	0.3991	678.1504	135.6301	1.195	0.3835	0.2565
Test1	1	0.2357	400.562	400.562	3.530	0.0930	0.3896
Test2	1	0.2414	410.2892	410.2892	3.616	0.0897	0.3974
Test3	1	0.0152	25.8466	25.8466	0.228	0.6445	0.0713
Test4	1	0.2832	481.3241	481.3241	4.242	0.0695	0.4522
Test5	1	0.0027	4.614109	4.614109	0.041	0.8447	0.0538
Error	9	0.6009	1021.183	113.4648			
Total(Adjusted)	14	1.0000	1699.333	121.381			

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

Model Term

This is the term from the design model.

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

DF

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

R2

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

Sum of Squares

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

Mean Square

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

F-Ratio

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

Prob Level

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

Power(5%)

Power is the probability of rejecting the null hypothesis that all the regression coefficients associated with this term are zero, assuming that the estimated values of these coefficients are their true values.

PRESS Section

PRESS Section

Parameter	From PRESS Residuals	From Regular Residuals
Sum of Squared Residuals	2839.941	1021.183
Sum of |Residuals|	169.6438	99.12155
R2	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 R2

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

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

Normality Tests Section

Normality Tests Section

Test Name	Test Value	Prob Level	Reject H0 At Alpha = 20%?
Shapiro Wilk	0.9076	0.124280	Yes
Anderson Darling	0.4581	0.263931	No
D'Agostino Skewness	2.0329	0.042064	Yes
D'Agostino Kurtosis	1.5798	0.114144	Yes
D'Agostino Omnibus	6.6285	0.036361	Yes

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

Serial-Correlation and Durbin-Watson Test

Serial Correlation of Residuals Section

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 0.516398.

Durbin-Watson Test For Serial Correlation

Parameter	Value	Did the Test Reject H0: Rho(1) = 0?
Durbin-Watson Value	1.0010	
Prob. Level: Positive Serial Correlation	0.0072	Yes
Prob. Level: Negative Serial Correlation	0.9549	No

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

Lag and Correlation

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

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

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.*

Durbin-Watson Value

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

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

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.

R-Squared Section

R-Squared Section

Independent Variable	Total R2 for This I.V. And Those Above	R2 Increase When This I.V. Added To Those Above	R2 Decrease When This I.V. Is Removed	R2 When This I.V. Is Fit Alone	Partial R2 Adjusted For All Other I.V.'s
Test1	0.0509	0.0509	0.2357	0.0509	0.2817
Test2	0.0990	0.0480	0.2414	0.0579	0.2866
Test3	0.1131	0.0142	0.0152	0.0055	0.0247
Test4	0.3964	0.2832	0.2832	0.1379	0.3203
Test5	0.3991	0.0027	0.0027	0.0034	0.0045

R^2 reflects the percent of variation in *Y* explained by the independent variables in the model. A value of R^2 near zero indicates a complete lack of fit between *Y* and the *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 R2 for This I.V. and Those Above

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

R2 Increase When This IV Added to Those Above

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

R2 Decrease When This IV is Removed

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

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

R2 When This IV Is Fit Alone

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

Partial R2 Adjusted For All Other IV's

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

Variable Omission Section

Variable Omission Section

Independent Variable	R2 When I.V. Omitted	MSE When I.V. Omitted	Mallow's Cp When I.V. Omitted	H0: B=0 Prob Level	R2 Of Regress. Of This I.V. On Other I.V.'s
Full Model	0.3991	113.4648			
Test1	0.1634	142.1745	7.5303	0.0930	0.9747
Test2	0.1576	143.1472	7.6160	0.0897	0.9717
Test3	0.3839	104.703	4.2278	0.6445	0.2280
Test4	0.1158	150.2507	8.2421	0.0695	0.9876
Test5	0.3964	102.5797	4.0407	0.8447	0.2329

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

Independent Variable

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

R2 When IV Omitted

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

MSE When IV Omitted

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

Mallow's Cp When IV Omitted

Another criterion for variable selection and importance is Mallow's *Cp* statistic. The optimum model will have a *Cp* value close to *p*+1, where *p* is the number of independent variables. A *Cp* greater than (*p*+1) indicates that the regression model is 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

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

H0: B=0 Prob Level

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

R2 Of Regress. Of This IV Other X's

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

Sum of Squares and Correlation Section

Sum of Squares and Correlation Section

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

This section provides the sum of squares and correlations equivalent to the *R-Squared Section*.

Independent Variable

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

Sequential Sum Squares

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

Incremental Sum Squares

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

Last Sum Squares

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

Simple Correlation

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

Partial Correlation

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

Sequential Models Section

Sequential Models Section

Independent Variable	Included R2	Omitted R2	Included F-Ratio	Included Prob>F	Omitted F-Ratio	Omitted Prob>F
Test1	0.0509	0.3482	0.697	0.4187	1.304	0.3390
Test2	0.0990	0.3001	0.659	0.5351	1.498	0.2801
Test3	0.1131	0.2859	0.468	0.7107	2.141	0.1735
Test4	0.3964	0.0027	1.641	0.2390	0.041	0.8447
Test5	0.3991	0.0000	1.195	0.3835		

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

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

Independent Variable

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

Included R2

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

Omitted R2

This is the R^2 for the full model minus the *Included* 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.

Included F-ratio

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

Included Prob>F

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

Omitted F-Ratio

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

Omitted Prob>F

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

Multicollinearity Section

Multicollinearity Section

Independent Variable	Variance Inflation Factor	R2 Versus Other I.V.'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

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

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

R2 Versus Other IV's

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

Tolerance

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

Diagonal of X'X Inverse

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

Eigenvalues of Centered Correlations Section

Eigenvalues of Centered Correlations

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Number
1	2.2150	44.299	44.299	1.000
2	1.2277	24.554	68.853	1.804
3	1.1062	22.124	90.978	2.002
4	0.4446	8.892	99.870	4.982
5	0.0065	0.130	100.000	340.939

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

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

Eigenvalue

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

Incremental Percent

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

Cumulative Percent

This is the running total of the Incremental Percent.

Condition Number

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

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

Eigenvector Percent of Regression-Coefficent-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	96.606	96.606	1.000
2	0.1041	1.735	98.340	55.686
3	0.0670	1.116	99.457	86.532
4	0.0214	0.357	99.814	270.830
5	0.0109	0.181	99.995	533.756
6	0.0003	0.005	100.000	17767.041

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

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

Eigenvalue

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

Incremental Percent

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

Cumulative Percent

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

Condition Number

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

Eigenvector Percent of Regression-Coefficent-Variance using Uncentered Correlations

Eigenvector Percent of Regression-Coefficent-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 of Means

Predicted Values with Confidence Limits of Means

Row	Actual IQ	Predicted IQ	Standard Error of Predicted	95% Lower Conf. Limit of Mean	95% Upper Conf. Limit of Mean
1	106.000	110.581	7.157	94.391	126.770
2	92.000	98.248	7.076	82.242	114.255
3	102.000	97.616	6.223	83.539	111.693
4	121.000	118.340	8.687	98.689	137.990
5	102.000	96.006	6.369	81.597	110.414
6	105.000	102.233	5.433	89.942	114.523
7	97.000	100.204	4.100	90.930	109.479
8	92.000	97.073	9.099	76.490	117.657
9	94.000	96.414	7.089	80.379	112.450
10	112.000	102.467	6.352	88.098	116.835
11	130.000	107.846	6.464	93.223	122.468
12	115.000	112.933	7.331	96.349	129.517
13	98.000	107.167	5.339	95.090	119.244
14	96.000	106.255	5.532	93.741	118.769
15	103.000	111.618	7.100	95.556	127.679
16		97.705	7.031	81.800	113.611
17		100.198	4.305	90.459	109.938

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

Actual

This is the actual value of *Y*.

Predicted

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

Standard Error of Predicted

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

Lower 95% C.L. of Mean

This is the lower limit of a 95% confidence interval estimate of the mean of *Y* for this observation.

Upper 95% C.L. of Mean

This is the upper limit of a 95% confidence interval estimate of the mean of *Y* for this observation. Note that you set the alpha level.

Predicted Values with Prediction Limits of Individuals

Predicted Values with Prediction Limits of Individuals

Row	Actual IQ	Predicted IQ	Standard Error of Predicted	95% Lower Pred. Limit of Individual	95% Upper Pred. Limit of Individual
1	106.000	110.581	12.833	81.551	139.611
2	92.000	98.248	12.788	69.320	127.177
3	102.000	97.616	12.336	69.709	125.523
4	121.000	118.340	13.745	87.247	149.433
5	102.000	96.006	12.411	67.930	124.081
6	105.000	102.233	11.958	75.183	129.283
7	97.000	100.204	11.414	74.385	126.024
8	92.000	97.073	14.009	65.383	128.764
9	94.000	96.414	12.795	67.470	125.359
10	112.000	102.467	12.402	74.411	130.522
11	130.000	107.846	12.460	79.659	136.032
12	115.000	112.933	12.931	83.681	142.185
13	98.000	107.167	11.915	80.213	134.120
14	96.000	106.255	12.003	79.103	133.407
15	103.000	111.618	12.801	82.659	140.576
16		97.705	12.763	68.833	126.578
17		100.198	11.489	74.208	126.189

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

Actual

This is the actual value of *Y*.

Predicted

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

Standard Error of Predicted

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

Lower 95% Pred. Limit of Individual

This is the lower limit of a 95% prediction interval of the individual value of *Y* for the values of the IV's for this observation.

Upper 95% Pred. Limit of Individual

This is the upper limit of a 95% prediction interval of the individual value of *Y* for the values of the IV's for this observation. Note that you set the alpha level.

Residual Report

Residual Report

Row	Actual IQ	Predicted IQ	Residual	Absolute Percent Error	Sqrt(MSE) Without This Row
1	106.000	110.581	-4.581	4.322	11.085
2	92.000	98.248	-6.248	6.792	10.905
3	102.000	97.616	4.384	4.298	11.136
4	121.000	118.340	2.660	2.199	11.181
5	102.000	96.006	5.994	5.877	10.984
6	105.000	102.233	2.767	2.635	11.241
7	97.000	100.204	-3.204	3.304	11.231
8	92.000	97.073	-5.073	5.515	10.759
9	94.000	96.414	-2.414	2.568	11.240
10	112.000	102.467	9.533	8.512	10.489
11	130.000	107.846	22.154	17.042	5.526
12	115.000	112.933	2.067	1.797	11.253
13	98.000	107.167	-9.167	9.354	10.659
14	96.000	106.255	-10.255	10.682	10.471
15	103.000	111.618	-8.618	8.367	10.533
16		97.705			
17		100.198			

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

Actual

This is the actual value of *Y*.

Predicted

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

Residual

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

Absolute Percent Error

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

Sqrt(MSE) Without This Row

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

Regression Diagnostics Section

Regression Diagnostics Section

Row	Standardized Residual	RStudent	Hat Diagonal	Cook's D	Dffits	CovRatio
1	-0.5806	-0.5579	0.4514	0.0462	-0.5061	2.9388
2	-0.7847	-0.7665	0.4413	0.0811	-0.6812	2.3714
3	0.5071	0.4851	0.3413	0.0222	0.3492	2.5863
4	0.4315	0.4111	0.6650	0.0616	0.5792	5.3387
5	0.7021	0.6808	0.3575	0.0457	0.5079	2.2506
6	0.3020	0.2862	0.2601	0.0053	0.1697	2.5777
7	-0.3259	-0.3091	0.1481	0.0031	-0.1289	2.2162
8	-0.9161	-0.9070	0.7297	0.3775	-1.4901	4.1684
9	-0.3037	-0.2878	0.4429	0.0122	-0.2566	3.4207
10	1.1149	1.1322	0.3556	0.1143	0.8410	1.2896
11	2.6167	5.0444	0.3683	0.6652	3.8514	0.0006
12	0.2675	0.2532	0.4737	0.0107	0.2402	3.6717
13	-0.9945	-0.9938	0.2512	0.0553	-0.5756	1.3465
14	-1.1265	-1.1460	0.2697	0.0781	-0.6964	1.1151
15	-1.0853	-1.0975	0.4443	0.1569	-0.9814	1.5725
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 σ^2 will be obtained. Some statisticians refer to these as the *studentized deleted residuals*.

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

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

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

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

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

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

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

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

DFBETAS Section

DFBETAS Section

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 $(X'X)^{-1}$.

Graphic Residual Analysis

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

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

Point Cloud

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

Wedge

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

Bowtie

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

Sloping Band

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

Curved Band

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

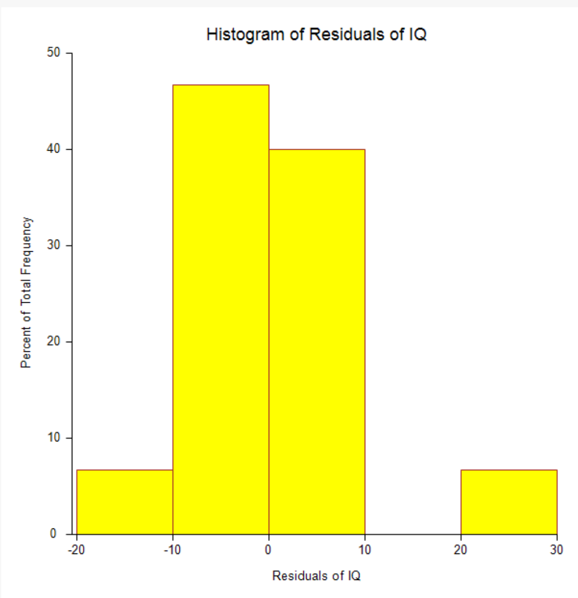
Curved Band with Increasing or Decreasing Variability

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

Histogram

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

Plots Section

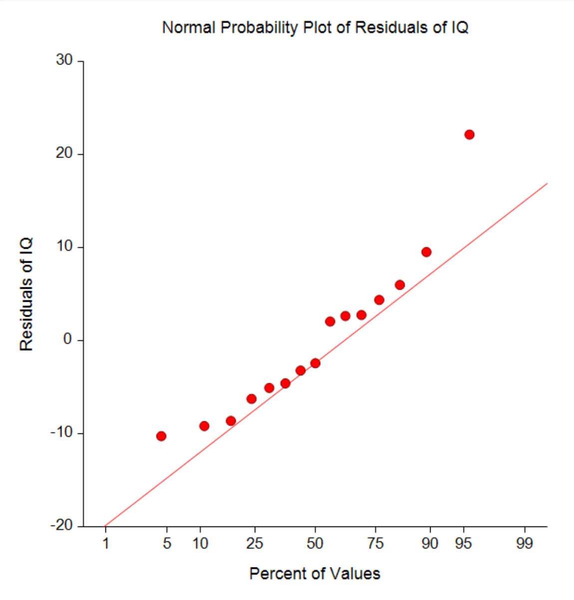


Probability Plot of Residuals

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

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

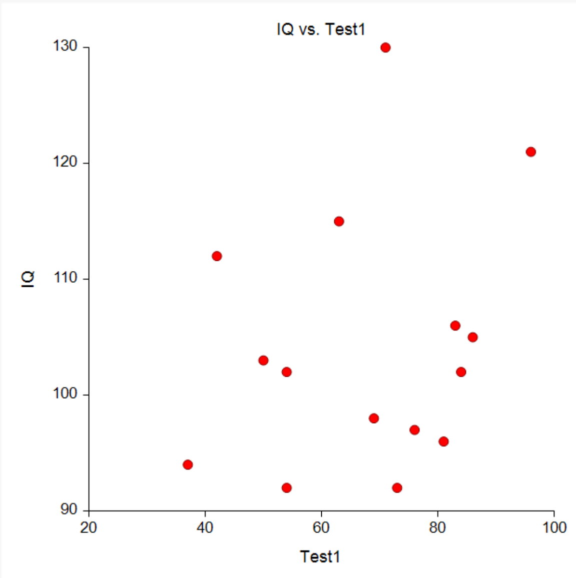
Plots Section

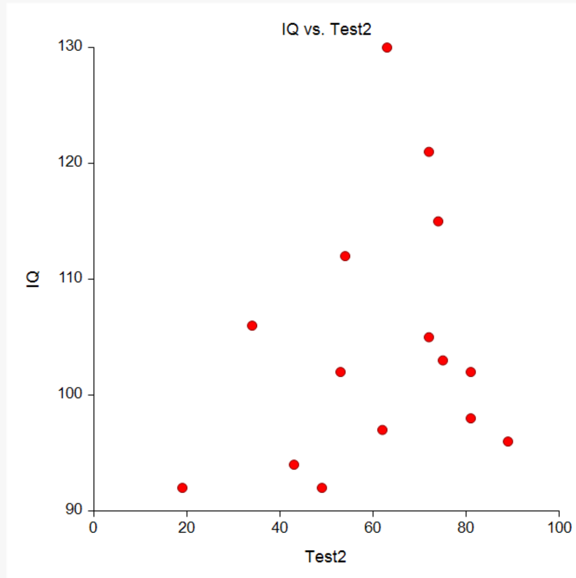


Plots of Y versus each IV

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

Plots Section





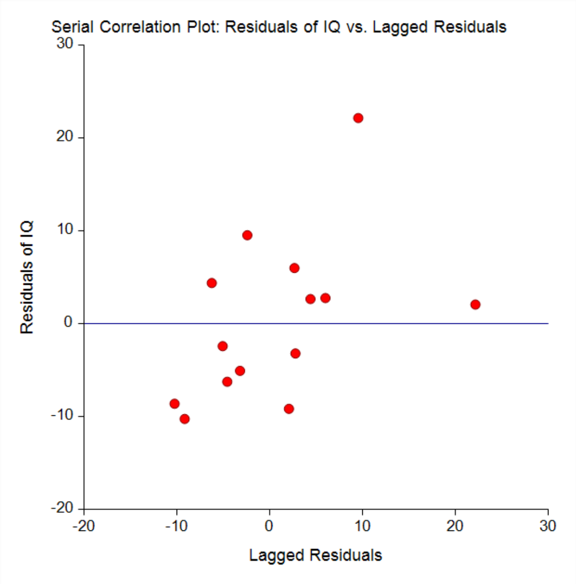
(More plots follow)

Serial Correlation of Residuals Plot

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

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

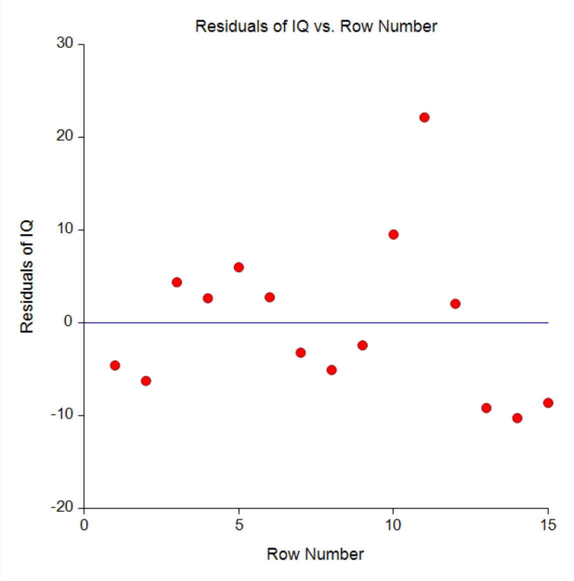
Plots Section



Sequence Plot

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

Plots Section

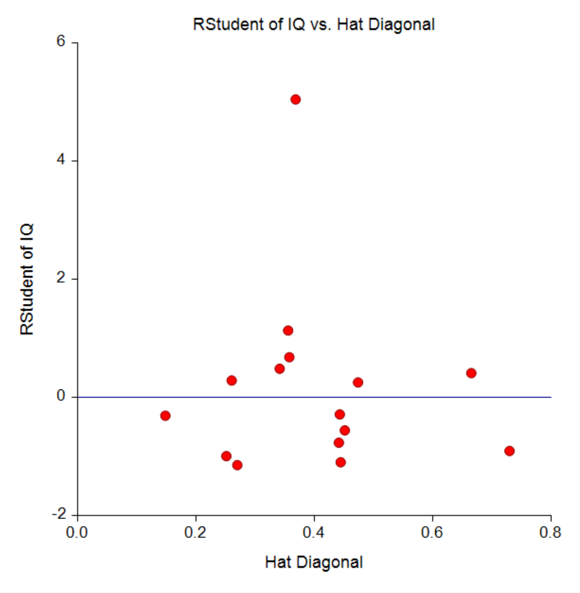


RStudent vs Hat Diagonal Plot

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.

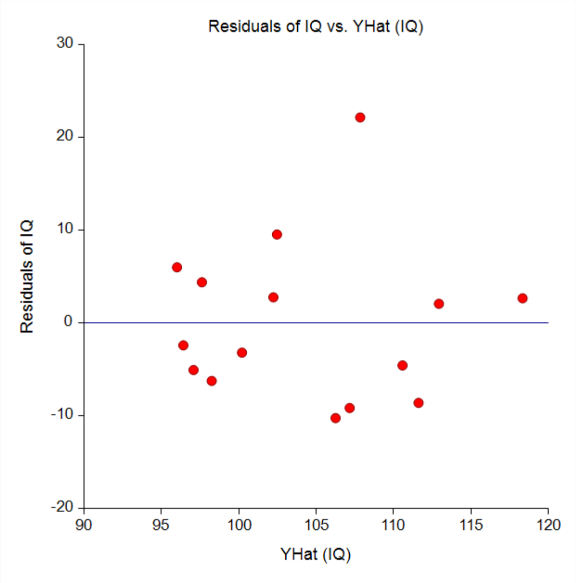
Plots Section



Residual vs Predicted Plot

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

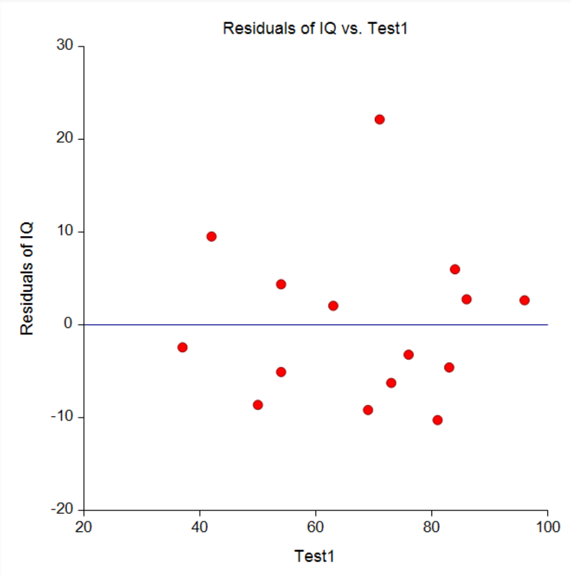
Plots Section

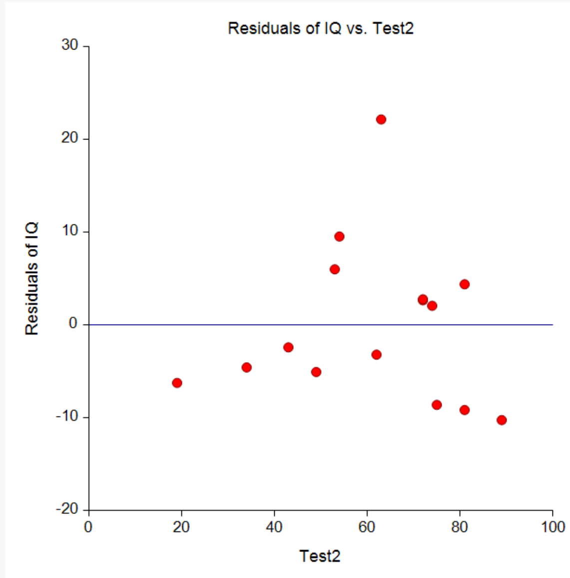


Residual vs Predictor(s) Plot

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

Plots Section



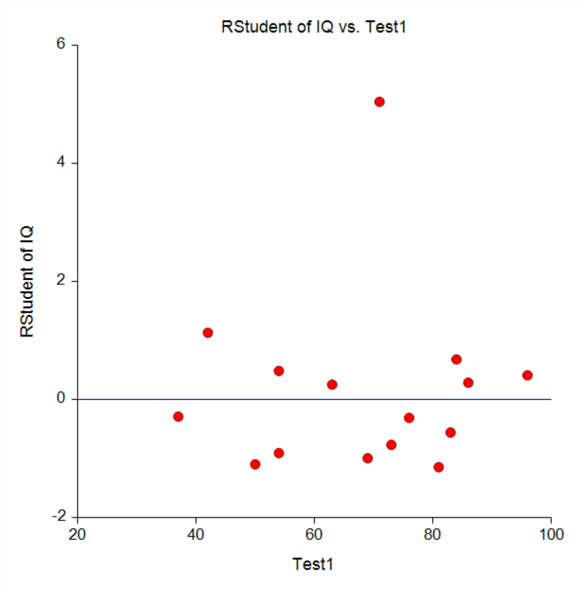


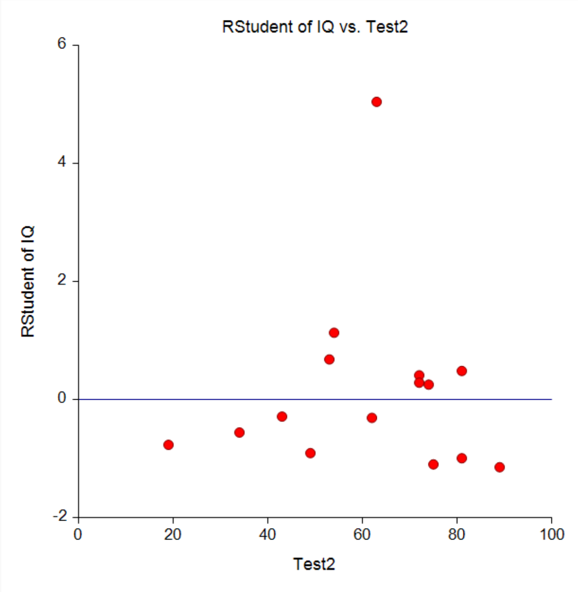
(More plots follow)

RStudent vs Predictor(s)

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

Plots Section



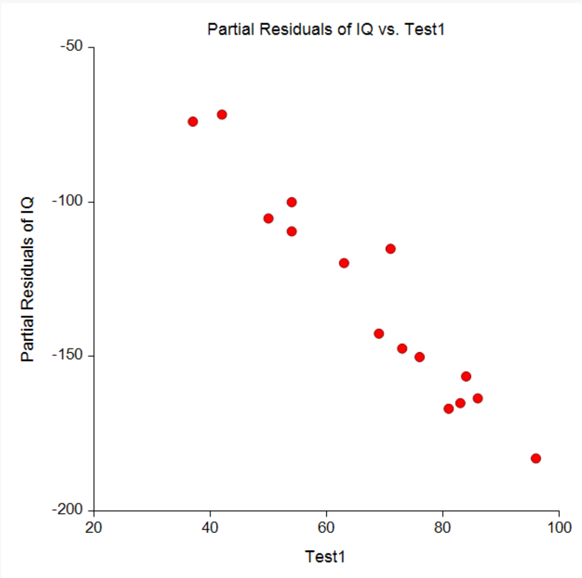


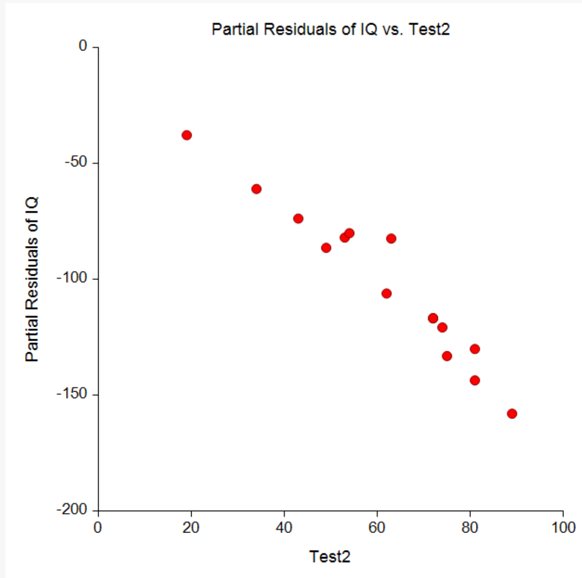
(More plots follow)

Partial Residual Plots

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

Plots Section





(More plots follow)

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 (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** 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 Tab

Y Dependent Variable(s) ....**IQ**
X's Numeric Independent Variables................**Test1-Test2,Test4**
Calculate Bootstrap C.I.'s ....**Checked**

Reports Tab

Select a Group of Reports and Plots ....**Display only those items that are CHECKED BELOW**
Regression Coefficients..................................**Checked**

Resampling Tab

Samples (N)..**3000**
Random Seed...**5768267** (for reproducibility)

3 Run the procedure

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

Regression Coefficient Section

Regression Coefficient Confidence Intervals Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper95% Conf. Limit of β(i)	Standardized Coefficient
Intercept	90.7327	12.8272	62.5003	118.9651	0.0000
Test1	-1.9650	0.9406	-4.0353	0.1053	-3.1020
Test2	-1.6485	0.7980	-3.4048	0.1078	-2.9024
Test4	3.7890	1.6801	0.0912	7.4869	4.7988

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

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

Bootstrap Section

Bootstrap Section

Estimation Results		Bootstrap Confidence Limits		
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
Intercept				
Original Value	90.7327	0.9000	67.5586	109.8079
Bootstrap Mean	92.1493	0.9500	60.9929	114.5635
Bias (BM - OV)	1.4166	0.9900	47.5995	130.1956
Bias Corrected	89.3161			
Standard Error	13.2971			
B(Test1)				
Original Value	-1.9650	0.9000	-3.0029	-0.0550
Bootstrap Mean	-2.1345	0.9500	-3.2805	0.5906
Bias (BM - OV)	-0.1695	0.9900	-4.1577	1.8078
Bias Corrected	-1.7955			
Standard Error	0.9760			
B(Test2)				
Original Value	-1.6485	0.9000	-2.5471	0.1086
Bootstrap Mean	-1.8259	0.9500	-2.7819	0.7748
Bias (BM - OV)	-0.1774	0.9900	-3.5368	2.0172
Bias Corrected	-1.4711			
Standard Error	0.8742			
B(Test4)				
Original Value	3.7890	0.9000	0.3325	5.7402
Bootstrap Mean	4.1124	0.9500	-0.9312	6.3442
Bias (BM - OV)	0.3234	0.9900	-3.1891	7.9426
Bias Corrected	3.4656			
Standard Error	1.7864			

Predicted Mean and Confidence Limits of IQ When Row = 16

Original Value	99.509	|	0.9000	92.703	105.490
Bootstrap Mean	99.749	|	0.9500	90.762	107.214
Bias (BM - OV)	0.240	|	0.9900	86.011	113.245
Bias Corrected	99.269				
Standard Error	4.078				

Predicted Mean and Confidence Limits of IQ When Row = 17

Original Value	101.264	|	0.9000	96.439	105.725
Bootstrap Mean	101.269	|	0.9500	95.552	106.861
Bias (BM - OV)	0.005	|	0.9900	92.836	110.324
Bias Corrected	101.259				
Standard Error	2.899				

Predicted Value and Prediction Limits of IQ When Row = 16

Original Value	99.509	|	0.9000	69.008	122.502
Bootstrap Mean	100.941	|	0.9500	63.042	128.629
Bias (BM - OV)	1.432	|	0.9900	50.647	141.444
Bias Corrected	98.077				
Standard Error	16.330				

Predicted Value and Prediction Limits of IQ When Row = 17

Original Value	101.264	|	0.9000	71.654	124.119
Bootstrap Mean	102.893	|	0.9500	66.029	129.958
Bias (BM - OV)	1.629	|	0.9900	52.307	143.231
Bias Corrected	99.635				
Standard Error	16.301				

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000,
User-Entered Random Seed = 5768267.

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

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

Standard Error

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

Conf. Level

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

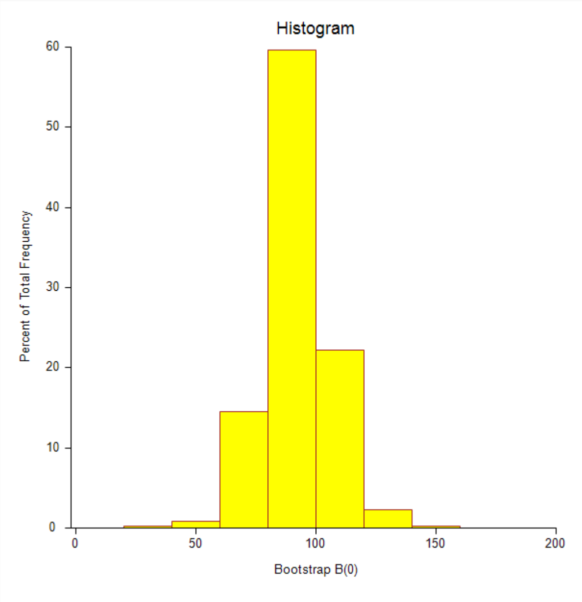
Bootstrap Confidence Limits - Lower and Upper

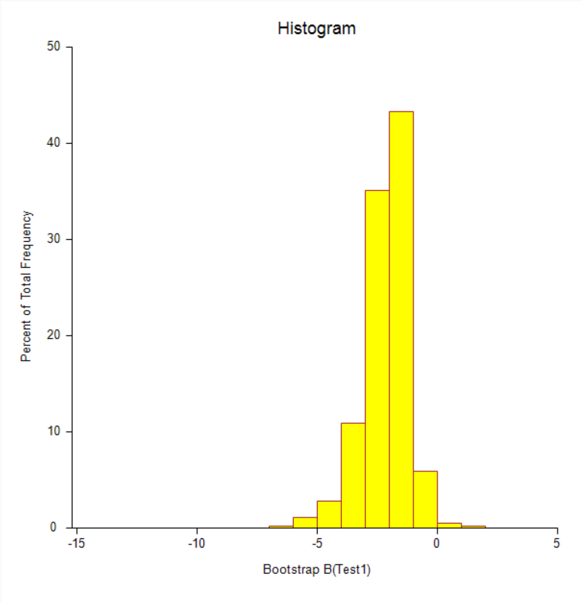
These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

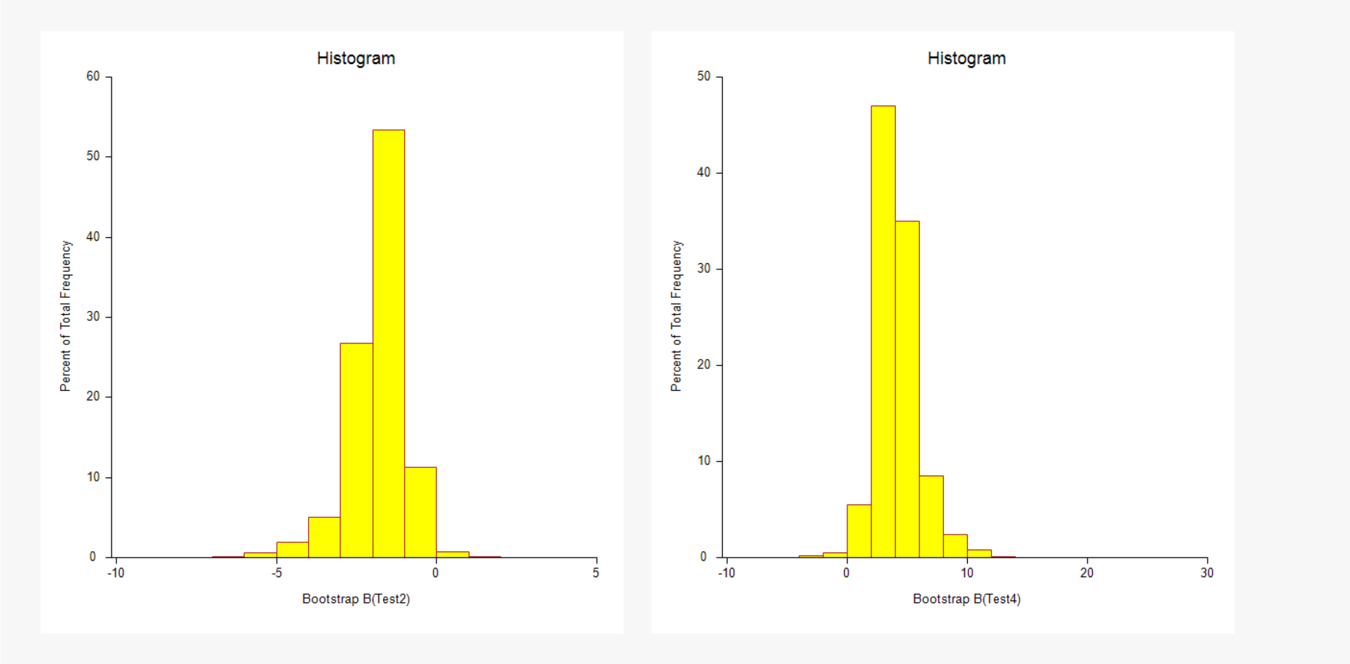
Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

Bootstrap Histograms Section

Bootstrap Histograms Section







Each histogram shows the distribution of the corresponding parameter estimate.

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

Example 3 – Robust Regression

This section presents an example of how to generate bootstrap confidence intervals with a multiple regression analysis. The tutorial will use the data are in the IQ database. This example will run a regression of IQ on Test1 through Test5.

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 (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** 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 Tab

Y Dependent Variable(s) ....**IQ**
X's Numeric Independent Variables................**Test1-Test5**
Perform Robust Regression............................**Checked**

Reports Tab

Select a Group of Reports and Plots ....**Display only those items that are CHECKED BELOW**
Equation ....**Checked**
Robust Coefficients..**Checked**
Robust Percentiles..**Checked**
Robust Residuals...**Checked**

Robust Tab

Robust Method ....**Huber's Method**
Minimum % Beta Change **1.0**

3 Run the procedure

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

Regression Equation Section

Regression Coefficient T-Tests Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0: β(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	60.7985	15.7492	3.860	0.0038	Yes	0.9285
Test1	-1.4085	0.6364	-2.213	0.0542	No	0.5063
Test2	-1.1785	0.5425	-2.173	0.0579	No	0.4919
Test3	0.1926	0.1403	1.373	0.2031	No	0.2332
Test4	2.8696	1.1329	2.533	0.0321	Yes	0.6173
Test5	0.1162	0.1328	0.874	0.4046	No	0.1229

This report gives the robust regression coefficients as well as *t*-tests. Note that the statistical tests are approximate because we are using robust regression. You could generate bootstrap robust confidence intervals to supplement these results.

Robust Regression Coefficient Section

Robust Regression Coefficients Section

Robust Iteration	Max % Change in any Beta	Robust B(0)	Robust B(1)	Robust B(2)	Robust B(3)
0		85.2404	-1.9336	-1.6599	0.1050
1	244.726	71.6768	-1.6799	-1.4283	0.1648
2	61.163	66.7707	-1.5881	-1.3446	0.1865
3	23.552	62.3507	-1.4718	-1.2368	0.1951
4	3.886	60.8935	-1.4180	-1.1887	0.1952
5	1.493	60.8642	-1.4135	-1.1832	0.1933
6	0.795	60.7985	-1.4085	-1.1785	0.1926

This report shows the largest percent change in any of the regression coefficients as well as the first four regression coefficients. The first iteration always shows the ordinary least squares estimates on the full dataset so that you can compare these value with those that occur after a few robust iterations.

This report allows you to determine if enough iterations have been run for the coefficients to have stabilized. In this example, the coefficients have stabilized. If they had not, we would decrease the value of the Minimum % Beta Change and rerun the analysis.

Robust Percentiles of Residuals Section

Robust Percentiles of Residuals Section

Iter. No.	Max % Change in any Beta	Percentiles of Absolute Residuals 25th	50th	75th	100th
0		2.767	5.073	9.167	22.154
1	244.726	1.726	4.446	7.637	27.573
2	61.163	1.573	3.093	7.084	29.533
3	23.552	1.511	2.599	7.083	30.626
4	3.886	1.564	2.285	7.296	30.714
5	1.493	1.569	2.271	7.387	30.604
6	0.795	1.581	2.252	7.440	30.553

The purpose of this report is to highlight the maximum percentage changes among the regression coefficients and to show the convergence of the absolute value of the residuals after a selected number of iterations.

Iter. No.

This is the robust iteration number.

Max % Change in any Beta

This is the maximum percentage change in any of the regression coefficients from one iteration to the next. This quantity can be used to determine if enough iterations have been run. Once this value is less than five percent, little is gained by further iterations.

Percentiles of Absolute Residuals

The absolute values of the residuals for this iteration are sorted and the percentiles are calculated. We want to terminate the iteration process when there is little change in median of the absolute residuals.

Robust Residuals and Weights Section

Robust Residuals and Weights

Row	Actual IQ	Predicted IQ	Residual	Absolute Percent Error	Robust Weight
1	106.000	104.565	1.435	1.354	1.0000
2	92.000	96.915	-4.915	5.343	1.0000
3	102.000	100.078	1.922	1.884	1.0000
4	121.000	121.703	-0.703	0.581	1.0000
5	102.000	98.551	3.449	3.381	1.0000
6	105.000	100.270	4.730	4.504	1.0000
7	97.000	98.450	-1.450	1.495	1.0000
8	92.000	94.252	-2.252	2.448	1.0000
9	94.000	96.007	-2.007	2.135	1.0000
10	112.000	103.875	8.125	7.255	0.6515
11	130.000	99.447	30.553	23.502	0.1716
12	115.000	113.419	1.581	1.375	1.0000
13	98.000	105.440	-7.440	7.592	0.7108
14	96.000	105.269	-9.269	9.655	0.5720
15	103.000	104.735	-1.735	1.684	1.0000
16		90.367			0.0000
17		96.281			0.0000

The predicted values, the residuals, and the robust weights are reported for the last iteration. These robust weights can be saved for use in a weighted regression analysis, or they can be used as a filter to delete observations with a weight less than some number, say 0.20, in an ordinary least squares regression analysis.

Note that in this analysis, row 11 appears to be an outlier.

Row

This is the number of the row. Rows whose weight is less than 0.1 are starred.

Actual

This is the actual value of the dependent variable.

Predicted

This is the predicted value of *Y* based on the robust regression equation from the final iteration.

Residual

The residual is the difference between the Actual and Predicted values of *Y*.

Robust Weight

Once the convergence criteria for the robust procedure have been met, these are the final weights for each observation.

These weights will range from zero to one. Observations with a low weight make a minimal contribution to the determination of the regression coefficients. In fact, observations with a weight of zero have been deleted from the analysis. These weights can be saved and used again in a weighted least squares regression.

Example 4 – Variable Subset Selection

This section presents an example of how to select a subset of the available IV's that are the most useful in predicting *Y*. The tutorial will use the data are in the IQ database. In this example, we will select a subset from the five IV's available.

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 (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Y Dependent Variable(s) ....**IQ**
X's Numeric Independent Variables................**Test1-Test5**

Model Tab

Subset Selection...**Hierarchical Forward with Switching**
Max Terms in Subset.....................................**6**
Which Model Terms..**Up to 2-Way**

Reports Tab

Select a Group of Reports and Plots ....**Display items appropriate for a STANDARD ANALYSIS**

3 Run the procedure

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

Subset Selection Summary Section

Subset Selection Summary Section

No. Terms	No. X's	R-Squared Value	R-Squared Change
1	1	0.1379	0.1379
2	2	0.1542	0.0163
3	3	0.2466	0.0924
4	4	0.3587	0.1121
5	5	0.5681	0.2094
6	6	0.5877	0.0196

This report shows the number of terms, number of IV's, and *R*-squared values for each subset size. This report is used to determine an appropriate subset size for a second run. You search the table for a subset size after which the *R*-squared increases only slightly as more variables are added.

In this example, there appears to be two places where a break occurs: from 1 to 2 terms and from 5 to 6 terms. Under normal circumstances, we would pick from a subset size of 5 for a second run. However, because the sample size in this example is only 15, we would not want to go above a subset size of 3 (our rule of thumb is *N*/#IV's > 5).

Subset Selection Detail Section

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	R2	Term Entered	Term Removed
0	Add	0	0	0.0000	Intercept	
1	Add	1	1	0.1379	Test4	
2	Add	2	2	0.1542	Test3	
3	Add	3	3	0.2466	Test3*Test3	
4	Add	4	4	0.3587	Test4*Test4	
5	Add	5	5	0.4149	Test2	
6	Switch	5	5	0.4203	Test1	Test3*Test3
7	Switch	5	5	0.5681	Test2*Test2	Test4*Test4
8	Add	6	6	0.5877	Test1*Test1	

This report shows the details of which variables were added or removed at each step in the search procedure. The final model for three IV's would include Test4, Test3, and Test3*Test3.

Because of the restrictions due to our use of hierarchical models, you might run an analysis using the Forward with Switching option as well as a search without 2-way interactions. Because of the small sample size, these options produce models with much larger *R*-squared values. However, it is our feeling that this larger *R*-squared values occur because the extra variables are actually fitting random error rather than a reproducible pattern.

Example 5 – Sales Price Prediction

This section presents an example of using multiple regression to construct an equation that predicts the sales price of a home based on a few basic IV's such as square footage, lot size, and so on. The Resale dataset contains several variables relating to the sales price of a house. These include year built, number of bedrooms, number of bathrooms, size of garage, number of fireplaces, overall quality rating, amount of building with brick, finished square footage, total square footage, and lot size.

The Resale dataset contains data on 150 sales that took place recently. Our task is to develop a mathematical model that relates sales price to the IV's listed about and then use this model to predict the eventual sales price for two additional properties.

Step 1 – View Scatter Plots

The starting point in such an analysis is to view individual scatter plots of sales price versus each of the potential IV's looking for outliers, curvilinear patterns, and other anomalies. Although we could create these scatter plots in other procedures, we will use the Multiple Regression procedure to do so.

Setup

To run this example, complete the following steps:

1 Open the Resale example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5-1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Y Dependent Variable(s) ....**Price**
X's Numeric Independent Variables................**Year-LotSize**

Reports Tab

Select a Group of Reports and Plots ....**Display only those items that are CHECKED BELOW**

Plots Tab

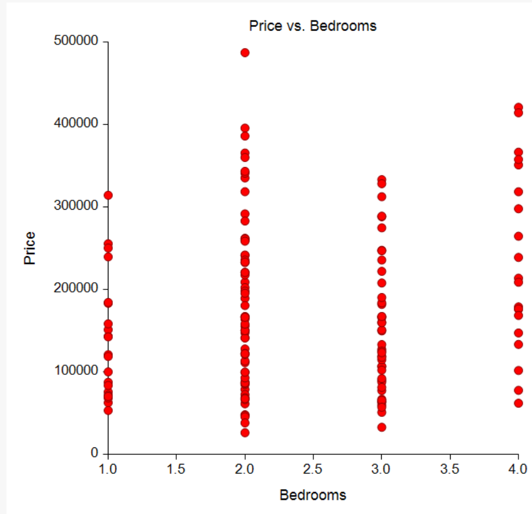
Y vs X ....**Checked**

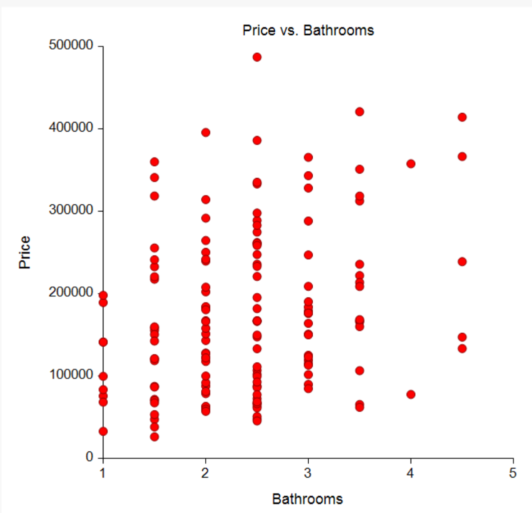
3 Run the procedure

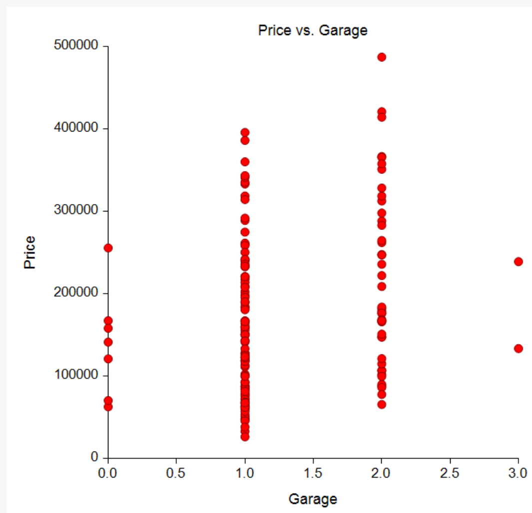
- Click the **Run** button to perform the calculations and generate the output.

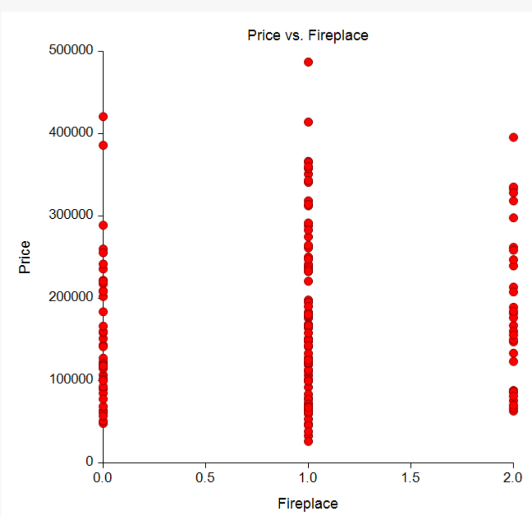
Scatter Plot Output

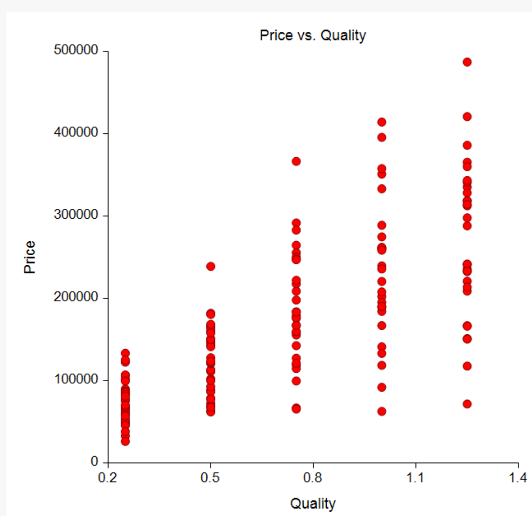
Plots Section

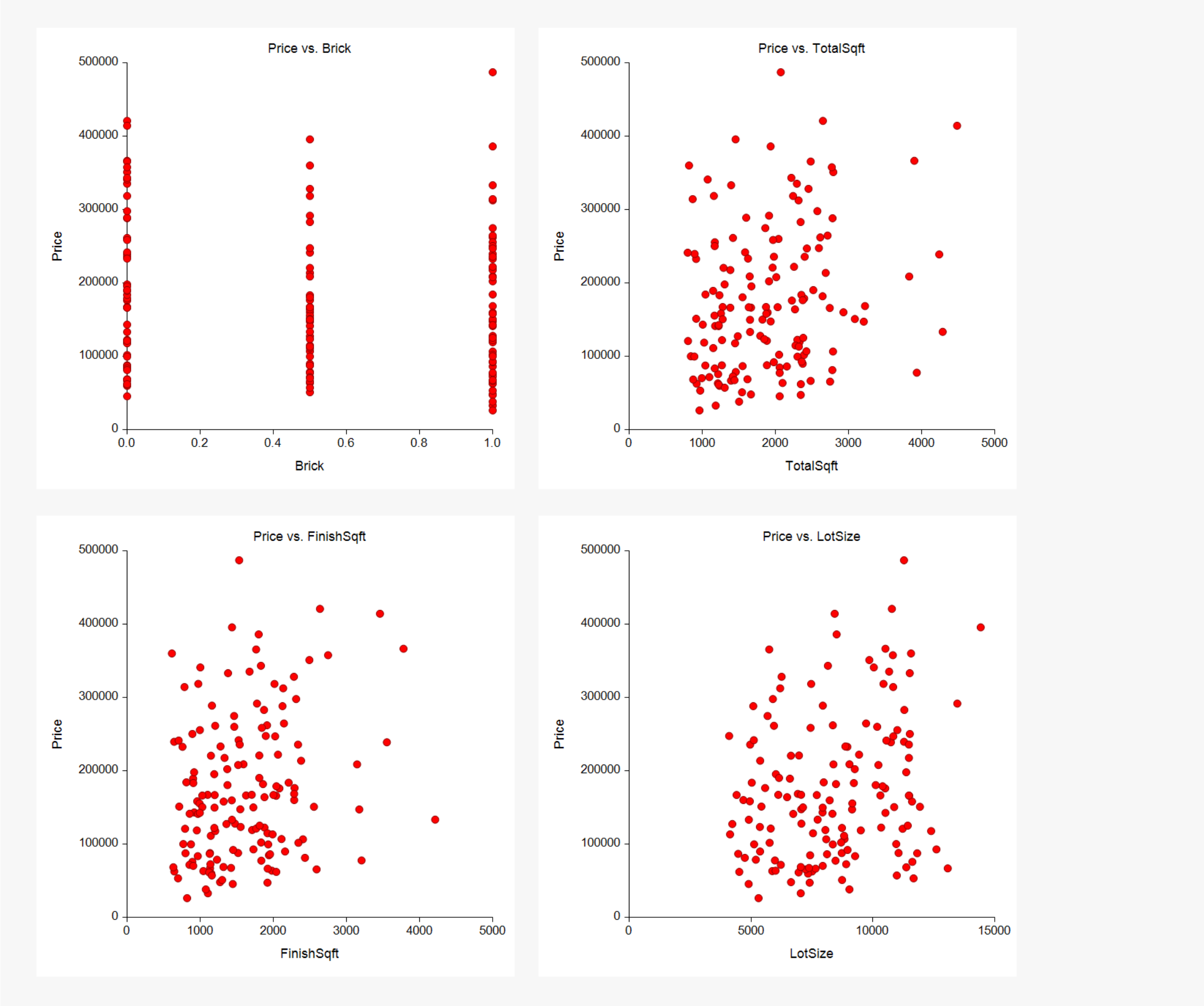












Looking at these plots, we notice that Bathrooms, Quality, and Year appear to have the most direct relationship with price. We cannot spot any outliers, so we procedure to the next step.

Step 2 – Use Robust Regression to Find Outliers

Although we could not spot any outliers on the scatter plots, it is important to make sure that we have not missed any. To do this, we run a robust regression analysis and search the robust weights for values less that 0.20 (which we define as an outlier).

This analysis assumes that you have just completed Example 5-1. You may follow along here by making the appropriate entries or load the completed settings file **Example 5-2** by clicking on Open Example Settings File from the File menu of the Multiple Regression window.

Setup

To run this example, complete the following steps:

1 Open the Resale example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5-2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Perform Robust Regression............................**Checked**

Reports Tab

Robust Coefficients...**Checked**
Robust Residuals...**Checked**

Robust Tab

Minimum % Beta Change **0.1**
Cutoff for Weight Report **0.40**

3 Run the procedure

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

Robust Regression Output

Robust Regression Coefficients Section

Robust Iteration	Max % Change in any Beta	Robust B(0)	Robust B(1)	Robust B(2)	Robust B(3)
0		-6975033.8132	3466.1399	9068.0709	-377.5098
1	146.222	-6907482.9980	3432.2573	8114.1221	174.4935
2	43.791	-6898667.1571	3427.6793	8059.9605	250.9067
3	6.395	-6896910.7470	3426.7382	8062.1699	266.9523
4	1.712	-6896384.8015	3426.4524	8065.1218	271.5213
5	0.608	-6896269.5148	3426.3862	8066.4706	273.0447
6	0.369	-6896265.1890	3426.3806	8066.9712	273.6238
7	0.206	-6896281.4591	3426.3874	8067.1474	273.8489
8	0.109	-6896295.9868	3426.3940	8067.2078	273.9371
9	0.056	-6896305.4524	3426.3985	8067.2279	273.9723

Robust Residuals and Weights

Row	Actual Sales Price	Predicted Sales Price	Residual	Absolute Percent Error	Robust Weight
55	32900.000	-70171.619	103071.619	313.288	0.3426
120	117800.000	210610.031	-92810.031	78.786	0.3805
150	487200.000	373849.490	113350.510	23.266	0.3115

From a perusal of these reports, we learn that there are three potential outliers: rows 55, 120, and 150. However, their robust weights are much larger than the cutoff value of 0.200 which we set as an indicator of when an observation is an outlier. So, even though these three observations are predicted poorly, we decide to leave them in the dataset for the rest of the analysis.

Step 3 – Variable Selection

The next step is to search for the most useful subset of the IV's. To do this, we made an initial search for each subset up to ten IV's. We will study the R-squared values to determine a reasonable subset size.

This analysis assumes that you have just completed Example 5-2. You may follow along here by making the appropriate entries or load the completed settings file **Example 5-3** by clicking on Open Example Settings File from the File menu of the Multiple Regression window.

Setup

To run this example, complete the following steps:

1 Open the Resale example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5-3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Perform Robust Regression............................**Unchecked**

Model Tab

Subset Selection...**Hierarchical Forward with Switching**
Max Terms in Subset......................................**10**
Which Model Terms..**Up to 2-Way**

Reports Tab

Subset Summary ....**Checked**
Subset Detail ....**Checked**

3 Run the procedure

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

Variable Selection Output

Subset Selection Summary Section

No. Terms	No. X's	R-Squared Value	R-Squared Change
1	1	0.5212	0.5212
2	2	0.7676	0.2464
3	3	0.8440	0.0764
4	4	0.8929	0.0489
5	5	0.8956	0.0027
6	6	0.8969	0.0014
7	7	0.9009	0.0039
8	8	0.9020	0.0011
9	9	0.9031	0.0011
10	10	0.9037	0.0006

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	R2	Term Entered	Term Removed
0	Add	0	0	0.0000	Intercept	
1	Add	1	1	0.5212	Quality Index	
2	Add	2	2	0.7676	Year Built	
3	Add	3	3	0.8440	Total Area (Sqft)	
4	Add	4	4	0.8929	Lot Size (Sqft)	
5	Add	5	5	0.8956	Bedrooms	
6	Add	6	6	0.8968	Brick Ratio	
7	Switch	6	6	0.8969	Brick Ratio*Brick Ratio	Bedrooms
8	Add	7	7	0.9009	Bedrooms	
9	Add	8	8	0.9020	Fireplaces	
10	Add	9	9	0.9031	Fireplaces*Fireplaces	
11	Add	10	10	0.9037	Fireplaces*Brick Ratio	

Scanning down the *R*-squared values, it is easy to see that the appropriate subset size is four. With four IV's, an *R*-squared of 0.8929 is achieved which is impressive for this type of data. From the Subset Selection Detail report, we learn that the four IV's are Quality, Year, TotalSqrt, and LotSize. These seem to be a reasonable basis for sales price estimation.

Step 4 – Standard Regression

The next step is to generate a standard regression analysis using the four IV's that were selected in the last step.

This analysis assumes that you have just completed Example 5-3. You may follow along here by making the appropriate entries or load the completed settings file **Example 5-4** by clicking on Open Example Settings File from the File menu of the Multiple Regression window.

Setup

To run this example, complete the following steps:

1 Open the Resale example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5-4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

X's Numeric Independent Variables................**Year,Quality,TotalSqft,LotSize**

Model Tab

Subset Selection...**None - No Search is Conducted**
Which Model Terms..**Up to 1-Way**

Reports Tab

Select a Group of Reports and Plots ....**Display items appropriate for a STANDARD ANALYSIS**

3 Run the procedure

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

Standard Regression Output

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	Sales Price	Rows Processed	150
Number Ind. Variables	4	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	0
R2	0.8929	Rows with Weight Missing	0
Adj R2	0.8899	Rows with Y Missing	0
Coefficient of Variation	0.1858	Rows Used in Estimation	150
Mean Square Error	1.049649E+09	Sum of Weights	150.000
Square Root of MSE	32398.29	Completion Status	Normal Completion
Ave Abs Pct Error	22.636		

We have only included the Run Summary report here. You can look at the complete output when you run this example. We note that the final *R*-squared value is 0.8929, which is pretty good, but the average absolute percent error is 22.636%, which is disturbing.

This completes this analysis. If you wanted to use these results to predict the sales price of additional properties, you would simple add the data to the bottom of the database, leaving the Price variable blank. The Predicted Individuals report will give the estimates and prediction limits for these additional properties.

Example 6 – Checking the Parallel Slopes Assumption in Analysis of Covariance

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

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

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

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 (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 6** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Variables Tab

Y Dependent Variable(s) ....**IQ**
X's Numeric Independent Variables................**Age**
X's Categorical Independent Variables...........**State**
Default Contrast Type....................................**Standard Set**

Model Tab

Which Model Terms.......................................**Full Model**

Reports Tab

Select a Group of Reports and Plots ....**Display only those items that are CHECKED BELOW**
ANOVA Detail..**Checked**

3 Run the procedure

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

Analysis of Variance Detail Section

Analysis of Variance Detail Section

Model Term	DF	R2	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1		313345.2	313345.2			
Model	5	0.2438	80.15984	16.03197	1.547	0.2128	0.4472
Age	1	0.0296	9.740934	9.740934	0.940	0.3419	0.1537
State	2	0.1417	46.57466	23.28733	2.248	0.1274	0.4123
Age*State	2	0.1178	38.72052	19.36026	1.869	0.1761	0.3500
Error	24	0.7562	248.6402	10.36001			
Total(Adjusted)	29	1.0000	328.8	11.33793			

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

Example 7 – Analyzing Pre-Post Data with both Categorical and Numeric IV's

The PrePost dataset contains the results of a study involving 144 subjects that were divided into three groups. The first group (Control) received a placebo, the second group (Dose20) received a small dose of the drug of interest, and the third group (Dose40) received a large dose of the drug of interest. Each subject response was measured before (Pre) and after (Post) the drug was administered, and the gain from Pre to Post was calculated. Also, each subject's propensity score was measured. This Propensity is a combined index created from several demographic variables. The age group (Age) of each subject was also recorded.

The goal of the research is to build a regression model from this data that will allow the gain scores to be predicted. The model should include all significant interaction terms.

Step 1 – Scan for Outliers Using Robust Regression

The first step is to scan for outliers using robust regression. Of course, you should also look at the scatter plots of *Y* versus each IV. The robust regression is useful because it provides a list of potential outliers even when interactions are included. It is often difficult to find true outliers when interactions are included.

Setup

To run this example, complete the following steps:

1 Open the PrePost example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 7-1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Y Dependent Variable(s) ....**Gain**
X's Numeric Independent Variables................**Pre,Propensity**
X's Categorical Independent Variables...........**Group-Age**
Perform Robust Regression............................**Checked**

Model Tab

Which Model Terms..**Up to 2-Way**

Reports Tab

Select a Group of Reports and Plots .... **Display only those items that are CHECKED BELOW**
Run Summary... **Checked**
Robust Coefficients... **Checked**
Robust Percentiles.. **Checked**
Robust Residuals.. **Checked**

Robust Tab

Minimum % Beta Change **1.0**
Maximum Iterations **20**
Cutoff for Weight Report.................................. **0.50**

3 Run the procedure

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

Robust Regression Output

Robust Regression Coefficients Section

Robust Iteration	Max % Change in any Beta	Robust B(0)	Robust B(1)	Robust B(2)	Robust B(3)
0		19.3828	-2.6061	0.8632	-7.1093
1	1238.621	18.1621	-2.4050	0.7838	-7.4761
2	25.133	17.1554	-2.1905	0.6801	-7.5543
3	15.109	16.4886	-2.0322	0.6002	-7.5867
4	15.095	16.1000	-1.9235	0.5419	-7.6036
5	12.774	15.8236	-1.8637	0.5130	-7.6174
6	9.683	15.6587	-1.8243	0.4931	-7.6274
7	7.007	15.5714	-1.7971	0.4780	-7.6319
8	4.955	15.5352	-1.7780	0.4660	-7.6323
9	2.864	15.5109	-1.7677	0.4598	-7.6308
10	1.494	15.4926	-1.7620	0.4566	-7.6305
11	0.995	15.4860	-1.7577	0.4538	-7.6301

Robust Residuals and Weights

Row	Actual Gain	Predicted Gain	Residual	Absolute Percent Error	Robust Weight
9	222.000	203.685	18.315	8.250	0.2893
16	174.000	158.661	15.339	8.815	0.3452
30	24.000	35.324	-11.324	47.183	0.4673
45	214.000	195.817	18.183	8.497	0.2914
53	5.000	-5.711	10.711	214.220	0.4941
54	57.000	69.484	-12.484	21.902	0.4240
99	260.000	232.035	27.965	10.756	0.1895
105	73.000	85.251	-12.251	16.783	0.4320
106	54.000	64.679	-10.679	19.776	0.4957
116	204.000	187.062	16.938	8.303	0.3128
144	6.000	-6.181	12.181	203.011	0.4346

There are only a few suspected outliers. Row 99 was especially suspicious since its weight is less than 0.20. We also looked at the Regression Diagnostics report and found that these rows also had large values RStudent and Dffits. However, since we could find nothing wrong with the data for these subjects and since we want our final equation to represent as wide of a population as possible, we decided to include these rows in the rest of the analysis.

Step 2 – Search for a Parsimonious Model

Once we have determined that our data is as free of large outliers as we wish, our next task is to conduct a variable selection phase to find a model with as few IV's as possible which still achieves a high *R*-squared value. The Run Summary report (not shown above) listed the an *R*-squared of 0.9894 with a total of 21 IV's. Our goal in this phase is to substantially decrease the number of IV's while achieving an *R*-squared near 0.9894. Because we are fitting interactions, we will conduct as hierarchical forward search with switching.

Note that the changes listed below assume that you have just completed Step 1. You may follow along here by making the appropriate entries or load the completed settings file **Example 7-2** by clicking on Open Example Settings File from the File menu of the Multiple Regression window.

Setup

To run this example, complete the following steps:

1 Open the PrePost example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 7-2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Perform Robust Regression............................**Unchecked**

Model Tab

Subset Selection...**Hierarchical Forward with Switching**
Max Terms in Subset.......................................**10**
Which Model Terms...**Up to 2-Way**

Reports Tab

Select a Group of Reports and Plots ....**Display only those items that are CHECKED BELOW**
Run Summary...**Checked**
Subset Summary ....**Checked**
Subset Detail ....**Checked**

3 Run the procedure

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

Variable Selection Output

Subset Selection Summary Section

No. Terms	No. X's	R-Squared Value	R-Squared Change
1	1	0.3514	0.3514
2	3	0.7334	0.3821
3	5	0.7433	0.0099
4	9	0.7618	0.0185
5	7	0.9854	0.2236
6	8	0.9862	0.0008
7	10	0.9879	0.0017
8	11	0.9880	0.0001
9	16	0.9885	0.0005
10	18	0.9889	0.0003

Subset Selection Detail Section

Step	Action	No. of Terms	No. of X's	R2	Term Entered	Term Removed
0	Add	0	0	0.0000	Intercept	
1	Add	1	2	0.3514	Propensity	
2	Add	2	2	0.7290	Group	
3	Switch	2	3	0.7334	Pre	Propensity
4	Add	3	4	0.7433	Age	
5	Add	4	9	0.7618	Group*Age	
6	Add	5	10	0.7690	Pre*Pre	
7	Switch	5	8	0.9822	Pre*Group	Group*Age
8	Switch	5	7	0.9854	Propensity	Age
9	Add	6	8	0.9862	Propensity*Propensity	
10	Add	7	10	0.9879	Propensity*Group	
11	Add	8	11	0.9880	Pre*Propensity	
12	Add	9	13	0.9880	Age	
13	Switch	9	14	0.9882	Pre*Age	Pre*Propensity
14	Switch	9	16	0.9884	Group*Age	Pre*Age
15	Switch	9	15	0.9884	Pre*Propensity	Propensity*Group
16	Switch	9	16	0.9885	Pre*Age	Pre*Propensity
17	Add	10	18	0.9889	Propensity*Age	

We notice from the Subset Selection Summary report that the first five terms achieve an *R*-squared of 0.9854. After that, additional terms increase *R*-squared very little. We decide to include the first five terms in our model.

The Subset Selection Detail report shows that these five terms are: Group, Pre, Propensity, Pre*Pre, and Group*Pre.

Step 3 – Estimate the Model

The next step is to estimate the regression equation and evaluate the residual plots. There are two ways to create the model. The first way is to reset the maximum number of terms to five and rerun the subset selection. The second way is enter the final model in the Custom Model box. This has the advantage that you can run other analyses, such as robust regression, which are not possible during a variable search. So we setup the analysis using the second method.

Note that the changes listed below assume that you have just completed Step 2. You may follow along here by making the appropriate entries or load the completed settings file **Example 7-3** by clicking on Open Example Settings File from the File menu of the Multiple Regression window.

Setup

To run this example, complete the following steps:

1 Open the PrePost example dataset

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

2 Specify the Multiple Regression (Old Version) procedure options

- Find and open the **Multiple Regression (Old Version)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 7-3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

X's Categorical Independent Variables............**Group**

Model Tab

Subset Selection...**None - No Search is Conducted**
Which Model Terms..**Custom Model**
Custom Model...**Group Pre Pre*Pre Group*Pre Propensity**

Reports Tab

Select a Group of Reports and Plots ....**Display only those items that are CHECKED BELOW**
Run Summary...**Checked**
Equation ....**Checked**
Regression Coefficients..................................**Checked**
ANOVA Detail...**Checked**

3 Run the procedure

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

Standard Regression Output

Run Summary Section

Parameter	Value	Parameter	Value
Dependent Variable	Gain	Rows Processed	144
Number Ind. Variables	7	Rows Filtered Out	0
Weight Variable	None	Rows with X's Missing	0
R2	0.9854	Rows with Weight Missing	0
Adj R2	0.9847	Rows with Y Missing	0
Coefficient of Variation	0.1496	Rows Used in Estimation	144
Mean Square Error	38.70051	Sum of Weights	144.000
Square Root of MSE	6.220973	Completion Status	Normal Completion
Ave Abs Pct Error	47.269		

Regression Coefficient T-Tests Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0: β(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	11.5547	2.5123	4.599	0.0000	Yes	0.9954
(Group="Dose20")	-5.1942	2.7863	-1.864	0.0645	No	0.4567
(Group="Dose40")	-35.5054	2.7570	-12.878	0.0000	Yes	1.0000
Pre	-2.0806	0.2045	-10.173	0.0000	Yes	1.0000
Propensity	0.7301	0.0818	8.924	0.0000	Yes	1.0000
Pre*Pre	0.0241	0.0019	12.591	0.0000	Yes	1.0000
(Group="Dose20")*Pre	0.6312	0.0708	8.915	0.0000	Yes	1.0000
(Group="Dose40")*Pre	3.2646	0.0730	44.724	0.0000	Yes	1.0000

Regression Coefficient Confidence Intervals Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper95% Conf. Limit of β(i)	Standardized Coefficient
Intercept	11.5547	2.5123	6.5866	16.5229	0.0000
(Group="Dose20")	-5.1942	2.7863	-10.7043	0.3159	-0.0489
(Group="Dose40")	-35.5054	2.7570	-40.9575	-30.0532	-0.3345
Pre	-2.0806	0.2045	-2.4850	-1.6761	-0.7314
Propensity	0.7301	0.0818	0.5683	0.8919	0.3453
Pre*Pre	0.0241	0.0019	0.0203	0.0278	0.6285
(Group="Dose20")*Pre	0.6312	0.0708	0.4912	0.7713	0.2465
(Group="Dose40")*Pre	3.2646	0.0730	3.1203	3.4090	1.1848

Note: The T-Value used to calculate these confidence limits was 1.978.

This concludes the regression analysis. We have estimated a regression equation that contains only seven IV's, yet accounts for over 98% of the variability in the Gain score.

Note that the interpretation of the regression coefficients is difficult because of the inclusion of the Group*Pre interaction term. For example, the equation seems to indicate that the Gain is reduced by 5.1942 for the Dose20 group as compared to the Control group. However, the (Group=DOSE2)*Pre regression coefficient of 0.6312 will more than offset this value for most subjects because typical pretest values are greater than 10. That is, 10*0.6312 = 6.312 which is greater than 5.1942.

For example, a subject in the Dose20 group with a pretest score of 50 has an estimated gain score which is 26.3658 = -5.1942+0.6312(50) higher than a similar subject in the Control group.

As a final note, you may wish to adjust the structure of the Group variable. If you wanted to change the reference value to *DOSE40* rather than the default of *CONTROL*, you would change the Default Reference Value on the Variables tab to *Last Value after Sorting* or the X's: Categorical Independent Variables box from *Group* to *Group(DOSE40)* and rerun the analysis. This would yield the following table (you can generate this table by loading the completed settings file **Example 7-4** by clicking on Open Example Settings File from the File menu of the Multiple Regression window).

Standard Regression Output

Regression Coefficient T-Tests Section

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Value to test H0: β(i)=0	Prob Level	Reject H0 at 5%?	Power of Test at 5%
Intercept	-23.9507	2.6621	-8.997	0.0000	Yes	1.0000
(Group="Control")	35.5054	2.7570	12.878	0.0000	Yes	1.0000
(Group="Dose20")	30.3112	2.8590	10.602	0.0000	Yes	1.0000
Pre	1.1841	0.2087	5.674	0.0000	Yes	0.9999
Propensity	0.7301	0.0818	8.924	0.0000	Yes	1.0000
Pre*Pre	0.0241	0.0019	12.591	0.0000	Yes	1.0000
(Group="Control")*Pre	-3.2646	0.0730	-44.724	0.0000	Yes	1.0000
(Group="Dose20")*Pre	-2.6334	0.0753	-34.989	0.0000	Yes	1.0000

Estimated Model

-23.9506614382102+35.50538794438*(Group="Control")+30.3112093048755*(Group="Dose20")+1.18406725423673*Pre+0.730124848180748*Propensity+0.024058147508663*Pre*Pre-3.26462104811021*(Group="Control")*Pre-2.63337520487248*(Group="Dose20")*Pre