Chapter 307

# Subset Selection in Multiple Regression

## Introduction

Multiple regression analysis is documented in *Chapter 305 – Multiple Regression*, so that information will not be repeated here. Refer to that chapter for in depth coverage of multiple regression analysis. This chapter will deal solely with the topic of 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 internal variables that are generated for a categorical independent variable. If such a variable has six categories, five internal variables are generated. You can see that with two or three categorical variables, a large number of internal variables will 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 internal variables. Thus, the whole set of internal 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 internal variables in a 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.

# Selection Methods

## 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 $R^2$. Enter this term into the model.
3. Continue adding terms until a target value for $R^2$ 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 $R^2$. 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.

# 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 – Subset Selection in Multiple Regression

This section presents an example of how to run a subset selection from the data presented earlier in this chapter. The data are in the IQ dataset. This example will find a subset of three IV's from the candidate pool of *Test1* through *Test5*. The dependent variable is IQ. This program outputs over thirty different reports and plots, many of which contain duplicate information. Only those reports that are specifically needed for a subset selection will be present here.

## 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 Subset Selection in Multiple Regression procedure options**

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

| Variables, Model Tab | |
|---|---|
| Y | **IQ** |
| Numeric X's | **Test1-Test5** |
| Terms | **1-Way** |
| Search Method | **Hierarchical Forward Selection** |
| Stop search when number of terms reaches | **3** |

| Reports Tab | |
|---|---|
| Run Summary | **Checked** |
| Descriptive Statistics | **Checked** |
| Correlation Matrix | **Checked** |
| Subset Summary | **Checked** |
| Subset Detail | **Checked** |
| Coefficient T-Tests | **Checked** |
| Coefficient C.I.'s | **Checked** |
| Estimated Equation | **Checked** |
| ANOVA Detail | **Checked** |
| Residual Normality Tests | **Checked** |
| Residuals | **Checked** |

Plots Tab

Histogram ..... **Checked**
Probability Plot..... **Checked**
Residuals vs X..... **Checked**

### 3 Run the procedure

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

## Run Summary

**Run Summary**

| Item | Value | Rows | Value |
|---|---|---|---|
| Dependent Variable (Y) | IQ | Rows Processed | 17 |
| Number of Independent Variables (X) | 3 | Rows Used in Estimation | 15 |
| Weight Variable | None | Rows with X's Missing | 0 |
| $R^2$ | 0.1591 | Rows with Y Missing | 2 |
| Adjusted $R^2$ | 0.0000 | | |
| Coefficient of Variation | 0.1092 | | |
| Mean Square Error (MSE) | 129.9059 | | |
| Square Root of MSE | 11.39763 | | |
| Average \|Percent Error\| | 6.933 | | |
| Completion Status | Normal Completion | | |

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

### $R^2$

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

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

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

1. *Additional independent variables*. It is possible to increase $R^2$ by adding more independent variables, but the additional independent variables may actually cause an increase in the mean square error, an unfavorable situation. This case happens when your sample size is small.
2. *Range of the independent variables*. $R^2$ is influenced by the range of each independent variable. $R^2$ increases as the range of the *X's* increases and decreases as the range of the *X's* decreases.
3. *Slope magnitudes*. $R^2$ does not measure the magnitude of the slopes.

4. *Linearity*. $R^2$ does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *x* and *Y* was a perfect circle. The $R^2$ value of this relationship would be zero.
5. *Predictability*. A large $R^2$ does not necessarily mean high predictability, nor does a low $R^2$ necessarily mean poor predictability.
6. *No-intercept model*. The definition of $R^2$ assumes that there is an intercept in the regression model. When the intercept is left out of the model, the definition of $R^2$ changes dramatically. The fact that your $R^2$ value increases when you remove the intercept from the regression model does not reflect an increase in the goodness of fit. Rather, it reflects a change in the underlying meaning of $R^2$.
7. *Sample size*. $R^2$ is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

### Adjusted $R^2$

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

### Coefficient of Variation

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

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

### 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

**Descriptive Statistics**

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

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

## Correlation Matrix

**Correlation Matrix**

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

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

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

## Subset Selection Summary

**Subset Selection Summary**

| Number of | | $R^2$ | |
|---|---|---|---|
| Terms | X's | Value | Change |
| 1 | 1 | 0.1379 | 0.1379 |
| 2 | 2 | 0.1542 | 0.0163 |
| 3 | 3 | 0.1591 | 0.0049 |

This report shows the number of terms, number of IV's, and $R^2$ 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 $R^2$ increases only slightly as more variables are added.

## Subset Selection Detail

**Subset Selection Detail**

| | | Number of | | | Term | |
|---|---|---|---|---|---|---|
| Step | Action | Terms | X's | $R^2$ | Entered | 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 | 0.1591 | Test2 | |

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 Test2, Test3, and Test4.

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.

# Regression Coefficients T-Tests

**Regression Coefficient T-Tests**

| Independent Variable | Regression Coefficient b(i) | Standard Error Sb(i) | Standardized Coefficient | T-Test of H0: β(i) = 0 T-Statistic | P-Value | Reject H0 at α = 0.05? |
|---|---|---|---|---|---|---|
| Intercept | 75.93027 | 23.07606 | 0.0000 | 3.290 | 0.0072 | Yes |
| Test2 | -0.05858721 | 0.2324377 | -0.1032 | -0.252 | 0.8056 | No |
| Test3 | 0.1089272 | 0.2146191 | 0.1457 | 0.508 | 0.6218 | No |
| Test4 | 0.368076 | 0.3258485 | 0.4662 | 1.130 | 0.2827 | No |

This report gives the coefficients, standard errors, and significance tests.

## Independent Variable

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

## Regression Coefficient b(i)

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

## Standard Error Sb(i)

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

## Standardized Coefficient

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

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

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

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

## T-Statistic

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

### P-Value

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

### Reject H0 at α = 0.05?

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

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

## Regression Coefficients Confidence Intervals

**Regression Coefficient Confidence Intervals**

| Independent Variable | Regression Coefficient b(i) | Standard Error Sb(i) | 95% Confidence Interval Limits for β(i) Lower | Upper |
|---|---|---|---|---|
| Intercept | 75.93027 | 23.07606 | 25.14022 | 126.7203 |
| Test2 | -0.05858721 | 0.2324377 | -0.5701791 | 0.4530047 |
| Test3 | 0.1089272 | 0.2146191 | -0.3634463 | 0.5813006 |
| Test4 | 0.368076 | 0.3258485 | -0.3491118 | 1.085264 |

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

This report gives the coefficients, standard errors, and confidence interval.

### Independent Variable

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

### Regression Coefficient

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

### Standard Error Sb(i)

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

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

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

The formulas for the lower and upper confidence limits are:

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

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

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

## Estimated Equation

**Estimated Equation**

IQ =
75.9302747014515 - 0.0585872131040306 * Test2 + 0.108927169070947 * Test3 + 0.368076016587041 * Test4

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

## Analysis of Variance Detail

**Analysis of Variance Detail**

| Source | DF | $R^2$ Lost If Term(s) Removed | Sum of Squares | Mean Square | F-Ratio | P-Value |
|---|---|---|---|---|---|---|
| Intercept | 1 | | 163281.7 | 163281.7 | | |
| Model | 3 | 0.1591 | 270.3687 | 90.1229 | 0.694 | 0.5748 |
| Test2 | 1 | 0.0049 | 8.253181 | 8.253181 | 0.064 | 0.8056 |
| Test3 | 1 | 0.0197 | 33.46297 | 33.46297 | 0.258 | 0.6218 |
| Test4 | 1 | 0.0975 | 165.7572 | 165.7572 | 1.276 | 0.2827 |
| Error | 11 | 0.8409 | 1428.965 | 129.9059 | | |
| Total (Adjusted) | 14 | | 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.

### Source

This is the term from the design model.

### DF

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

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

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

### Sum of Squares

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

### Mean Square

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

### F-Ratio

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

### P-Value

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

## Residual Normality Tests

**Residual Normality Tests**

| | Test of H0: Residuals Normally Distributed | | |
|---|---|---|---|
| Test Name | Test Statistic Value | P-Value | Reject H0 at α = 0.2? |
| Shapiro-Wilk | 0.896 | 0.0833 | Yes |
| Anderson-Darling | 0.593 | 0.1220 | Yes |
| D'Agostino Skewness | 2.274 | 0.0230 | Yes |
| D'Agostino Kurtosis | 1.765 | 0.0775 | Yes |
| D'Agostino Omnibus | 8.287 | 0.0159 | 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.

# Residuals

**Residuals**

| Row | IQ Actual | IQ Predicted | Residual | Absolute Percent Error | Sqrt(MSE) Without This Row |
|---|---|---|---|---|---|
| 1 | 106 | 104.20740 | 1.79263600 | 1.69116600 | 11.934340 |
| 2 | 92 | 100.43640 | -8.43645000 | 9.17005400 | 11.420430 |
| 3 | 102 | 104.04170 | -2.04167900 | 2.00164700 | 11.931160 |
| 4 | 121 | 114.01510 | 6.98494300 | 5.77268000 | 11.512880 |
| 5 | 102 | 105.69710 | -3.69707800 | 3.62458600 | 11.888930 |
| 6 | 105 | 107.65240 | -2.65241200 | 2.52610700 | 11.919180 |
| 7 | 97 | 104.66650 | -7.66645100 | 7.90355800 | 11.679990 |
| 8 | 92 | 96.88332 | -4.88332200 | 5.30795900 | 11.757060 |
| 9 | 94 | 97.78729 | -3.78728900 | 4.02903100 | 11.850500 |
| 10 | 112 | 100.89120 | 11.10878000 | 9.91855100 | 11.142430 |
| 11 | 130 | 103.30070 | 26.69926000 | 20.53789000 | 7.294193 |
| 12 | 115 | 105.78880 | 9.21117100 | 8.00971400 | 11.553870 |
| 13 | 98 | 107.72240 | -9.72244000 | 9.92085600 | 11.470170 |
| 14 | 96 | 108.97380 | -12.97381000 | 13.51439000 | 10.973150 |
| 15 | 103 | 102.93590 | 0.06414504 | 0.06227674 | 11.953910 |
| 16 | | 96.85107 | | | |
| 17 | | 101.03520 | | | |

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

## Actual Y

This is the actual value of *Y*.

## Predicted Y

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

## Residual

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

## Absolute Percent Error

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

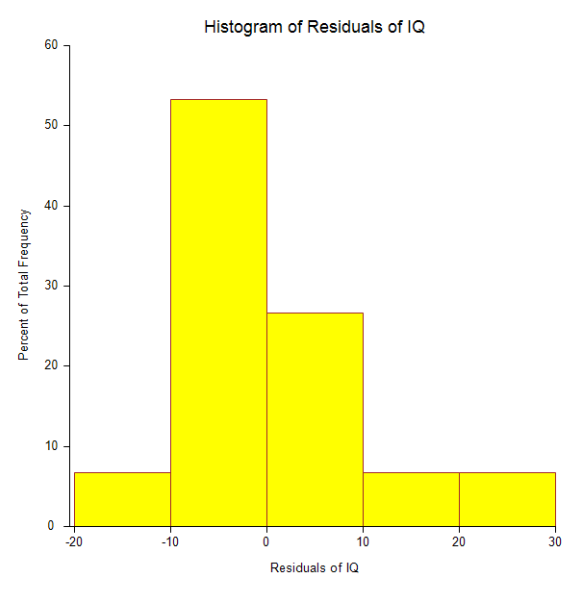
## Sqrt(MSE) Without This Row

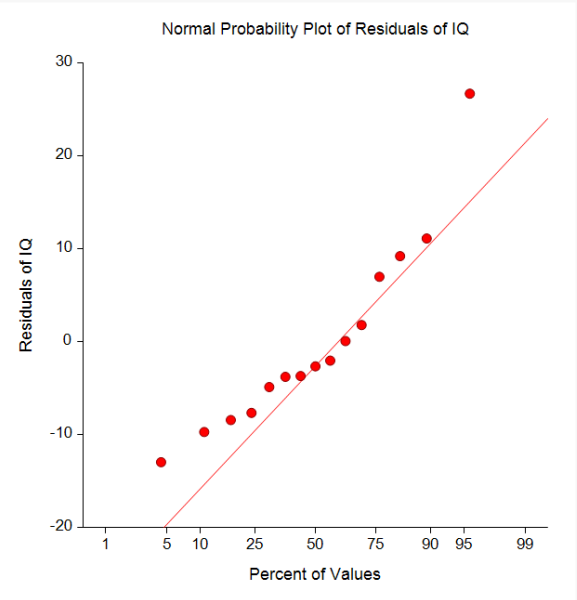
This is the value of the square root of the mean square error that is obtained if this row is deleted. A perusal of this statistic for all observations will highlight observations that have an inflationary impact on mean square error and could be outliers.

# Residual Plots

These plots let you assess the residuals. Any nonrandom pattern may require a redefining of the regression model.

### Residual Distribution Plots




### Residuals vs X Plots