Chapter 335

Ridge Regression

Introduction

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable. Another biased regression technique, principal components regression, is also available in **NCSS**. Ridge regression is the more popular of the two methods.

Multicollinearity

Multicollinearity, or collinearity, is the existence of near-linear relationships among the independent variables. For example, suppose that the three ingredients of a mixture are studied by including their percentages of the total. These variables will have the (perfect) linear relationship: P1 + P2 + P3 = 100. During regression calculations, this relationship causes a division by zero which in turn causes the calculations to be aborted. When the relationship is not exact, the division by zero does not occur and the calculations are not aborted. However, the division by a very small quantity still distorts the results. Hence, one of the first steps in a regression analysis is to determine if multicollinearity is a problem.

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 (and that's just for starters).

Sources of Multicollinearity

To deal with multicollinearity, you must be able to identify its source. The source of the multicollinearity 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 have been collected from a narrow subspace of the independent variables. The multicollinearity has been created by the sampling methodology—it does not exist in the population. Obtaining more data on an expanded range would cure this multicollinearity problem. The extreme example of this is when you try to fit a line to a single point.

2. *Physical constraints* of the linear model or population. This source of multicollinearity 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 multicollinearity.

3. *Over-defined model*. Here, there are more variables than observations. This situation should be avoided.

4. *Model choice or specification*. This source of multicollinearity comes from using independent variables that are powers or interactions of an original set of variables. It should be noted that if the sampling subspace of independent variables is narrow, then any combination of those variables will increase the multicollinearity problem even further.

5. *Outliers*. Extreme values or outliers in the *X*-space can cause multicollinearity as well as hide it. We call this outlier-induced multicollinearity. This should be corrected by removing the outliers before ridge regression is applied.

Detection of Multicollinearity

There are several methods of detecting multicollinearity. We mention a few.

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. Consider the variance inflation factors (VIF). VIFs over 10 indicate collinear variables.

3. Eigenvalues of the correlation matrix of the independent variables near zero indicate multicollinearity. Instead of looking at the numerical size of the eigenvalue, use the condition number. Large condition numbers indicate multicollinearity.

4. Investigate the signs of the regression coefficients. Variables whose regression coefficients are opposite in sign from what you would expect may indicate multicollinearity.

Correction for Multicollinearity

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

When these steps are not possible, you might try ridge regression.

Ridge Regression Models

Following the usual notation, suppose our regression equation is written in matrix form as

$$\underline{Y} = X\underline{B} + \underline{e}$$

where $\underline{Y}$ is the dependent variable, **X** represents the independent variables, $\underline{B}$ is the regression coefficients to be estimated, and $\underline{e}$ represents the errors are residuals.

Standardization

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation, since we must somehow indicate whether the variables in a particular formula are standardized or not. To keep the presentation simple, we will make the following general statement and then forget about standardization and its confusing notation.

As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is in a standardized scale.

Ridge Regression Basics

In ordinary least squares, the regression coefficients are estimated using the formula

$$\hat{\underline{B}} = (X'X)^{-1}X'\underline{Y}$$

Note that since the variables are standardized, **X'X** = **R**, where **R** is the correlation matrix of independent variables. These estimates are unbiased so that the expected value of the estimates are the population values. That is,

$$E(\hat{\underline{B}}) = \underline{B}$$

The variance-covariance matrix of the estimates is

$$V(\hat{\underline{B}}) = \sigma^2 \mathbf{R}^{-1}$$

and since we are assuming that the y's are standardized, $\sigma^2 = 1$.

From the above, we find that

$$V(\hat{b}_j) = r^{jj} = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared value obtained from regression X$_j$ on the other independent variables. In this case, this variance is the VIF. We see that as the R-squared in the denominator gets closer and closer to one, the variance (and thus VIF) will get larger and larger. The rule of thumb cut-off value for VIF is 10. Solving backwards, this translates into an R-squared value of 0.90. Hence, whenever the R-squared value between

one independent variable and the rest is greater than or equal to 0.90, you will have to face multicollinearity.

Now, ridge regression proceeds by adding a small value, k, to the diagonal elements of the correlation matrix. (This is where ridge regression gets its name since the diagonal of ones in the correlation matrix may be thought of as a ridge.) That is,

$$\underline{\widetilde{\mathbf{B}}} = (\mathbf{R} + k\mathbf{I})^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

k is a positive quantity less than one (usually less than 0.3).

The amount of bias in this estimator is given by

$$E(\underline{\widetilde{\mathbf{B}}} - \underline{\mathbf{B}}) = [(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X} - \mathbf{I}]\underline{\mathbf{B}}$$

and the covariance matrix is given by

$$V(\underline{\widetilde{\mathbf{B}}}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}$$

It can be shown that there exists a value of k for which the mean squared error (the variance plus the bias squared) of the ridge estimator is less than that of the least squares estimator. Unfortunately, the appropriate value of k depends on knowing the true regression coefficients (which are being estimated) and an analytic solution has not been found that guarantees the optimality of the ridge solution. We will discuss more about determining k later.

Alternative Interpretations of Ridge Regression

1. Ridge regression may be given a Bayesian interpretation. If we assume that each regression coefficient has expectation zero and variance *1/k*, then ridge regression can be shown to be the Bayesian solution.

2. Another viewpoint is referred to by detractors as the "phoney data" viewpoint. It can be shown that the ridge regression solution is achieved by adding rows of data to the original data matrix. These rows are constructed using 0 for the dependent variables and the square root of k or zero for the independent variables. One extra row is added for each independent variable. The idea that manufacturing data yields the ridge regression results has caused a lot of concern and has increased the controversy in its use and interpretation.

Choosing k

Ridge Trace

One of the main obstacles in using ridge regression is in choosing an appropriate value of *k*. Hoerl and Kennard (1970), the inventors of ridge regression, suggested using a graphic which they called the *ridge trace.* This plot shows the ridge regression coefficients as a function of *k*. When viewing the ridge trace, the analyst picks a value for *k* for which the regression coefficients have stabilized. Often, the regression coefficients will vary widely for small values of *k* and then stabilize. Choose the smallest value of *k* possible (which introduces the smallest bias) after which the regression coefficients have seem to remain constant. Note that increasing *k* will eventually drive the regression coefficients to zero. Following is an example of a ridge trace.
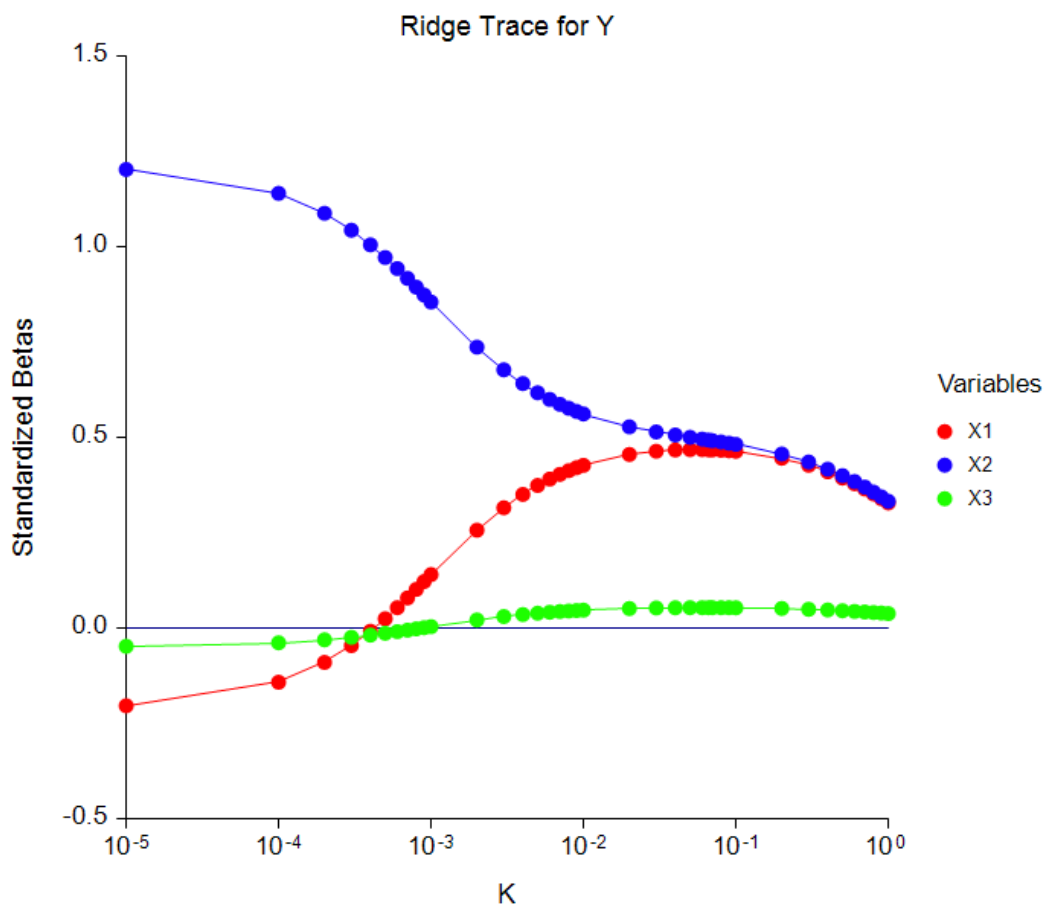
In this example, the values of *k* are shown on a logarithmic scale. We have drawn a vertical line at the selected value of *k* which is 0.006. A few notes are in order here.

First of all, the vertical axis contains the points for the least squares solution. These are labeled as 0.000001. This was done so that these coefficients may be seen. In actual fact, the logarithm of zero is minus infinity, so the least squares values cannot be displayed when the horizontal axis is put in a log scale.

We have displayed a large range of values. We see that adding *k* has little impact until *k* is about 0.0001. The action seems to stop somewhere near 0.006.

Analytic k

Hoerl and Kinnard (1976) proposed an iterative method for selecting k. This method is based on the formula

$$k = \frac{ps^2}{\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}}}$$

To obtain the first value of k, we use the least squares coefficients. This produces a value of k. Using this new k, a new set of coefficients is found, and so on. Unfortunately, this procedure does not necessarily converge. In **NCSS**, we have modified this routine so that if the resulting k is greater than one, the new value of k is equal to the last value of k divided by two.

This calculated value of k is often used because humans tend to pick a k from the ridge trace that is too large.

Assumptions

The assumptions are the same as those used in regular multiple regression: linearity, constant variance (no outliers), and independence. Since ridge regression does not provide confidence limits, normality need not be assumed.

What the Professionals Say

Ridge regression remains controversial. In this section we will present the comments made in several books on regression analysis.

Neter, Wasserman, and Kutner (1983) state:

"Ridge regression estimates tend to be stable in the sense that they are usually little affected by small changes in the data on which the fitted regression is based. In contrast, ordinary least squares estimates may be highly unstable under these conditions when the independent variables are highly multicollinear. Also, the ridge estimated regression function at times will provide good estimates of mean responses or predictions of new observations for levels of the independent variables outside the region of the observations on which the regression function is based. In contrast, the estimated regression function based on ordinary least squares may perform quite poorly in such instances. Of course, any estimation or prediction well outside the region of the observations should always be made with great caution.

"A major limitation of ridge regression is that ordinary inference procedures are not applicable and exact distributional properties are not known. Another limitation is that the choice of the biasing constant k is a judgmental one. While formal methods have been developed fo making this choice, these methods have their own limitations."

John O. Rawlings (1988) states:

"While multicollinearity does not affect the precision of the estimated responses (and predictions) at the observed points in the X-space, it does cause variance inflation of estimated responses at other points. Park shows that the restrictions on the parameter estimates implicit in principal component regression are also optimal in MSE sense for estimation of responses over certain regions of the X-space. This suggests that biased regression methods may be beneficial in certain cases for estimation of responses also. The biased regression methods do not seem to have much to offer when the objective is to assign some measure of "relative importance" to the independent variables involved in a multicollinearity… Ridge regression attacks the multicollinearity by reducing the apparent magnitude of the correlations."

Raymond H. Myers (1990) states:

"Ridge regression is one of the more popular, albeit controversial, estimation procedures for combating multicollinearity. The procedures discussed in this and subsequent sections fall into the category of biased estimation techniques. They are based on this notion: though ordinary least squares gives unbiased estimates and indeed enjoy the minimum variance of all linear unbiased estimators, there is no upper bound on the variance of the estimators and the presence of multicollinearity may produce large variances. As a result, one can visualize that, under the condition of multicollinearity, a huge price is paid for the unbiasedness property that one achieves by using ordinary least squares. Biased estimation is used to attain a substantial reduction in variance with an accompanied increase in stability of the regression coefficients. The coefficients become biased and, simply put, if one is successful, the reduction in variance is of greater magnitude than the bias induced in the estimators…

"Although, clearly, ridge regression should not be used to solve all model-fitting problems involving multicollinearity, enough positive evidence about ridge regression exists to suggest that it should be a part of any model builder's arsenal of techniques."

Draper and Smith (1981) state:

"From this discussion, we can see that the use of ridge regression is perfectly sensible in circumstances in which it is believed that large beta-values are unrealistic from a practical point of view. However, it must be realized that the choice of *k* is essentially equivalent to an expression of how big one believes those betas to be. In circumstances where one cannot accept the idea of restrictions on the betas, ridge regression would be completely inappropriate."

"Overall, however, we would advise against the indiscriminate use of ridge regression unless its limitations are fully appreciated."

Thomas Ryan (1997) states:

"The reader should note that, for all practical purposes, the ordinary least squares (OLS) estimator will also generally be biased because we can be certain that it is unbiased only when the model that is being used is the correct model. Since we cannot expect this to be true, we similarly cannot expect the OLS estimator to be unbiased. Therefore, although the choice between OLS and a ridge estimator is often portrayed as a choice between a biased estimator and an unbiased estimator, that really isn't the case."

"Ridge regression permits the use of a set of regressors that might be deemed inappropriate if least squares were used. Specifically, highly correlated variables can be used together, with ridge regression used to reduce the multicollinearity. If, however, the multicollinearity were extreme, such as when regressors are almost perfectly correlated, we would probably prefer to delete one or more regressors before using the ridge approach."

Data Structure

The data are entered as three or more variables. One variable represents the dependent variable. The other variables represent the independent variables. An example of data appropriate for this procedure is shown below. These data were concocted to have a high degree of multicollinearity as follows. We put a sequence of numbers in X1. Next, we put another series of numbers in X3 that were selected to be unrelated to X1. We created X2 by adding X1 and X3. We made a few changes in X2 so that there was not perfect correlation. Finally, we added all three variables and some random error to form Y.

The data are contained in the RidgeReg dataset. We suggest that you open this dataset now so that you can follow along with the example.

RidgeReg Dataset (Subset)

X1	X2	X3	Y
1	2	1	3
2	4	2	9
3	6	4	11
4	7	3	15
5	7	2	13
6	7	1	13
7	8	1	17
8	10	2	21
9	12	4	25
10	13	3	27

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 is generated for that row.

Example 1 – Ridge Regression Analysis

This section presents an example of how to run a ridge regression analysis of the data presented earlier in this chapter. The data are in the RidgeReg dataset. In this example, we will run a regression of *Y* on *X1 - X3*.

Setup

To run this example, complete the following steps:

1   Open the RidgeReg example dataset
- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **RidgeReg** and click **OK**.

2   Specify the Ridge Regression procedure options
- Find and open the **Ridge 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 Tab

Y: Dependent Variable......................................**Y**
X's: Independent Variables..............................**X1-X3**
Final K (On Reports)..**Optimum** (This will cause the optimum value found in the
　　　　　　　　　　　　　　　　　　　　　　　search procedure to be used in all of the reports.)
0.0001 to 0.0009...**Checked** (to include these values of *k*.)
K Search...**Checked** (so that the optimal value of *k* will be found.)

Reports Tab

All Reports and Plots ....**Checked** (All are selected so they can be documented here.)

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

Descriptive Statistics Section

Descriptive Statistics Section

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
X1	18	9.5	5.338539	1	18
X2	18	11.5	5.404247	2	19
X3	18	2.166667	1.098127	1	4
Y	18	23.11111	10.87841	3	39

For each variable, the descriptive statistics of the nonmissing values are computed. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix Section

Correlation Matrix Section

	X1	X2	X3	Y
X1	1.000000	0.987841	-0.015051	0.985544
X2	0.987841	1.000000	0.133813	0.995574
X3	-0.015051	0.133813	1.000000	0.116539
Y	0.985544	0.995574	0.116539	1.000000

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 multicollinearity problems.

Least Squares Multicollinearity Section

Least Squares Multicollinearity Section

Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance
X1	477.2665	0.9979	0.0021
X2	485.8581	0.9979	0.0021
X3	11.7455	0.9149	0.0851

Since some VIF's are greater than 10, multicollinearity is a problem.

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 independent variables. A VIF of 10 or more for large data sets indicates a multicollinearity problem since the R_x^2 with the remaining X's is 90 percent. For small data sets, even VIF's of 5 or more can signify multicollinearity.

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

R-Squared vs Other X'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 1- R_x^2, the denominator of the variance inflation factor.

Eigenvalues of Correlations

Eigenvalues of Correlations

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Number
1	1.994969	66.50	66.50	1.00
2	1.004003	33.47	99.97	1.99
3	0.001027	0.03	100.00	1941.85

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

This section gives an eigenvalue analysis of the independent variables after they have been centered and scaled. Notice that in this example, the third eigenvalue is very small.

Eigenvalue

The eigenvalues of the correlation matrix. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero indicate a multicollinearity problem 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 indicate a multicollinearity problem 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. Condition numbers greater than 1000 indicate a severe multicollinearity problem while condition numbers between 100 and 1000 indicate a mild multicollinearity problem.

Eigenvector of Correlations

Eigenvector of Correlations

No.	Eigenvalue	X1	X2	X3
1	1.994969	0.701391	0.707741	0.084573
2	1.004003	-0.134162	0.014553	0.990853
3	0.001027	0.700036	-0.706322	0.105159

This report displays the eigenvectors associated with each eigenvalue. The notion behind eigenvalue analysis is that the axes are rotated from the ones defined by the variables to a new set defined by the variances of the variables. Rotating is accomplished by taking weighted averages of the original variables. Thus, the first new axis could be the average of X1 and X2. The first new variable is constructed to account for the largest amount of variance possible from a single axis.

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 multicollinearity in your data. The eigenvalues represent the spread (variance) in the direction defined by this new axis. Hence, small eigenvalues indicate directions in which there is no spread. Since regression analysis seeks to find trends across values, when there is not a spread, the trends cannot be computed.

Table-Values

The table values give the eigenvectors. The eigenvectors give the weights that are used to create the new axis. By studying the weights, you can gain an understanding of what is happening in the data.

In the example above, we can see that the first factor (new variable associated with the first eigenvalue) is constructed by adding X1 and X2. Note that the weights are almost equal. X3 has a small weight, indicating that it does not play a role in this factor.

Factor 2 seems to be created complete from X3. X1 and X2 play only a small role in its construction.

Factor 3 seems to be the difference between X1 and X2. Again, X3 plays only a small role. Hence, the interpretation of these eigenvectors leads to the following statements:

1. Most of the variation in X1, X2, and X3 can be accounted for by considering only two variables: Z = X1+X2 and X3.

2. The third dimension, calculated as X1-X2, is almost negligible and might be ignored.

# Ridge Trace Plot
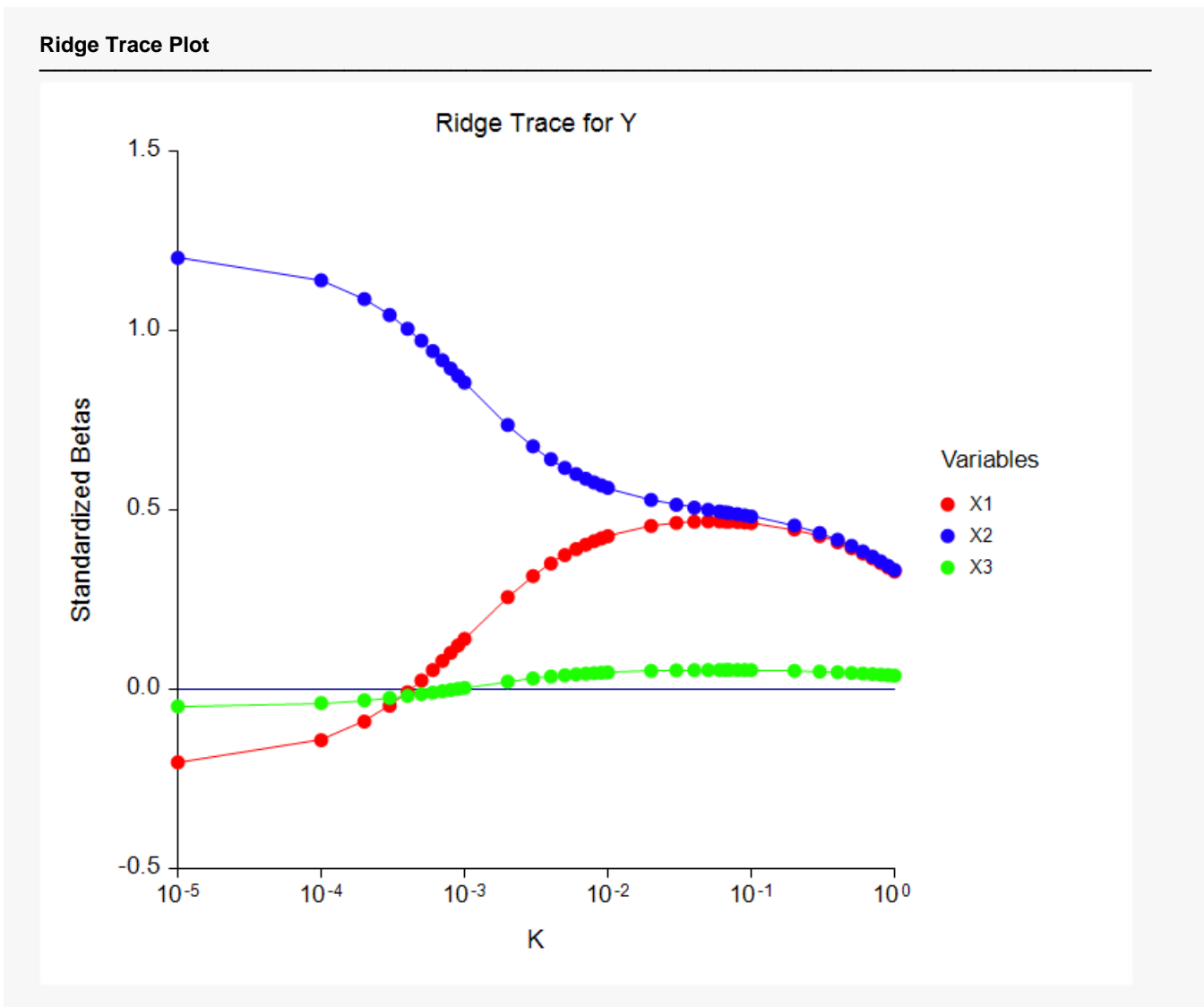
Ridge Trace Plot
_____

This is the famous *ridge trace* that is the signature of this technique. The plot is really very straight forward to read. It presents the standardized regression coefficients on the vertical axis and various values of *k* along the horizontal axis. Since the values of *k* span several orders of magnitude, we adopt a logarithmic scale along this axis.

The points on the left vertical axis (the left ends of the lines) are the ordinary least squares regression values. These occur for *k* equal zero. As *k* is increased, the values of the regression estimates change, often wildly at first. At some point, the coefficients seem to settle down and then gradually drift towards zero.

The task of the ridge regression analyst is to determine at what value of *k* these coefficients are at their stable values. A vertical line is drawn at the value selected for reporting purposes. It is anticipated that you would run the program several times until an appropriate value of *k* is determined. In this example, our search would be between 0.0001 and 0.1. The value selected on this graph happens to be 0.066237, the value obtained from the analytic search. We might be inclined to use an even smaller value of *k* such as 0.01. Remember, the smaller the value of *k*, the smaller the amount of bias that is included in the estimates.

Variance Inflation Factor Plot

Variance Inflation Factor Plot
_____
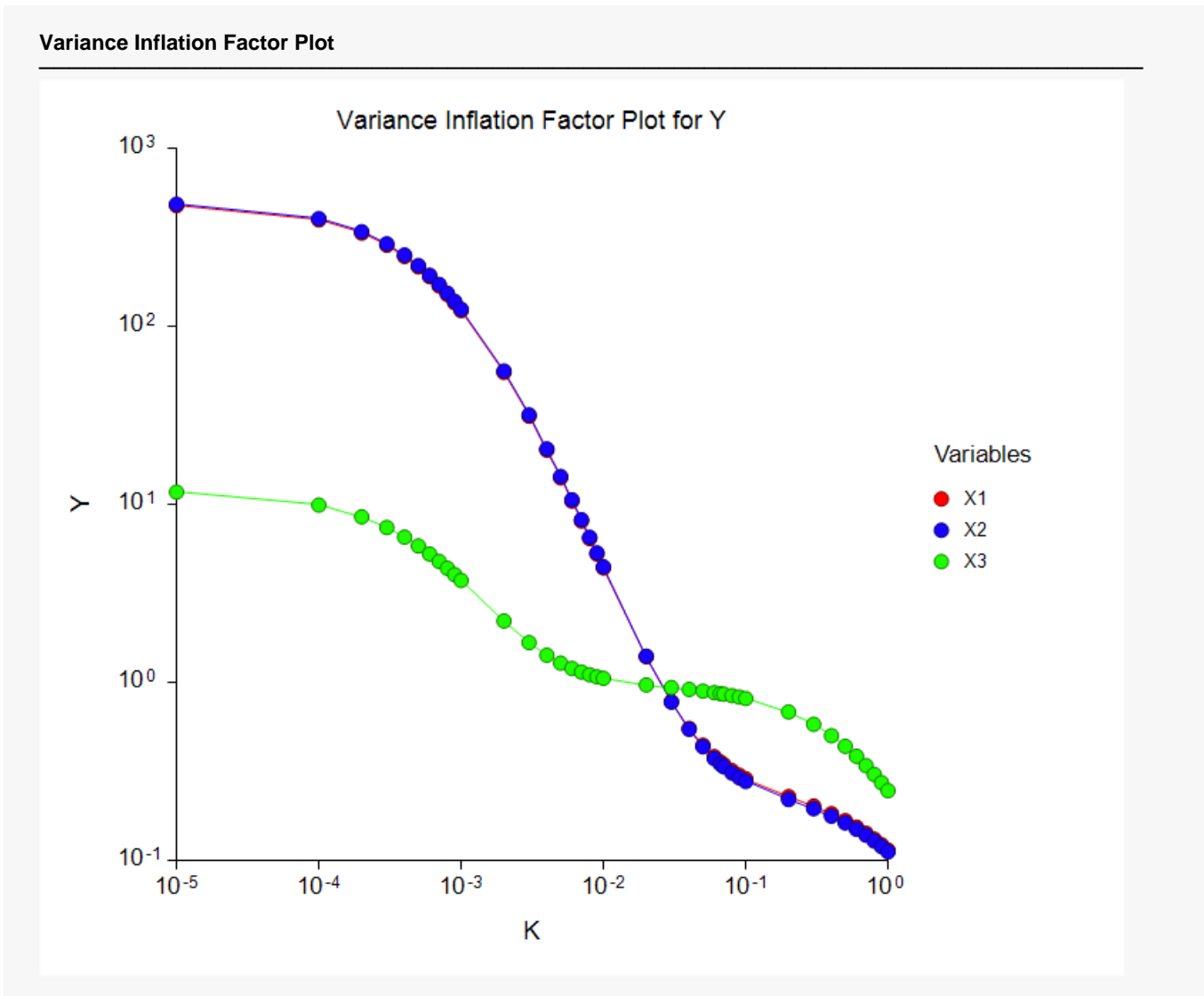
This is a plot that we have added that shows the impact of k on the variance inflation factors. Since the major goal of ridge regression is to remove the impact of multicollinearity, it is important to know at what point multicollinearity has been dealt with. This plot shows this.

The currently selected value of k is shown by a vertical line.

Since the rule-of-thumb is that multicollinearity is not a problem once all VIFs are less than 10, we inspect the graph for this point. In this example, it appears that all VIFs are small enough once k is greater than 0.007. Hence, this is the value of k that this plot would indicate we use.

Since this plot indicates $k = 0.007$ and the ridge trace indicates a value near 0.01, we would select 0.007 as our final result. The rest of the reports are generated for this value of k.

Standardized Ridge Regression Coefficients Section

Standardized Ridge Regression Coefficients Section

k	X1	X2	X3
0.000000	-0.2034	1.2029	-0.0475
0.000100	-0.1415	1.1404	-0.0382
0.000200	-0.0897	1.0881	-0.0304
0.000300	-0.0457	1.0436	-0.0238
0.000400	-0.0079	1.0054	-0.0181
0.000500	0.0249	0.9722	-0.0132
0.000600	0.0538	0.9431	-0.0088
0.000700	0.0793	0.9173	-0.0050
0.000800	0.1019	0.8944	-0.0016
0.000900	0.1223	0.8738	0.0015
0.001000	0.1406	0.8553	0.0042
.	.	.	.
.	.	.	.
.	.	.	.

This report gives the values that are plotted on the ridge trace. Note that the value found by the analytic search (0.066237) sticks out as you glance down the first column because it does not end in zeros.

Variance Inflation Factor Section

Variance Inflation Factor Section

k	X1	X2	X3
0.000000	477.2665	485.8581	11.7455
0.000100	396.3965	403.5292	9.9204
0.000200	334.4756	340.4914	8.5229
0.000300	286.0151	291.1566	7.4291
0.000400	247.3784	251.8230	6.5570
0.000500	216.0793	219.9592	5.8505
0.000600	190.3708	193.7870	5.2702
0.000700	168.9966	172.0273	4.7877
0.000800	151.0345	153.7411	4.3822
0.000900	135.7951	138.2268	4.0381
0.001000	122.7546	124.9510	3.7436
.	.	.	.
.	.	.	.
.	.	.	.

This report gives the values that are plotted on the variance inflation factor plot. Note how easy it is to determine when all three VIFs are less than 10.

K Analysis Section

K Analysis Section

k	R2	Sigma	B'B	Ave VIF	Max VIF
0.000000	0.9915	1.1028	1.4905	324.9567	485.8581
0.000100	0.9914	1.1119	1.3219	269.9487	403.5292
0.000200	0.9913	1.1199	1.1929	227.8300	340.4914
0.000300	0.9912	1.1272	1.0918	194.8669	291.1566
0.000400	0.9911	1.1339	1.0113	168.5862	251.8230
0.000500	0.9910	1.1401	0.9460	147.2964	219.9592
0.000600	0.9909	1.1459	0.8924	129.8093	193.7870
0.000700	0.9908	1.1513	0.8478	115.2705	172.0273
0.000800	0.9907	1.1565	0.8103	103.0526	153.7411
0.000900	0.9906	1.1614	0.7785	92.6867	138.2268
0.001000	0.9905	1.1661	0.7513	83.8164	124.9510
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This report provides a quick summary of the various statistics that might go into the choice of k.

k

This is the actual value of k. Note that the value found by the analytic search (0.066237) sticks out as you glance down this column because it does not end in zeros.

R2

This is the value of R-squared. Since the least squares solution maximizes R-squared, the largest value of R-squared occurs when k is zero. We want to select a value of k that does not stray very much from this value.

Sigma

This is the square root of the mean squared error. Least squares minimizes this value, so we want to select a value of k that does not stray very much from the least squares value.

B'B

This is the sum of the squared standardized regression coefficients. Ridge regression assumes that this value is too large and so the method tries to reduce this. We want to find a value for k at which this value has stabilized.

Ave VIF

This is the average of the variance inflation factors.

Max VIF

This is the maximum variance inflation factor. Since we are looking for that value of k which results in all VIFs being less than 10, this value is very helpful in your selection of k.

Ridge vs. Least Squares Comparison Section

This report provides a detailed comparison between the ridge regression solution and the ordinary least squares solution to the estimation of the regression coefficients.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of b_0.

Regular Ridge (and L.S.) Coeff's

These are the estimated values of the regression coefficients b_0, b_1, ..., b_p. The first column gives the values for ridge regression and the second column gives the values for regular least squares regression.

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 also called partial-regression coefficients since the effect of the other *X's* is removed.

Stand'zed Ridge (and L.S.) Coeff's

These are the estimated values of the standardized regression coefficients. The first column gives the values for ridge regression and the second column gives the values for regular least squares regression.

Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and 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 there are vastly different units involved for the variables, this is a way of making comparisons between 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 corresponding j^{th} independent variable.

Ridge (and L.S.) Standard Error

These are the estimated standard errors (precision) of the regression coefficients. The first column gives the values for ridge regression and the second column gives the values for regular least squares regression.

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. Since one of the objects of ridge regression is to reduce this (make the estimates more precise), it is of interest to see how much reduction has taken place.

R-Squared

R-squared is the coefficient of determination. It represents the percent of variation in the dependent variable explained by the independent variables in the model. The R-squared values of both the ridge and regular regressions are shown.

Sigma

This is the square root of the mean square error. It provides a measure of the standard deviation of the residuals from the regression model.

 It represents the percent of variation in the dependent variable explained by the independent variables in the model. The R-squared values of both the ridge and regular regressions are shown.

Ridge Regression Coefficient Section

Ridge Regression Coefficient Section for k = 0.066237

Independent Variable	Regression Coefficient	Standard Error	Standardized Regression Coefficient	VIF
Intercept	1.471949			
X1	0.9535758	0.06622486	0.4680	0.3581
X2	0.9931737	0.06455679	0.4934	0.3487
X3	0.5347818	0.5008289	0.0540	0.8664

Model

1.471949 + 0.9535758*X1 + 0.9931737*X2 + 0.5347818*X3

This report provides the details of the ridge regression solution.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of b_0.

Regression Coefficient

These are the estimated values of the regression coefficients b_0, b_1, ..., b_p. 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 also called partial-regression coefficients since the effect of the other X's is removed.

Standard Error

These are the estimated standard errors (precision) of the ridge regression coefficients. The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. In regular regression, we divide the coefficient by the standard error to obtain a t statistic. However, this is not possible here because of the bias in the estimates.

Stand'zed Regression Coefficient

These are the estimated values of the standardized regression coefficients. Standardized regression coefficients are the coefficients that would be obtained if you standardized each independent and 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 there are vastly different units involved for the variables, this is a way of making comparisons between 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 corresponding j^{th} independent variable.

VIF

These are the values of the variance inflation factors associated with the variables. When multicollinearity has been conquered, these values will all be less than 10. Details of VIF were given earlier.

Analysis of Variance Section

Analysis of Variance Section for k = 0.066237

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Intercept	1	9614.223	9614.223		
Model	3	1928.694	642.8979	108.3308	0.000000
Error	14	83.08412	5.93458		
Total(Adjusted)	17	2011.778	118.3399		

Additional Model Information

Mean of Dependent	23.11111
Root Mean Square Error	2.436099
R-Squared	0.9587
Coefficient of Variation	0.1054081

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

Source

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

DF

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

Mean Square

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

F-Ratio

This is the F statistic for testing the null hypothesis that all β_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.

Since ridge regression produces biased estimates, this F-Ratio is not a valid test. It serves as an index, but it would not stand up under close scrutiny.

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.

Root Mean Square Error

This is the square root of the mean square error. It is an estimate of σ, the standard deviation of the e_i's.

Mean of Dependent Variable

This is the arithmetic mean of the dependent variable.

R-Squared

This is the coefficient of determination. It is defined in full in the Multiple Regression chapter.

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.

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

Predicted Values and Residuals Section

Predicted Values and Residuals Section for k = 0.066237

Row	Actual	Predicted	Residual
1	3	4.946654	-1.946654
2	9	8.421359	0.5786409
3	11	12.43085	-1.430846
4	15	13.84281	1.157186
5	13	14.26161	-1.261608
6	13	14.6804	-1.680402
7	17	16.62715	0.3728486
8	21	20.10186	0.8981435
9	25	24.11134	0.8886567
10	27	25.52331	1.476689
11	25	25.9421	-0.9421052
12	27	26.3609	0.6391008
13	29	28.30765	0.6923512
14	33	31.78235	1.217646
15	35	35.79184	-0.7918407
16	37	37.20381	-0.2038085
17	37	37.6226	-0.6226026
18	39	38.0414	0.9586034

This section reports the predicted values and the sample residuals, or e_i's. When you want to generate predicted values for individuals not in your sample, add their values to the bottom of your database, leaving the dependent variable blank. Their predicted values will be shown on this report.

Actual

This is the actual value of Y for the ith row.

Predicted

The predicted value of Y for the ith row. It is predicted using the levels of the $X's$ for this row.
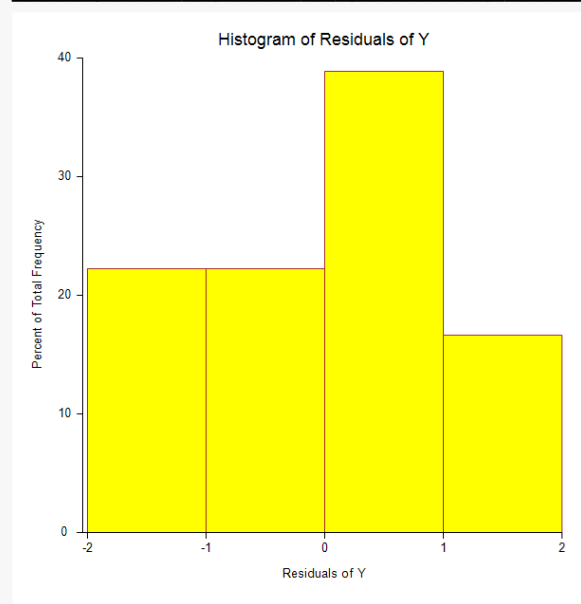
Residual

This is the estimated value of e_i. This is equal to the *Actual* minus the *Predicted.*

Histogram

The purpose of the histogram and density trace of the residuals is to display the distribution of the residuals.
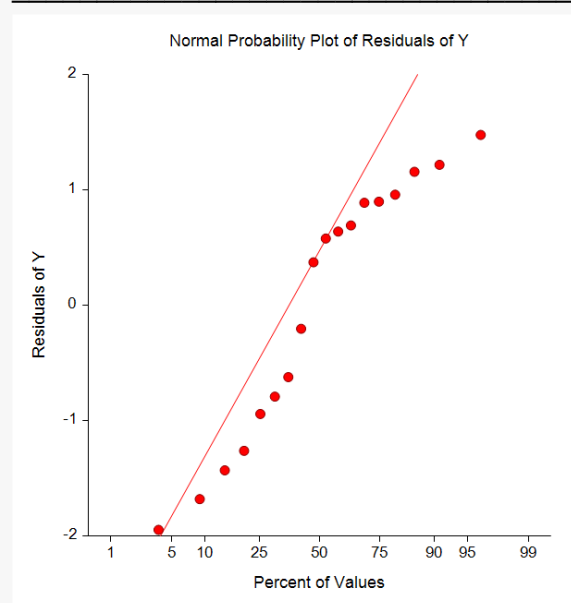
Distributional Plots of Residuals

The odd shape of this histogram occurs because of the way in which these particular data were manufactured.

Probability Plot of Residuals

Distributional Plots of Residuals

Residuals vs Yhat 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. A sloping or curved band signifies inadequate specification of the model. A sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.

Residuals vs Yhat Plot
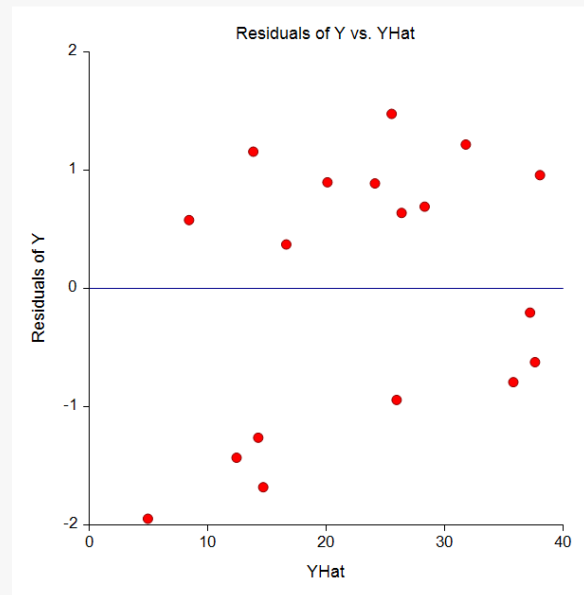
Residuals vs X's Plots

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.

Residuals vs X's Plots