Chapter 340

Principal Components Regression

Introduction

Principal Components 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, principal components regression reduces the standard errors. It is hoped that the net effect will be to give more reliable estimates. Another biased regression technique, ridge regression, is also available in **NCSS**. Ridge regression is the more popular of the two methods.

Multicollinearity

Multicollinearity is discussed both in the Multiple Regression chapter and in the Ridge Regression chapter, so we will not repeat the discussion here. However, it is important to understand the impact of multicollinearity so that you can decide if some evasive action (like pc regression) would be beneficial.

Principal Components Regression Models

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

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

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

Standardization

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 calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back to their original scale.

PC Regression Basics

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

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

Note that since the variables are standardized, **X'X** = **R**, where **R** is the correlation matrix of independent variables.

To perform principal components (PC) regression, we transform the independent variables to their principal components. Mathematically, we write

$$\mathbf{X}'\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}' = \mathbf{Z}'\mathbf{Z}$$

where **D** is a diagonal matrix of the eigenvalues of **X'X**, **P** is the eigenvector matrix of **X'X**, and **Z** is a data matrix (similar in structure to **X**) made up of the principal components. **P** is orthogonal so that **P'P** = **I**.

We have created new variables **Z** as weighted averages of the original variables **X**. This is nothing new to us since we are used to using transformations such as the logarithm and the square root on our data values prior to performing the regression calculations. Since these new variables are principal components, their correlations with each other are all zero. If we begin with variables X1, X2, and X3, we will end up with Z1, Z2, and Z3.

Severe multicollinearity will be detected as very small eigenvalues. To rid the data of the multicollinearity, we omit the components (the z's) associated with small eigenvalues. Usually, only one or two relatively small eigenvalues will be obtained. For example, if only one small eigenvalue were detected on a problem with three independent variables, we would omit Z3 (the third principal component).

When we regress **Y** on Z1 and Z2, multicollinearity is no longer a problem. We can then transform our results back to the **X** scale to obtain estimates of **B**. These estimates will be biased, but we hope that the size of this bias is more than compensated for by the decrease in variance. That is, we hope that the mean squared error of these estimates is less than that for least squares.

Mathematically, the estimation formula becomes

$$\underline{\widehat{\mathbf{A}}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\underline{\mathbf{Y}} = \mathbf{D}^{-1}\mathbf{Z}'\underline{\mathbf{Y}}$$

because of the special nature of principal components. Notice that this is ordinary least squares regression applied to a different set of independent variables.

The two sets of regression coefficients, **A** and **B**, are related using the formulas

$$\underline{\mathbf{A}} = \mathbf{P}'\underline{\mathbf{B}}$$

and

$$\underline{\mathbf{B}} = \mathbf{P}\underline{\mathbf{A}}$$

Omitting a principal component may be accomplished by setting the corresponding element of **A** equal to zero. Hence, the principal components regression may be outlined as follows:

1. Complete a principal components analysis of the **X** matrix and save the principal components in **Z**.
2. Fit the regression of **Y** on **Z** obtaining least squares estimates of **A**.
3. Set the last element of **A** equal to zero.
4. Transform back to the original coefficients using **B** = **PA**.

Alternative Interpretation of PC Regression

It can be shown that omitting a principal component amounts to setting a linear constraint on the regression coefficients. That is, in the case of three independent variables, we add the constraint

$$p_{13}b_1 + p_{23}b_2 + p_{33}b_3 = 0$$

Note that this is a constraint on the coefficients, not a constraint on the dependent variable. Essentially, we have avoided the multicollinearity problem by avoiding the region of the solution space in which it occurs.

How Many PC's Should Be Omitted

Unlike the selection of *k* in ridge regression, the selection of the number of PC's to omit is relatively straight forward. We omit the PC's corresponding to small eigenvalues. Since the size of the typical eigenvalue of a correlation matrix is one, we omit those that are much smaller than one. Usually, the choice will be obvious.

Assumptions

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

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 database 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 – Principal Components Regression

This section presents an example of how to run a principal components regression analysis of the data presented above. 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 Principal Components Regression procedure options

- Find and open the **Principal Components 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**
Reports Tab	
All Reports	**Checked** (All are selected here for documentation purposes.)
Plots Tab	
All Plots	**Checked** (Some of the aspects of the plot axes may be modified for improved viewing.)

3 Run the procedure

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

Descriptive Statistics

Descriptive Statistics

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

Correlation Matrix

	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

Least Squares Multicollinearity

Independent Variable	Variance Inflation Factor (VIF)	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 Factor (VIF)

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-squared 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.

Eigenvalues of Correlations

Eigenvalues of Correlations

		Percent		
Number	Eigenvalue	Incremental	Cumulative	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.

Eigenvectors of Correlations

Eigenvectors of Correlations

Number	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 those defined by the variables to a new set defined by the variances of the variables. Rotation is accomplished by taking weighted averages of the standardized original variables. The first new variable is constructed to account for the largest amount of variance possible from a single axis.

Number

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 indicate 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 accurately.

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.

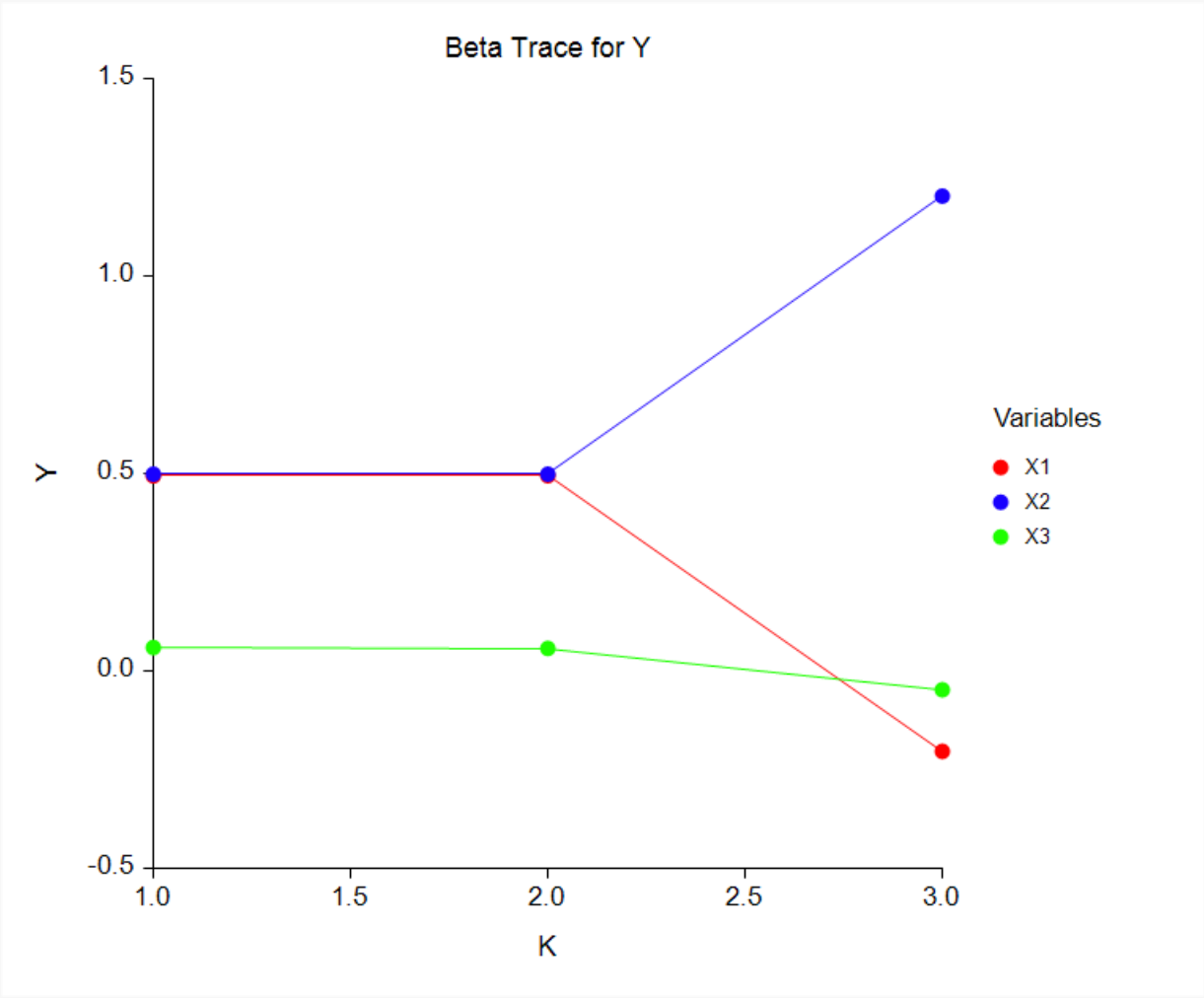
Factor 2 seems to be completely created from X3. X1 and X2 play only a small role in its construction.

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.

Beta Trace Plot

Beta Trace Plot

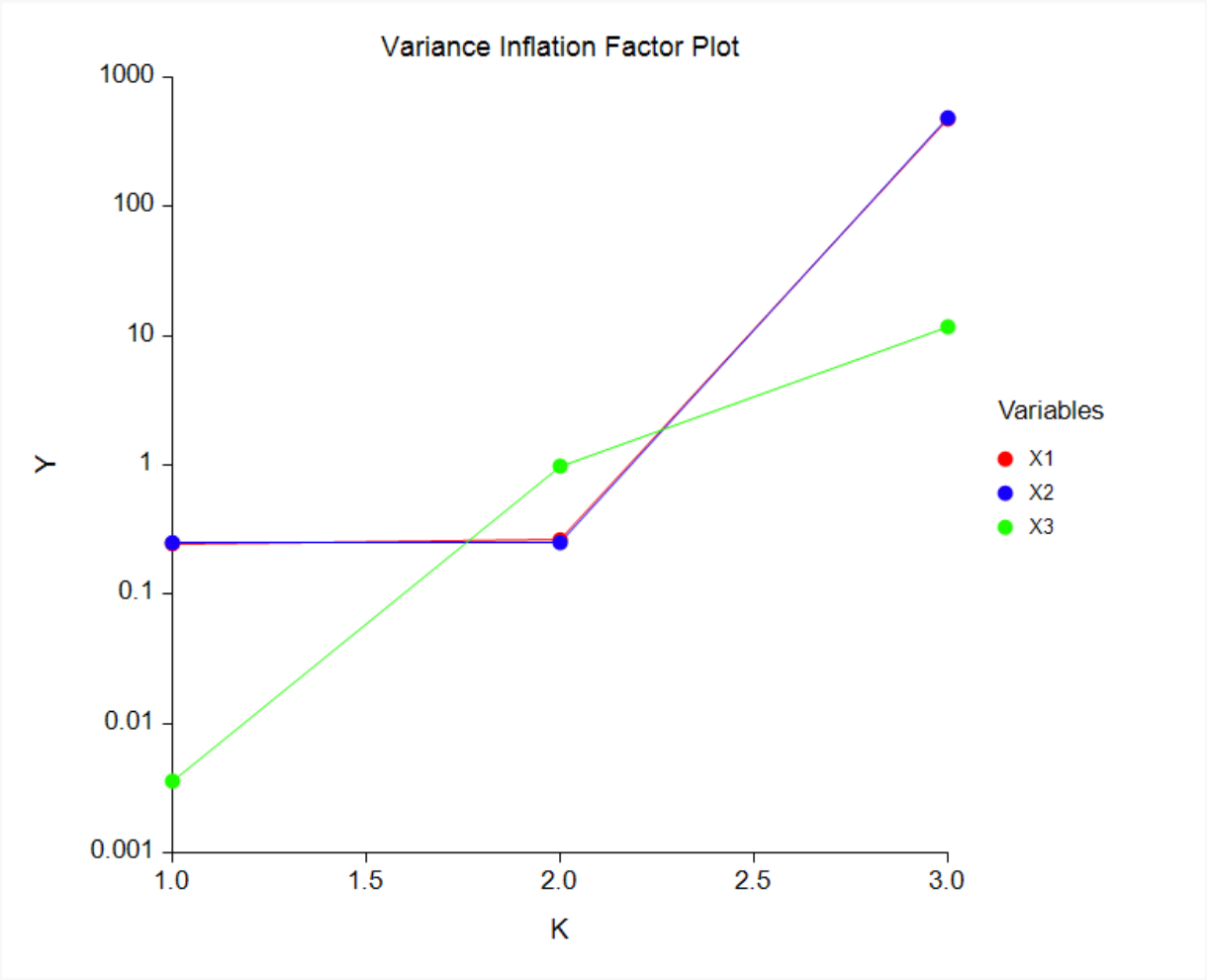


This plot shows the standardized regression coefficients (often referred to as the betas) on the vertical axis and the number of principal components (PC's) included along the horizontal axis. Thus, the set on the right is the least squares set.

By studying this plot, you can determine what omitting a certain number of PC's has done to the estimated regression coefficients.

Variance Inflation Factor Plot

Variance Inflation Factor Plot



This is a plot that shows the effect of the omitted PC's on the variance inflation factors. Since the major goal of PC 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.

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 less than 10 if only two of the three PC's are included.

Standardized Principal Components Regression Coefficients

Standardized Principal Components Regression Coefficients

# of PC's	X1	X2	X3
1	0.4942	0.4987	0.0596
2	0.4945	0.4987	0.0574
3	-0.2034	1.2029	-0.0475

This report gives the values that are plotted on the beta trace.

Variance Inflation Factors

Variance Inflation Factors

# of PC's	X1	X2	X3
1	0.2466	0.2511	0.0036
2	0.2645	0.2513	0.9815
3	477.2665	485.8581	11.7455

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.

Principal Components Analysis

Principal Components Analysis

				Variance Inflation Factor (VIF)	
# of PC's	R-Squared	SSE	B'B	Average	Maximum
1	0.9905	1.1677	0.4965	0.1671	0.2511
2	0.9905	1.1674	0.4965	0.4991	0.9815
3	0.9915	1.1028	1.4905	324.9567	485.8581

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

of PC's

This is the number of principal components included in the regression reported on this line.

R-Squared

This is the value of R-squared. Since the least squares solution maximizes R-squared, the largest value of R-squared occurs at bottom of the report (when all PC's are included).

Sigma

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

B'B

This is the sum of the squared standardized regression coefficients. PC regression assumes that this value is too large and so the method tries to reduce this. We want to find the number of PC's at which this value has stabilized.

Average Variance Inflation Factor (VIF)

This is the average of the variance inflation factors.

Maximum Variance Inflation Factor (VIF)

This is the maximum variance inflation factor. Since we are looking for the number of PC's which results in all VIFs being less than 10, this value is very helpful.

Principal Components vs. Least Squares Regression Comparison

Principal Components vs. Least Squares Regression Comparison with 1 Component Omitted

	Regular Coefficients		Standardized Coefficients		Standard Error	
Independent Variable	Principal Components	Least Squares	Principal Components	Least Squares	Principal Components	Least Squares
Intercept	0.763326	0.2230599				
X1	1.007698	-0.4144863	0.4945	-0.2034	0.0272776	1.094502
X2	1.003778	2.421286	0.4987	1.2029	0.02626337	1.090883
X3	0.568248	-0.4703622	0.0574	-0.0475	0.2554352	0.8347205
R-Squared	0.9905	0.9915				
Sigma	1.1674	1.1028				

This report provides a detailed comparison between the PC 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 Principal Components and Least Squares Coefficients

These are the estimated values of the regression coefficients $b_0, b_1, ..., b_p$. The first column gives the values for PC 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.

Standardized Principal Components and Least Squares Coefficients

These are the estimated values of the standardized regression coefficients. The first column gives the values for PC 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, respectively.

Principal Components and Least Squares Standard Error

These are the estimated standard errors (precision) of the regression coefficients. The first column gives the values for PC 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 PC 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 PC and regular regressions are shown.

Sigma

This is the square root of the mean squared 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 PC and regular regressions are shown.

Principal Components Regression Coefficients

Principal Components Regression Coefficients

Principal Component	Regression Coefficient	Individual R-Squared	Eigenvalue
PC1	7.6653	0.9905	1.994969
PC2	-0.0245	0.0000	1.004003
PC3	-10.8457	0.0010	0.001027

This report provides the details of the regression based on the principal components (the Z's).

Principal Component

This is the number of the principal component being reported about on this line. The order here corresponds to the order of the eigenvalues. Thus, the first is associated with the largest eigenvalue and the last is associated with the smallest.

Regression Coefficient

These are the estimated values of the regression coefficients $a_1, ..., a_p$. The value indicates how much change in Y occurs for a one-unit change in z when the remaining z's are held constant.

Individual R-Squared

This is the amount contributed to R-squared by this component.

Eigenvalue

This is the eigenvalue of this component.

Principal Components Regression Coefficients with Component(s) Omitted

Principal Components Regression Coefficients with 1 Component Omitted

Independent Variable	Regression Coefficient	Standard Error	Standardized Regression Coefficient	VIF
Intercept	0.763326			
X1	1.007698	0.0272776	0.4945	0.2645
X2	1.003778	0.02626337	0.4987	0.2513
X3	0.568248	0.2554352	0.0574	0.9815

Model

0.763326 + 1.007698*X1 + 1.003778*X2 + 0.568248*X3

This report provides the details of the PC 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 PC 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.

Standardized 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:

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

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 what VIF were given earlier.

Analysis of Variance

Analysis of Variance with 1 Component Omitted

Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1	9614.223	9614.223		
Model	3	1992.698	664.2327	487.3907	0.000000
Error	14	19.07968	1.362834		
Total(Adjusted)	17	2011.778	118.3399		

Additional Model Information

Mean of Dependent Variable	23.11111
Root Mean Square Error	1.167405
R-Squared	0.9905
Coefficient of Variation	0.05051271

An analysis of variance (ANOVA) table summarizes the information related to the sources of variation in the 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*. The formulas for each are:

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

Since PC 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.

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 α, say 0.05, the null hypothesis is rejected. If the p-value is greater than α, then the null hypothesis is accepted.

Mean of Dependent Variable

This is the arithmetic mean of the dependent variable.

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.

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

Predicted Values and Residuals with 1 Component Omitted

	Y		
Row	Actual	Predicted	Residual
1	3	4.346828	-1.346828
2	9	7.930331	1.069669
3	11	12.08208	-1.082081
4	15	13.52531	1.47469
5	13	13.96476	-0.9647598
6	13	14.40421	-1.40421
7	17	16.41569	0.5843138
8	21	19.99919	1.000811
9	25	24.15094	0.849061
10	27	25.59417	1.405833
11	25	26.03362	-1.033618
12	27	26.47307	0.5269322
13	29	28.48454	0.5154559
14	33	32.06805	0.9319535
15	35	36.2198	-1.219797
16	37	37.66302	-0.6630252
17	37	38.10247	-1.102475
18	39	38.54193	0.4580744

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 Y

This is the actual value of *Y* for the i^{th} row.

Predicted Y

The predicted value of *Y* for the i^{th} row. It is predicted using the levels of the *X's* for this row.

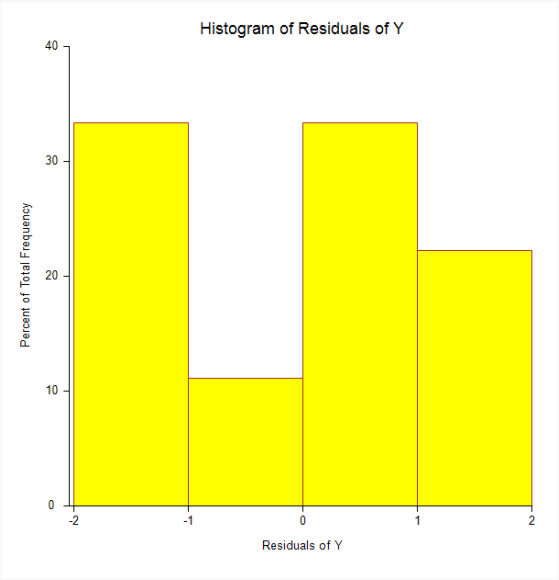
Residual

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

Histogram of Residuals

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

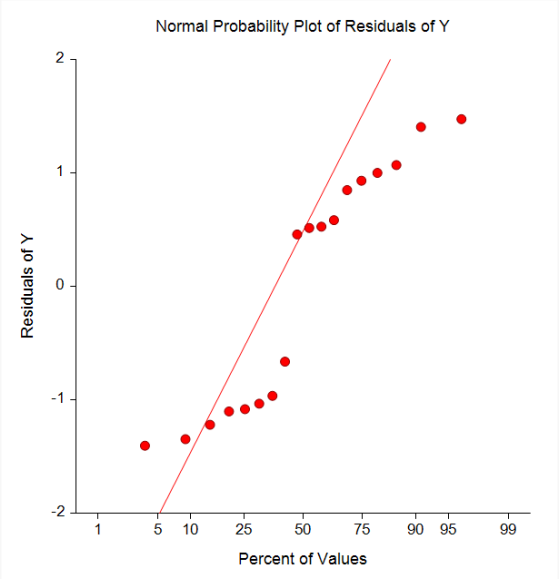
Distributional Plots of Residuals



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

Probability Plot of Residuals

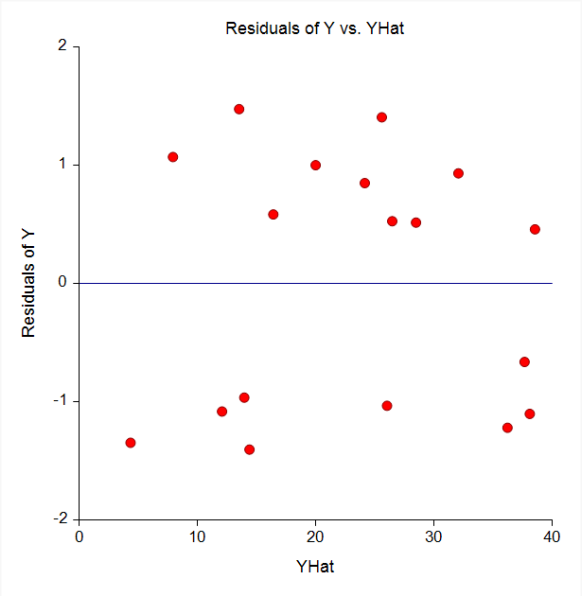
Distributional Plots of Residuals



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