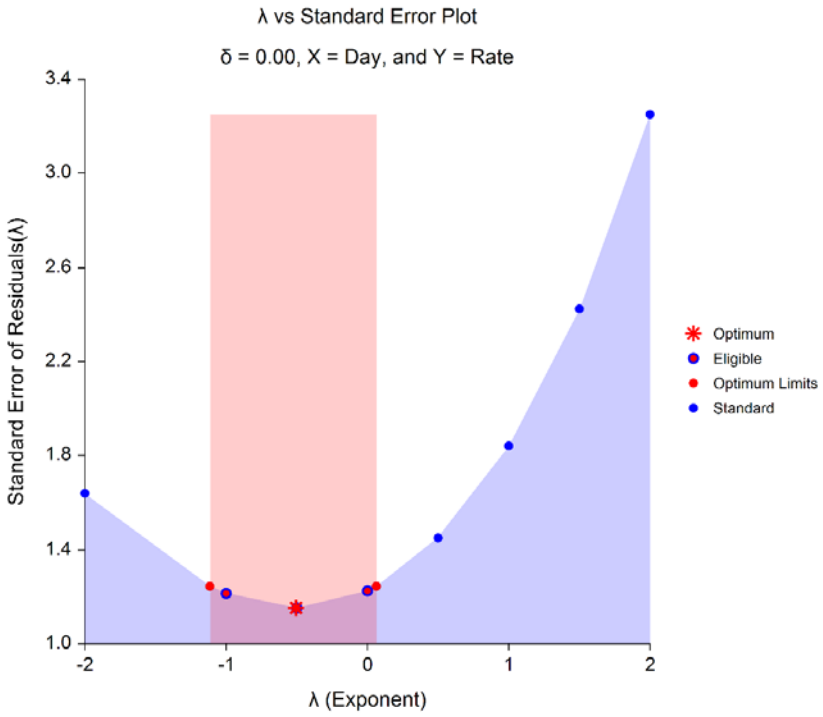
Chapter 192

Box-Cox Transformation for Simple Linear Regression

Introduction

This procedure finds the appropriate Box-Cox power transformation (1964) for a dataset containing a pair of variables that are to be analyzed by simple linear regression. This procedure is often used to modify the distributional shape of the response variable so that the residuals are more normally distributed. This is done so that tests and confidence limits that require normality can more appropriately be used. It cannot correct every data ill. For example, data that contain outliers may not be properly adjusted by this technique.

Example of the Box-Cox λ Plot



The Box-Cox transformation has the following mathematical form

$$Z = (Y + \delta)^{\lambda}$$

where λ is the exponent (power) and δ is a shift amount that is added when *Y* is zero or negative. When λ is zero, the above definition is replaced by

$$Z = \ln(Y + \delta)$$

Usually, the standard λ values of -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, and 2 are investigated to determine which, if any, is most suitable. The program will also solve for the optimum value of λ using maximum likelihood estimation. The program also calculates confidence limits about the optimum value. The usual procedure is to adopt the most convenient standard value between the confidence limits. For example, if the confidence limits were 0.4 to 1.1, λ would be set to the standard value of '1' (no transformation) since this is the most convenient. Care must be used when using the confidence limits, because they are heavily dependent on the sample size.

Box-Cox Algorithm

Suppose you have a sample of *n* observation pairs (Y_1, X_1), (Y_2, X_2), …, (Y_n, X_n). Further suppose you visually determine a value of δ the will keep all $X + \delta > 0$. Calculate a set of Z_i's corresponding to the Y_i's using

$$Z = \begin{cases} \left[(Y + \delta)^{\lambda} - 1\right] / \left[\lambda H^{\lambda - 1}\right] & \lambda \neq 0 \\ H \ln(Y + \delta) & \lambda = 0 \end{cases}$$

where *H* is the geometric mean of $Y + \delta$. That is,

$$H = \sqrt[n]{\prod_{i=1}^{n} (Y + \delta)}$$

Scaling by *H* is intended to keep the standard deviation of the *Z*'s approximately the same as the standard deviation of the *Y*'s so that the standard deviations can be compared at various values of λ.

Maximum Likelihood Estimation of λ

In this case, the likelihood for a given λ is inversely proportional to the square root of the mean square error of the residuals from the linear regression. The likelihood function is maximized when this value is minimized. A bracketing search algorithm is conducted that continues to tighten the boundaries until a specified precision (bracket width) is reached.

Approximate Confidence Interval for λ

An approximate confidence interval for λ is based on likelihood function which in turn is proportional to the sum of the squared residuals. The confidence limits correspond to the two values of λ at which

$$SD_{\lambda}^2 = SD_{\hat{\lambda}}^2 \exp\left(\frac{\chi_1^2(1-\alpha)}{n}\right)$$

where $\hat{\lambda}$ is the maximum likelihood estimate of λ and $\chi_1^2(1-\alpha)$ is the percentage point of the chi-squared distribution with one degree of freedom.

Data Structure

The data is entered in the standard columnar format in which the dependent variable (Y) is entered in one column and the independent variable (X) is entered in a second column. You can specify multiple Y's, but only one X on any one run of the program.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Data Variables

Y: Dependent Variable(s)

Specify the column(s) containing the dependent (response, predicted, or Y) variable. This is the variable to be predicted by the independent variables.

If more than one column is specified, a separate analysis is displayed for each column.

X: Independent Variable

Specify the column containing the independent variable (X). Only a single variable may be specified.

Correction for Non-Positive Data Values

All values must be positive to use the Box-Cox transformation. When negative values are encountered, an amount δ is added to each observation so that all resulting values are positive. This option controls if and how δ is specified.

- **Do Nothing (δ = 0)**
  No corrective action is taken. Non-positive values will be treated as missing values.
- **Add a fixed amount δ to each data value**
  Specify a value for δ which will always be added.
- **Add δ = |Min| + 1% of Range to each data value**
  Automatically set δ equal to the mini8mum plus one percent of the data range. This will insure that the result data values are always positive.

δ (Shift)

Specify an amount that will be added to each observation so that all values analyzed are positive. Hence, when one or more values are negative, δ should be less than the absolute value of the data minimum.

Any values that are negative or zero after δ has been added are treated as missing values in the analysis.

Maximum Likelihood Estimation of λ

Search for λ from Minimum to Maximum

The maximum likelihood estimate of λ is found using a special search algorithm that looks between two boundaries for the optimum value. Set the minimum boundary and the left and the maximum boundary on the right.

Typically, λ is between -5 and 5, so the minimum is often set to -5 and the maximum is often set to 5.

Width Around λ is Less Than

The search for the optimum value of λ uses a simple bracketing strategy. As the algorithm progresses, the distance between the upper and lower boundaries is decreased. Once this distance is less than the amount specified here, the algorithm is considered to have converged.

The scale is in terms of λ.

Since λ is usually rounded to the nearest half, a precision of 0.001 will be more than adequate.

Width Around C.L. is Less Than

The search for the confidence limits of the optimum λ uses a simple bracketing strategy. As the algorithm progresses, the distance between the upper and lower boundaries is decreased. Once this distance is less than the amount specified here, the algorithm is considered to have converged.

The scale is in terms of the residuals of the regression of Z (Y + δ) on X.

In most cases, a value of 0.0001 will be more than adequate.

Number of Iterations is More Than

Specify the maximum number of iterations used in the search for the optimum λ and in the search for the confidence limits about that optimum λ.

Usually, the algorithm will converge in 20 to 30 iterations, so we recommend 50 just to be on the safe side.

Standard λ's (Shown on Reports and Plots)

Input Type

Specify the input format of the standard λ values.

- **Range of λ values**
  Enter the minimum, maximum, and interval width from which a set of λ's may be calculated.

- **List of λ values**
  Specify a list of λ's directly.

Generate λ's from Min to Max

Enter the minimum and maximum values of λ. A set of λ's between the minimum and maximum is generated at points indicated by the Interval parameter. The maximum value is always included in the set, even if it does not match perfectly with the minimum and interval width.

The scaled standard deviation and normality test probability level are reported for each λ. Note that λ is the power (exponent) of the variable transformation. If λ is zero, the logarithmic transformation is used. Usually, λ is between -3 and 3.

Example

The settings minimum = 0, maximum = 2.5, and interval = 1 generates the set

0 1 2 2.5

Interval

Enter the interval between successive λ values. Typically, the interval is set to 0.5 or 0.25. Other values may be used, but most authors recommend either 0.5 or 0.25.

This value must be a positive number (zero is not allowed).

List of λ Values

Enter a list of λ values separated by blanks or commas. The scaled standard deviation and normality test probability level are displayed for each λ in the list.

Usually, λ is between -3 and 3.

Examples

-3 -2 -1 -0.5 0 0.5 1 2 3
-3 -2 -1 -0.5 -0.1
0 0.5 1 2 3

Reports Tab

The following options control the formatting of the reports.

Select Reports

Run Summary

Check to display this report.

Optimum λ and Confidence Limits

Check to display this report.

Standard λ's

Check to display this report.

Confidence Intervals

Confidence Level

Enter the confidence level of the confidence interval for the optimum λ. This value is entered as a percentage between 50 and 99.99. Usually, 95 is entered.

Reports Options Tab

The following options control the formatting of the reports.

Report Options

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Decimal Places

λ's – Test Statistics

These options allow you to specify the number of decimal places directly or based on the significant digits. If one

of the Auto options is used, the ending zero digits are not shown. For example, if 'Auto (Up to 7)' is chosen,

0.0500 is displayed as 0.05, 1.314583689 is displayed as 1.314584

The output formatting system is not always designed to accommodate 'Auto (Up to 13)', and if chosen, this will likely lead to lines that run on to a second line. This option is included, however, for the rare case when a very large number of decimals is needed.

Plots Tab

These options let you specify which plots are displayed.

Select Plots

Histogram of Original Data

Check to output this histogram.

Histogram of using Optimum λ

Check to output this histogram.

Histogram of using Standard λ's

Check to output these histograms.

Histogram Format

Click this button to specify the format of these histograms. A plot using sample data (not your actual data) will be displayed. If you would like to edit the plot with your actual data loaded into the plot preview, check the *Edit During Run* box in the upper right-hand corner of the plot format button.

Probability Plot of Original Data

Check to output this probability plot.

Probability Plot of using Optimum λ

Check to output this probability plot.

Probability Plot of using Standard λ's

Check to output these probability plots.

Probability Plot Format

Click this button to specify the format of these probability plots. A plot using sample data (not your actual data) will be displayed. If you would like to edit the plot with your actual data loaded into the plot preview, check the *Edit During Run* box in the upper right-hand corner of the plot format button.

λ vs SE Plot

Check to output this plot.

λ vs SE Plot Format

Click this button to specify the format of this plot. A plot using sample data (not your actual data) will be displayed. If you would like to edit the plot with your actual data loaded into the plot preview, check the *Edit During Run* box in the upper right-hand corner of the plot format button.

Plots Options Tab

These options let you specify where to store various row-wise statistics.

Decimal Places for Plot Titles and Label

λ in Labels and Titles

These options allow you to specify the number of decimal places directly or based on the significant digits. If one of the Auto options is used, the ending zero digits are not shown. For example, if 'Auto (Up to 7)' is chosen,

0.0500 is displayed as 0.05 and 1.314583689 is displayed as 1.314584.

δ in Labels and Titles

These options allow you to specify the number of decimal places directly or based on the significant digits. If one of the Auto options is used, the ending zero digits are not shown. For example, if 'Auto (Up to 7)' is chosen,

0.0500 is displayed as 0.05.

Text for Plot Titles and Label

λ vs SE Plot Title Line 1 – Legend-Standard

Enter the text to be used in plot titles, labels, and legends.

The following substitutions will take place as the entered text is processed.

{Z} will be replaced by the current Y variable name.

{W} will be replaced by the current X variable name.

{D} will be replaced by the current value of δ.

Example 1 – Box-Cox for Linear Regression

This section presents an example of how to run a Box-Cox transformation analysis on a set of simple linear regression data. The data used are found in the *Box Cox Lin Reg* dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the procedure window.

1 Open the BoxCoxLinReg dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **BoxCoxLinReg**.
- Click **Ok**.

2 Open the Box-Cox Transformation for Simple Linear Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Box-Cox Transformation for Simple Linear Regression** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the procedure window, select the **Variables tab**.
- Double-click in the **Y: Dependent Variable(s)** box. This will bring up the variable selection window.
- Select **Rate** from the list of variables and then click **Ok**. *Rate* will appear in the Y: Dependent Variable(s) box.
- Double-click in the **X: Independent Variable** box. This will bring up the variable selection window.
- Select **Dat** from the list of variables and then click **Ok**. *Day* will appear in the X: Independent Variable box.
- That's it. All other options can stay at their default values.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary Section

Run Summary for X = Day and Y = Rate

Parameter	Value	Parameter	Value
δ (Shift)	0	Rows Processed	25
Optimum λ (Power)	-0.5049	Rows Used	25
Minimum λ Searched	-5		
Maximum λ Searched	5	Geometric Mean (with δ)	8.5163
Target Search Width of λ	0.0001	Minimum	4.8600
MLE Iterations Used	26	Maximum	20.0900
Max MLE Iterations	50	Max/Min (> 10?)	4.13

This report summarizes the run by showing main results as well as the input settings that were used. You should pay particular attention to the *Rows* lines to make sure that they are as you expect. Also, if the number of MLE Iterations is equal to the Max MLE Iterations, the search algorithm may not have converged properly.

When the ratio of the maximum to the minimum is greater than 10, the Box-Cox transformation is often useful.

Optimum (Maximum Likelihood) Estimate of λ

Optimum (Maximum Likelihood) Estimate of λ for X = Day and Y = Rate
Power Transformation: Y = (X + δ) ^ λ

Item	Power λ	Shift δ	Square Root of MSE	R-Squared	Shapiro-Wilk Normality Test Prob Level
Optimum (MLE)	-0.5049	0	1.1527	0.8665	0.9732
Lower 95% C. L.	-1.1157	0	1.2447	0.8528	0.6770
Upper 95% C. L.	0.0629	0	1.2447	0.8501	0.4799

This report gives the results for the maximum likelihood estimation portion of the analysis.

Item

The name of item being reported on this line of the report.

Power λ

The value of λ for this item. This is the transformation exponent.

Shift δ

The value of δ, the shift value.

Square Root of MSE

This is the square root of the mean square error of the linear regression using the transformed data values. Actually, the data have not only been shifted and raised to the indicated power, but they have also been scaled by the geometric mean so that these values are directly comparable. Note the geometric mean is not used when using the λ that has been found by this algorithm.

R-Squared

This column gives the R-squared value for this transformation. Obviously, you want to maximize this value.

Shapiro-Wilk Normality Test Prob Level

The probability level of the Shapiro-Wilk normality test. Since the desire is to transform the data to be more normally distributed, you are looking for large (non-significant) values. Remember that this value is not only influenced by the normality of the data, but also by the sample size.

Standard λ's

Standard λ's for X = Day and Y = Rate
Power Transformation: Y = (X + δ) ^ λ

Item	Power λ	Shift δ	Square Root of MSE	R-Squared	Shapiro-Wilk Normality Test Prob Level
1	-2.0000	0	1.6399	0.7968	0.3210
2	**-1.0000**	**0**	**1.2142**	**0.8575**	**0.6438**
3	**-0.5000**	**0**	**1.1527**	**0.8665**	**0.9757**
4	**0.0000**	**0**	**1.2251**	**0.8535**	**0.6784**
5	0.5000	0	1.4502	0.8154	0.0219
6	1.0000	0	1.8413	0.7532	0.0011
7	1.5000	0	2.4242	0.6736	0.0001
8	2.0000	0	3.2507	0.5860	0.0000

λ's between the maximum likelihood confidence limits are bolded.

This report displays the results for each of the standard λ's.

Item

The number of item being reported on this line of the report.

Power λ

The value of λ for this item. This is the transformation exponent.

Shift δ

The value of δ, the shift value.

Square Root of MSE

This is the square root of the mean square error of the linear regression using the transformed data values. Actually, the data have not only been shifted and raised to the indicated power, but they have also been scaled by the geometric mean so that these values are directly comparable. Note the geometric mean is not used when using the λ that has been found by this algorithm.

R-Squared

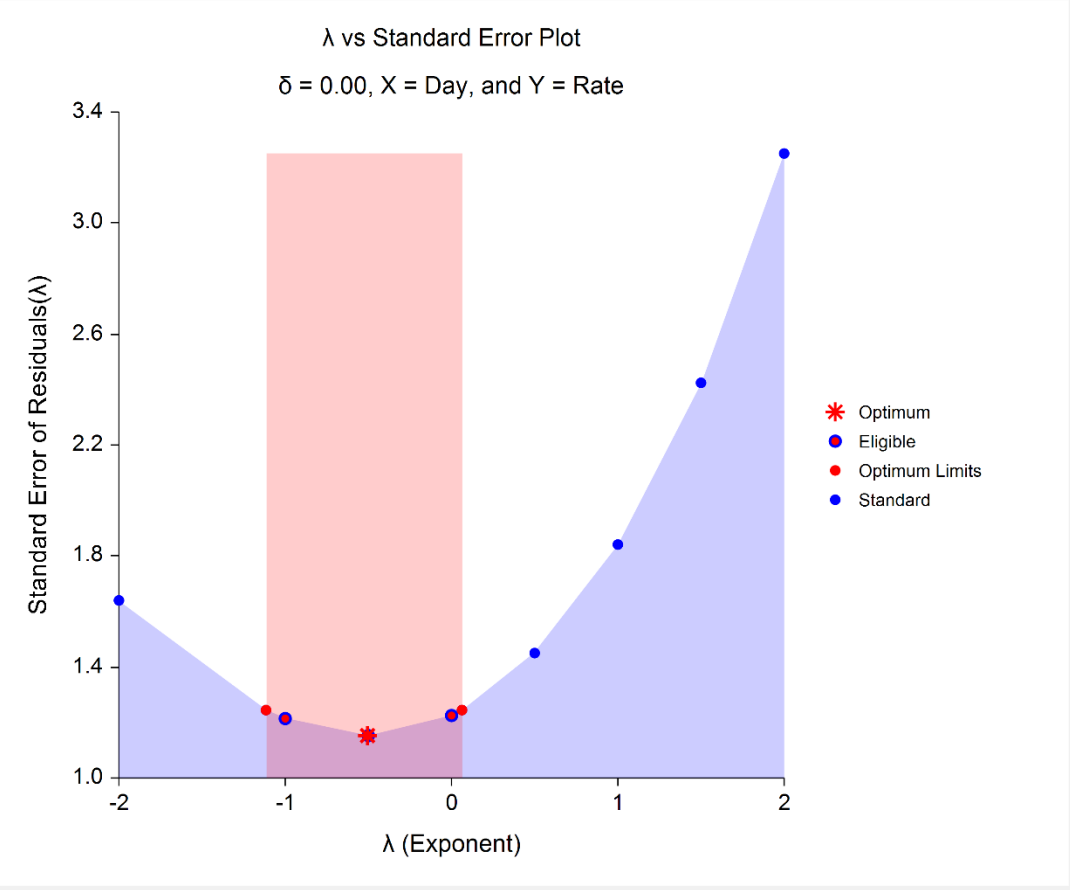
This column gives the R-squared value for this transformation. Obviously, you want to maximize this value.

Shapiro-Wilk Normality Test Prob Level

The probability level of the Shapiro-Wilk normality test. Since the desire is to transform the data to be more normally distributed, you are looking for large (non-significant) values. Remember that this value is not only influenced by the normality of the data, but also by the sample size.

Plots

Plot for Choosing λ

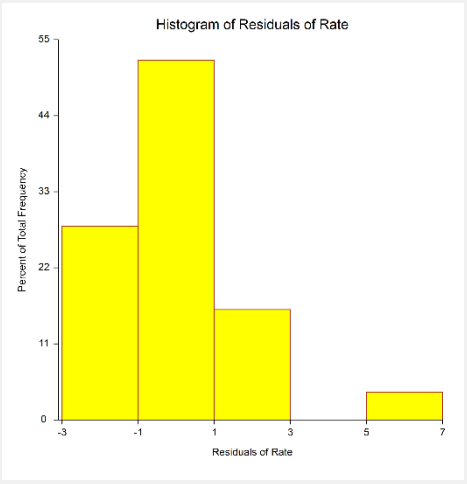


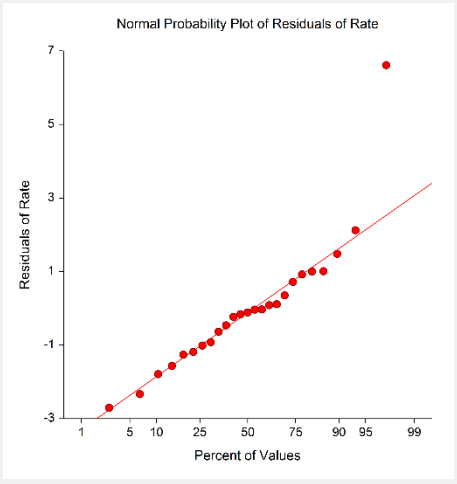
Box-Cox Transformation for Simple Linear Regression

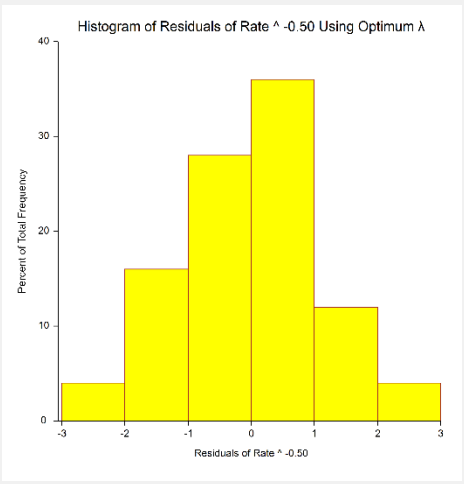
This plot gives a visual representation that will help you select the value of λ that you want to use. The optimum value found by maximum likelihood is plotted with a large, red asterisk. This value is usually inconvenient to use, so a convenient (standard) value is sought for that is close to the optimum value. These convenient values are plotted using a blue circle with a red center. In this example, it is obvious that λ = -0.5 (1/square root) is certainly a reasonable choice. The large shaded area in the middle of the plot highlights the values of λ that are within the confidence interval for the optimum.

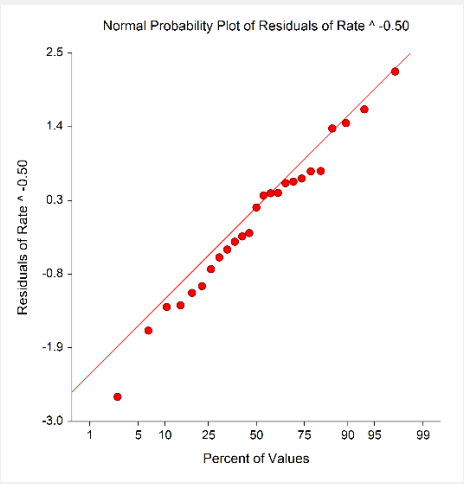
Note that this plot was created using the Scatter Plot procedure. The shading effects and different plot symbols were made by making several groups of data.

Plots for Assessing Normality at Various λ's









These plots let you see the improvement towards normality achieved by the power transformation. The top row shows the histogram and probability plot of the original data. The lack of normality is evident in both plots. The problem appears to be an outlier.

The bottom row of plots shows the same two plots applied to the data that has been transformed by the optimum λ. The histogram is now much closer to being bell shaped.