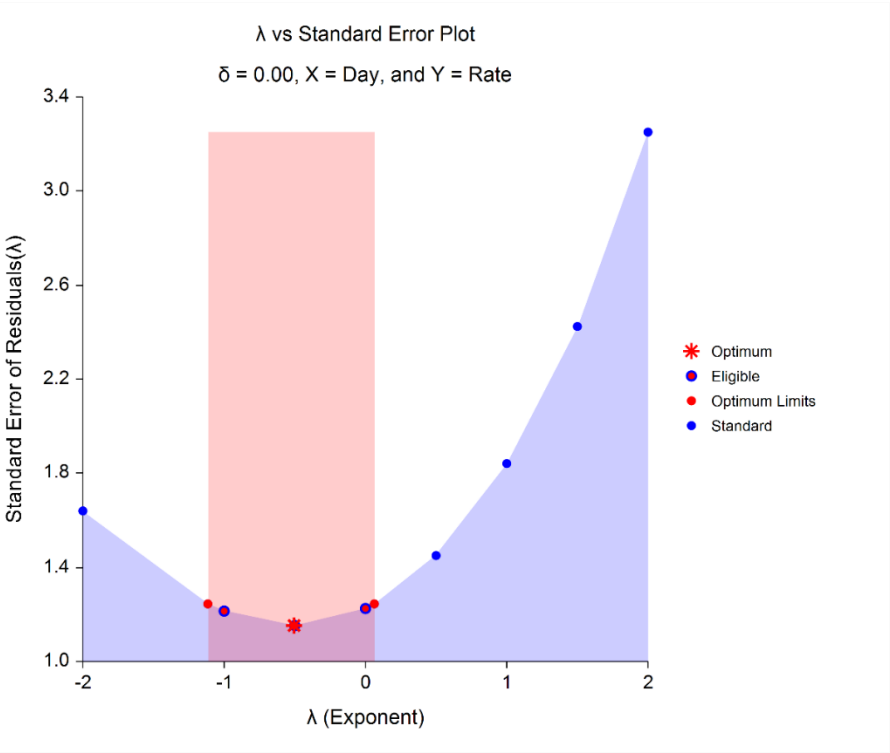
Chapter 192

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 that 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.

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.

Setup

To run this example, complete the following steps:

1 Open the BoxCoxLinReg example dataset

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

2 Specify the Box-Cox Transformation for Simple Linear Regression procedure options

- Find and open the **Box-Cox Transformation for Simple Linear 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(s) .....**Rate**
X Independent Variable .....**Day**

3 Run the procedure

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

Run Summary

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	4.86
MLE Iterations Used	26	Maximum	20.09
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: Z = (Y + δ) ^ λ

Item	Power λ	Shift δ	Square Root of MSE	R-Squared	Shapiro-Wilk Normality Test P-Value
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 P-Value

The p-value 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: Z = (Y + δ) ^ λ

Item	Power λ	Shift δ	Square Root of MSE	R-Squared	Shapiro-Wilk Normality Test P-Value
1	-2.0	0	1.6399	0.7968	0.3210
2	**-1.0**	**0**	**1.2142**	**0.8575**	**0.6438**
3	**-0.5**	**0**	**1.1527**	**0.8665**	**0.9757**
4	**0.0**	**0**	**1.2251**	**0.8535**	**0.6784**
5	0.5	0	1.4502	0.8154	0.0219
6	1.0	0	1.8413	0.7532	0.0011
7	1.5	0	2.4242	0.6736	0.0001
8	2.0	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 items 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

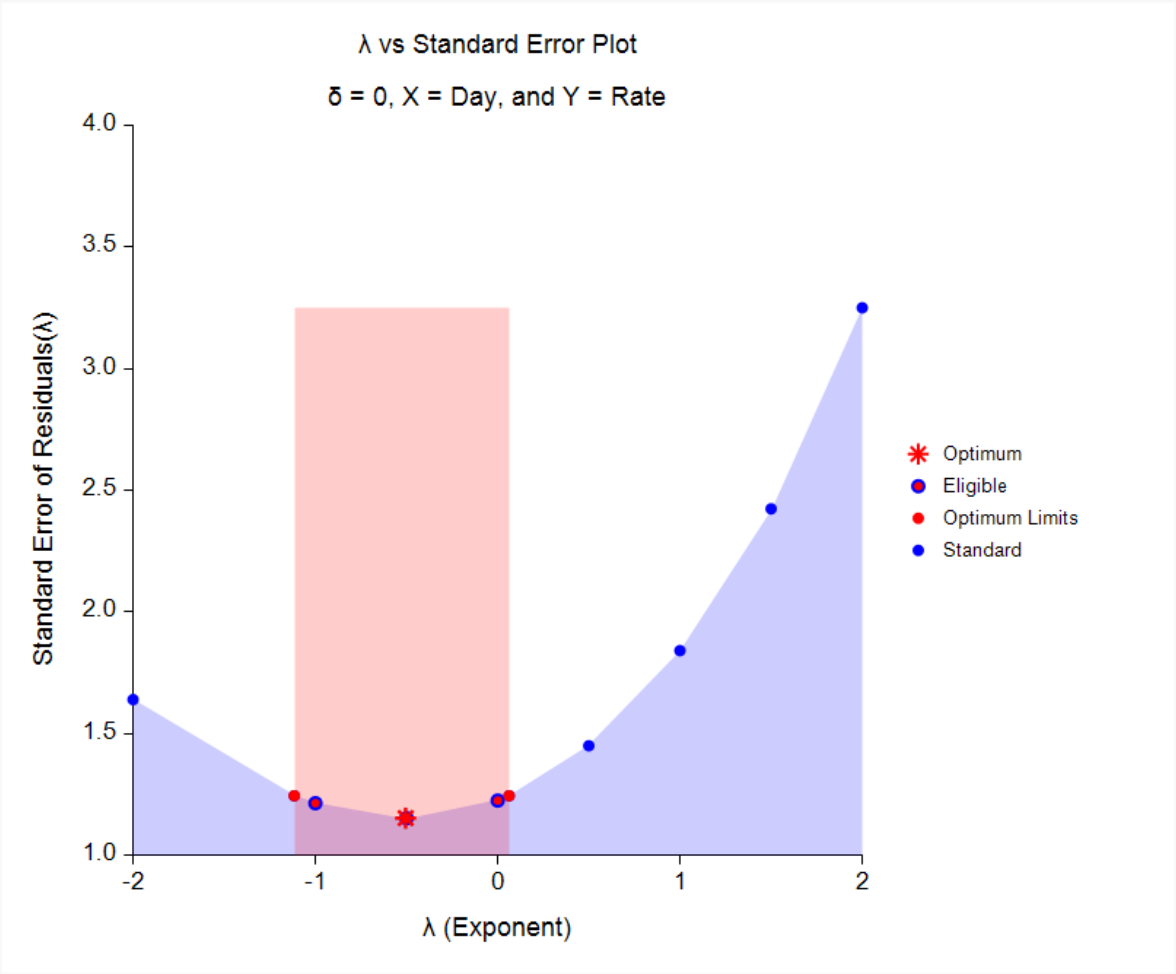
This column gives the R-squared value for this transformation. Obviously, you want to maximize this value.

Shapiro-Wilk Normality Test P-Value

The p-value 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

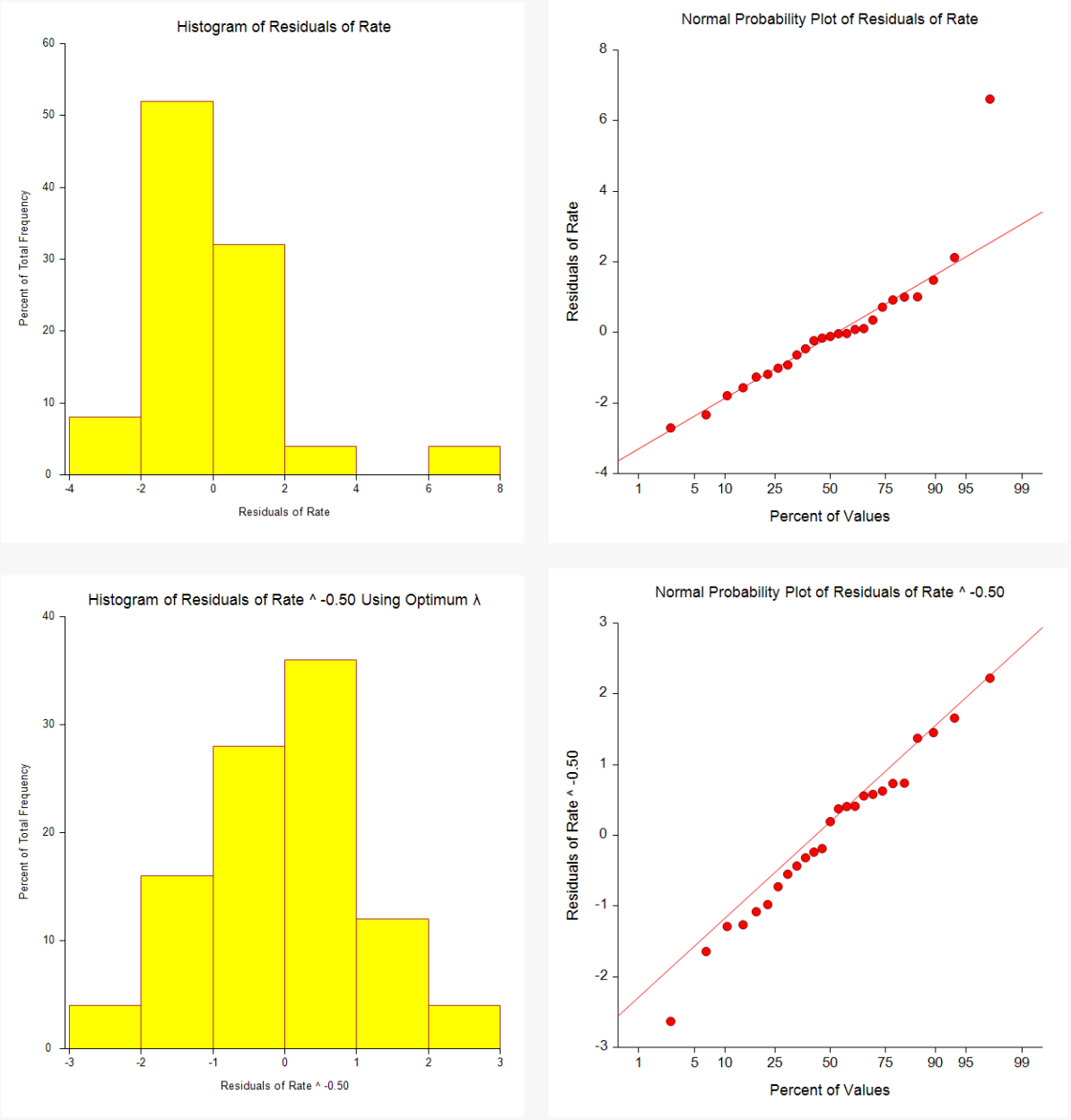
Plot for Choosing λ



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.