How do I perform a Box-Cox transformation in SAS?

Authored by

stats writer

June 23, 2024

A Box-Cox transformation is a statistical technique used to transform non-normally distributed data into a more normal distribution. This can be done in SAS by using the PROC TRANSREG procedure and specifying the BOXCOX option. This will automatically calculate and apply the appropriate transformation to the data. The output will include the transformed variable and the lambda value, which indicates the level of transformation used. The transformed data can then be used for further analysis or modeling. It is a useful tool for improving the accuracy and validity of statistical models.

Perform a Box-Cox Transformation in SAS

A box-cox transformation is a commonly used method for transforming a non-normally distributed dataset into a more normally distributed one.

The basic idea behind this method is to find some value for λ such that the transformed data is as close to normally distributed as possible, using the following formula:

y(λ) = (yλ - 1) / λ if y ? 0y(λ) = log(y) if y = 0

We can identify the optimal value to use for λ in SAS by using the PROC TRANSREG procedure.

The following example shows how to use this procedure in practice.

Example: Box-Cox Transformation in SAS

Suppose we have the following dataset in SAS:

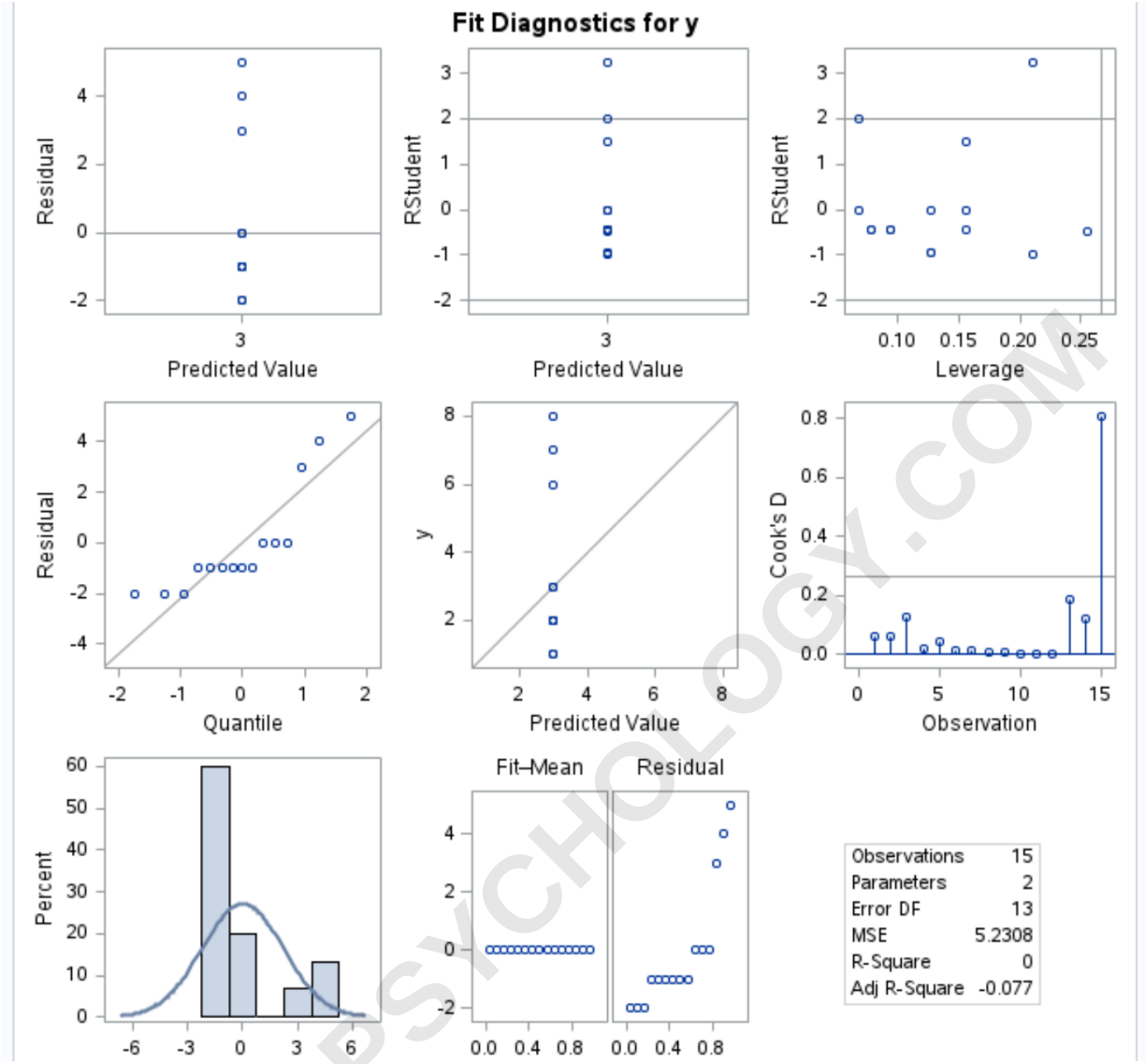
```
/*create dataset*/
data my_data;
input x y;
datalines;
7 1
7 1
8 1
3 2
2 2
4 2
4 2
6 2
6 2
7 3
5 3
3 3
3 6
5 7
8 8
;
run;
```

```
/*view dataset*/
proc printdata=my_data;
```

Suppose we use PROC REG to fit a simple linear regression model to this dataset, using x as the predictor variable and y as the response variable.

```
/*fit simple linear regression model*/
proc regdata=my_data;
model y = x;
run;
```

In the diagnostic plots in the output, we can view the Residual vs. Quantile plot (left-most plot in the middle row) to see if the residuals are roughly normally distributed in the model:



If the fall roughly along the straight diagonal line in the plot, then we generally assume that the residuals are normally distributed.

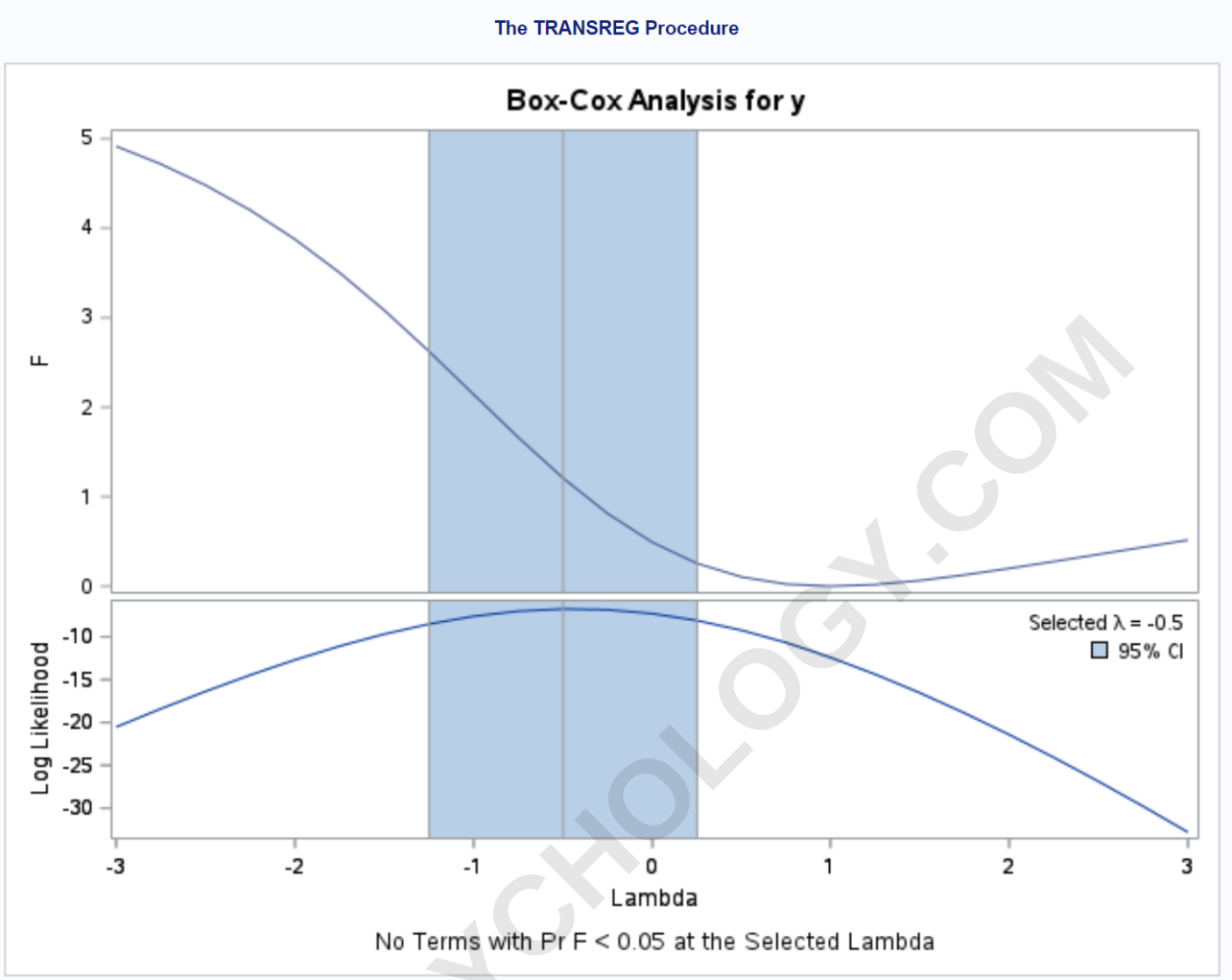
From the plot we can see that the residuals do not fall along the straight diagonal line much.

This is an indication that the response variable in the

regression model likely is not normally distributed.

Since the response variable is not normally distributed, we can use PROC TRANSREG to identify a value for λ that we can use to transform the response variable to be more normally distributed:

```
/*perform box-cox transformation*/
proc transregdata=my_data;
modelboxcox(y) = identity(x);
run;
```



The output tells us that the selected value to use for λ is -0.5.

Thus, we can fit a simple linear regression model by replacing the original response variable y with the variable y = (y-0.5 - 1) / -0.5.

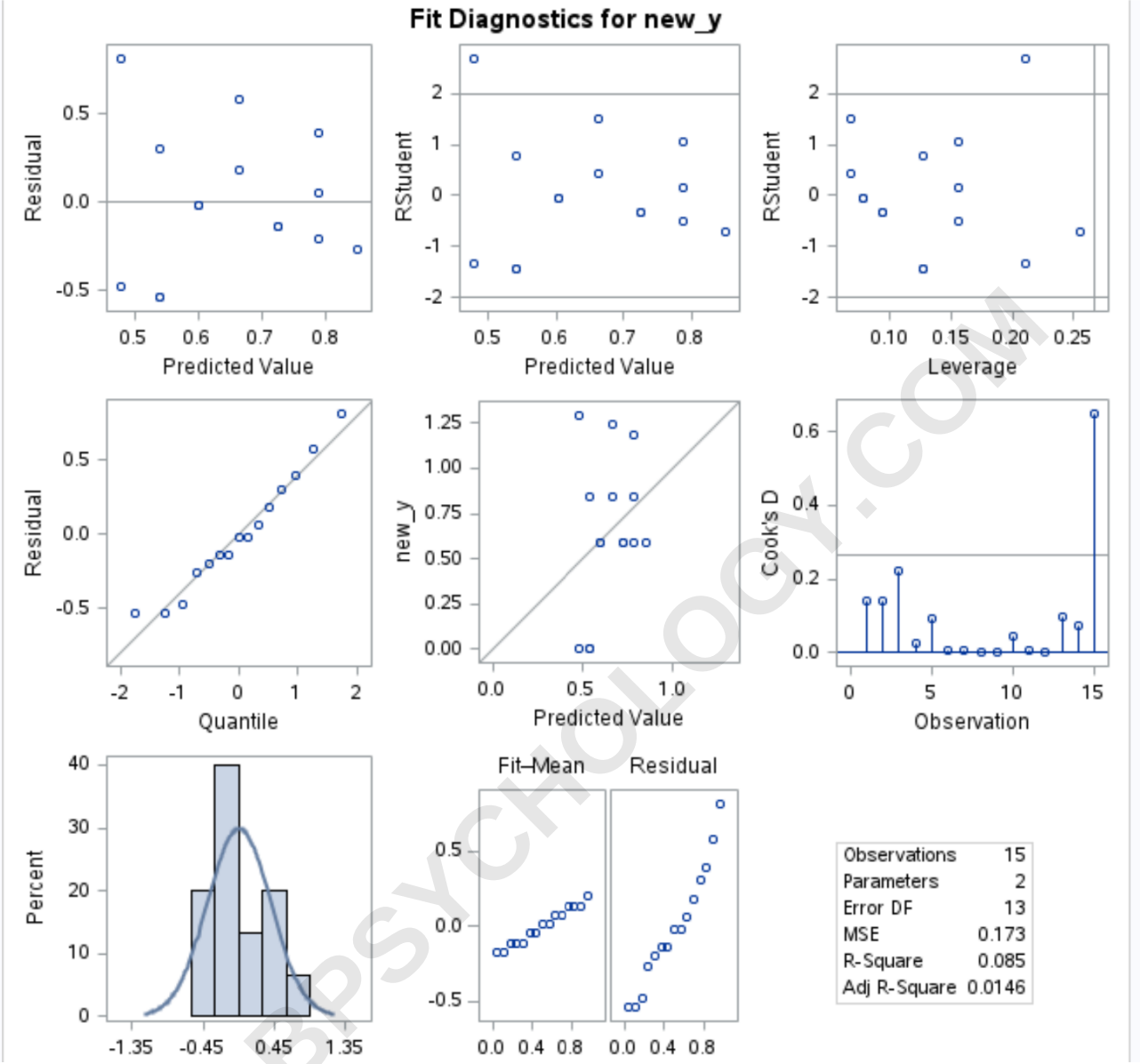
The following code shows how to do so:

/*create new dataset that uses box-cox transformation

```
to create new y*/
data new_data;
set my_data;
new_y = (y**(-0.5) - 1) / -0.5;
run;

/*fit simple linear regression model using new response variable*/
proc regdata=new_data;
model new_y = x;
run;
```

In the Residual vs. Quantile plot in this model output we can see that the residuals fall along the straight diagonal line much more closely:



This is an indication that the residuals of the box-cox transformed model are much more normally distributed, which satisfies one of the main assumptions of linear regression.

The following tutorials explain how to perform other

common tasks in SAS: