

# How do you perform a Box-Cox transformation in R, and can you provide examples?

Authored by  
**stats writer**

April 20, 2024

## RECOMMENDED CITATION

stats writer (2024). *How do you perform a Box-Cox transformation in R, and can you provide examples?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=137479>

The Box-Cox transformation is a statistical technique used to transform non-normally distributed data into a more normally distributed form. This transformation is commonly used in data analysis and modeling to improve the accuracy of statistical tests and predictions.

To perform a Box-Cox transformation in R, the "boxcox" function from the "MASS" package can be used. This function takes in the data to be transformed as well as a lambda value, which determines the type of transformation to be applied. A lambda value of 0 indicates a log transformation, while a lambda value of 1 indicates no transformation. The function then returns the transformed data.

For example, let's say we have a dataset of income values that are highly skewed. We can perform a Box-Cox transformation with a lambda value of 0.5 using the following code:

```
transformed_data <- boxcox(income, lambda = 0.5)
```

This will transform the income values and store the transformed data in the "transformed\_data" variable.

Additionally, the "boxcox" function also has an option to automatically select the optimal lambda value based on the data. This can be done by setting the "lambda" parameter to "NULL" in the function call.

In summary, the Box-Cox transformation is a powerful tool for improving the normality of non-normally distributed data. It can easily be performed in R using the "boxcox" function, which allows for both manual and automatic selection of the transformation parameters.

## Perform a Box-Cox Transformation in R (With Examples)

**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  $\lambda$  such that the transformed data is as close to**

normally distributed as possible, using the following formula:

$$y(\lambda) = (y^\lambda - 1) / \lambda \text{ if } y > 0$$
$$y(\lambda) = \log(y) \text{ if } y = 0$$

We can perform a box-cox transformation in R by using the `boxcox()` function from the `MASS()` library. The following example shows how to use this function in practice.

*Refer to [this paper](#) from the University of Connecticut for a nice summary of the development of the Box-Cox transformation.*

Example: Box-Cox Transformation in R

The following code shows how to fit a linear regression model to a dataset, then use the `boxcox()` function to find an optimal lambda to transform the response variable and fit a new model.

```
library(MASS)
```

```
#create data
```

```
y=c(1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 6, 7, 8)
```

```
x=c(7, 7, 8, 3, 2, 4, 4, 6, 6, 7, 5, 3, 3, 5, 8)
```

```
#fit linear regression model
```

```
model <- lm(y~x)
```

```
#find optimal lambda for Box-Cox transformation
```

```
bc <- boxcox(y ~ x)
```

```
(lambda <- bc$x)
```

```
-0.4242424
```

```
#fit new linear regression model using the Box-Cox transformation
```

```
new_model <- lm(((y^lambda-1)/lambda) ~ x)
```

The optimal lambda was found to be -0.4242424. Thus, the new regression model replaced the original response variable  $y$  with the variable  $y = (y - 0.4242424) / -0.4242424$ .

The following code shows how to create two Q-Q plots in R to visualize the differences in residuals between the two regression models:

```
#define plotting area
```

```
op <- par(pty = "s", mfrow = c(1, 2))
```

```
#Q-Q plot for original model
```

```
qqnorm(model$residuals)
```

```
qqline(model$residuals)
```

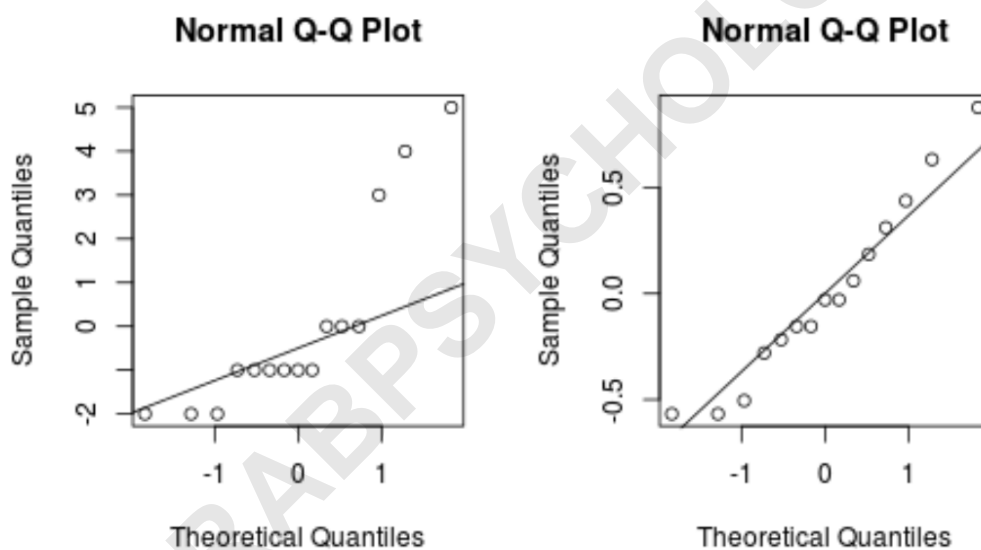
```
#Q-Q plot for Box-Cox transformed model
```

```
qqnorm(new_model$residuals)
```

```
qqline(new_model$residuals)
```

```
#display both Q-Q plots
```

```
par(op)
```



As a rule of thumb, if the data points fall along a straight diagonal line in a Q-Q plot then the dataset likely follows a normal distribution.

Notice how the box-cox transformed model produces a

**Q-Q plot with a much straighter line than the original regression model.**

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

**How to Transform Data in R (Log, Square Root, Cube Root)**

**How to Create & Interpret a Q-Q Plot in R**

**How to Perform a Shapiro-Wilk Test for Normality in R**