# Box-Cox Transformation

Dr. Kiah Wah Ong

ILLINOIS TECH

# Model Assumptions

Recall the major assumptions we have made in linear regression models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

are

- The relationship between the response and regressors is linear.
- The error terms $\epsilon_i$ have mean zero.
- The error terms $\epsilon_i$ have constant variance $\sigma^2$ (homoscedasticity)
- The error terms $\epsilon_i$ are normally distributed.
- The error terms $\epsilon_i$ and $\epsilon_j$ are uncorrelated for $i \neq j$.
- The regressors $x_1, \cdots, x_k$ are nonrandom.
- The regressors $x_1, \cdots, x_k$ are measured without error.
- The regressors are linearly independent.

# Box-Cox Transformation

Box and Cox (1964) gave us a procedure which we can employ if we wish to transform $y$ to correct for nonnormality and/or nonconstant variance.

This is done through the **power transformation** $y^\lambda$, where $\lambda$ is a parameter to be determined.

# Box-Cox Transformation

After the power transformation, the regression model becomes

$$y_i^{\lambda} = \beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik} + \epsilon_i$$

where

$$y_i^{\lambda} = \begin{cases} \frac{y_i^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln y_i & \text{for } \lambda = 0 \end{cases}$$

The optimal transformation parameter $\lambda$ and the parameters of the least square regression model

$$\mathbf{y}^{\lambda} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

are computed together through a maximum likelihood estimation.

# Box-Cox Transformation

Note that this family $y^\lambda$ encompasses the following simple transformations:

| $\lambda$ | $y' = y^\lambda$ |
|---|---|
| $\lambda = 2$ | $y' = y^2$ |
| $\lambda = 1/2$ | $y' = \sqrt{y}$ |
| $\lambda = 0$ | $y' = \ln y$ (by definition) |
| $\lambda = -1/2$ | $y' = 1/\sqrt{y}$ |
| $\lambda = -1$ | $y' = 1/y$ |

Note that the Box-Cox method works only if the response variable $y$ takes only positive values.

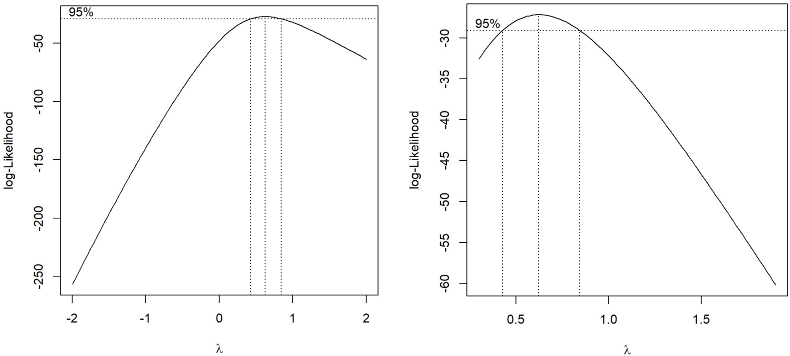
If the response takes on some negative values, a pre-transformation of data is needed to be performed, such as adding an appropriately positive number to all observations.

# Box-Cox Transformation

We can perform the Box-Cox transformation Using MASS in R pretty easily.

```
VST2<-read.csv("VST2.CSV", header=TRUE, sep=",")
x<-VST2$x
y<-VST2$y
model1=lm(y~x)
library(MASS)
bc=boxcox(model1)
```

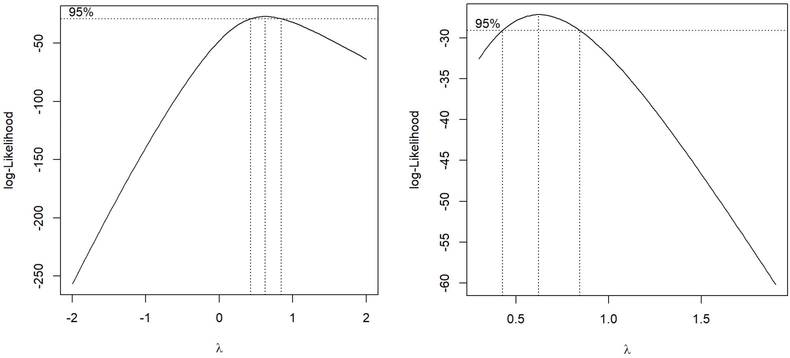
# Box-Cox Transformation




Comparing the default 95% confidence interval for $\lambda$ using

$\mathrm{bc} = \mathrm{boxcox}(\mathrm{model1})$ and $\mathrm{boxcox}(\mathrm{model1}, \mathrm{seq}(0.3, 1.9, 0.1))$

# Box-Cox Transformation




The instruction $\mathrm{best.lam} = \mathrm{bc\$x[which(bc\$y == max(bc\$y))]}$ gives the best value of $\lambda$, which is 0.626262... .

We can use $y' = \sqrt{y}$ as our transformation just like what we have done before. Notice that $\lambda = 0.5$ is still in the 95%-confidence interval of $\lambda$.