Lecture 15:
Box-Cox Transformations
STAT 632, Spring 2020

Transforming the Response

The Box-Cox method is a popular way to estimate a transformation for the response variable.

For a strictly positive response variable Y, the Box-Cox family of transformations is given by:

$$h(Y, \lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(Y), & \text{if } \lambda = 0 \end{cases}$$

Transformations, previously discussed, such as the square root ($\lambda = 1/2$) and the logarithm ($\lambda = 0$) are members of the Box-Cox family. Note that the log() function is assumed to be base e (also denoted as ln(), the natural logarithm).

Properties of Box-Cox transformations:

- $h(Y, \lambda)$ is a continuous function of λ
- The logarithmic transformation is a member of this family, since

$$\lim_{\lambda \to 0} \frac{Y^{\lambda} - 1}{\lambda} = \log(Y)$$

- Dividing by λ preserves the direction of Y, which would otherwise be reversed when λ is negative. As an illustration

Y	Y^{-1}	$\frac{Y^{-1}}{-1}$	$\frac{Y^{-1}-1}{-1}$
1	1	-1	0
2	1/2	-1/2	1/2
3	1/3	-1/3	2/3
4	1/4	-1/4	3/4

The parameter λ is estimated by maximizing the log-likelihood function[1]:

$$l(\lambda) = -\frac{n}{2}\log(\text{RSS}(\lambda)/n) + (\lambda - 1)\sum_{i=1}^{n}\log y_i$$

where $\text{RSS}(\lambda)$ is the residuals sum of squares when $h(y_i, \lambda)$ is the response. More specifically,

$$\text{RSS}(\lambda) = \sum_{i=1}^{n}[h(y_i, \lambda) - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}]^2$$

where $\hat{\beta}_0, \cdots, \hat{\beta}_p$ are the least squares estimates of the parameters for transformed regression model $h(Y_i, \lambda) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e$.

[1] See Sheather, pp. 228–230, for some background on maximum likelihood estimation in the context of multiple linear regression.

Example: Modeling Defective Rates (Sheather, Ch. 6)

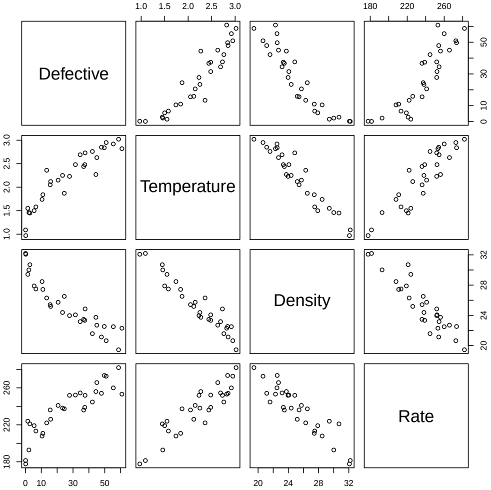
- For this example, the response variable is `Defect`, Y, the average number of defects per 1000 parts produced.
- The predictors are
  - `Temperate`, x_1, the standard deviation of the temperature of the production process
  - `Density`, x_2, the density of the product
  - `Rate`, x_3, the production rate
- Data were collected for $n = 30$ production runs.
- We begin by fitting the regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

and then consider a Box-Cox transformation $h(Y, \lambda)$ for the response.

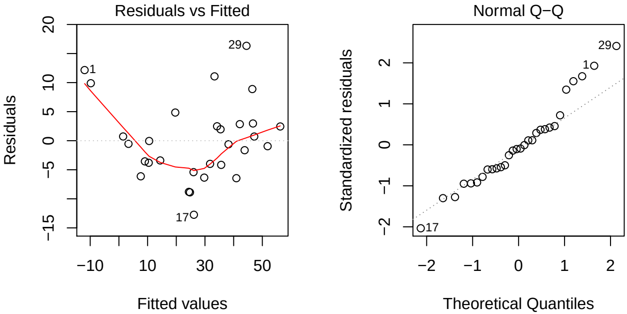
Defective Widgets

```
> defects <- read.csv("https://ericwfox.github.io/data/defects.csv")
> pairs(Defective ~ Temperature + Density + Rate, data=defects)
```

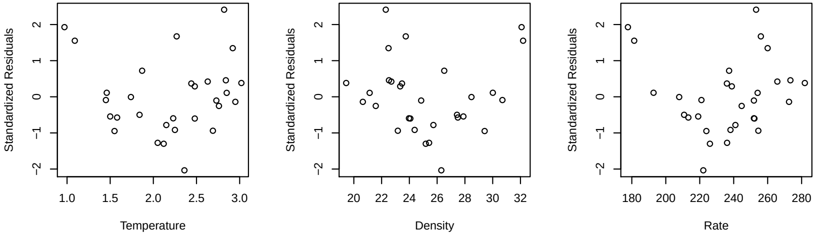


The scatterplot matrix and residual versus fitted plot show nonlinearity in the data that is not accounted for by the linear regression model.

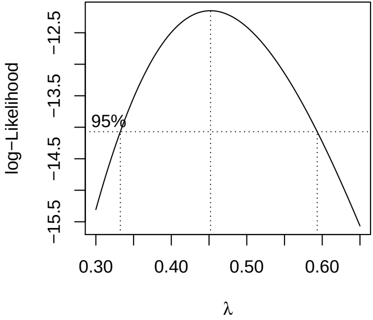
```
> lm1 <- lm(Defective ~ Temperature + Density + Rate, data=defects)
> par(mfrow=c(1,2), mar=c(4.5, 4.5, 2, 2))
> plot(lm1, 1:2)
```



```
> par(mfrow=c(1,3), mar=c(4.5, 4.5, 2, 2))
> plot(defects$Temperature, rstandard(lm1),
      xlab="Temperature", ylab="Standardized Residuals")
> plot(defects$Density, rstandard(lm1),
      xlab="Density", ylab="Standardized Residuals")
> plot(defects$Rate, rstandard(lm1),
      xlab="Rate", ylab="Standardized Residuals")
```



```
> library(MASS)
> library(car)
> boxcox(lm1, lambda=seq(0.3, 0.65, by=0.05))
> summary(powerTransform(lm1))
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.4519         0.5       0.3253       0.5785
```



- Using the Box-Cox procedure, the estimated value of the parameter is $\hat{\lambda} = 0.45$.
- The 95% confidence interval for λ is between 0.33 and 0.58.
- For interpretability, we will round and use $Y^{0.5}$, the square root transformation.
- Thus, the regression model with the transformed response is

$$\sqrt{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

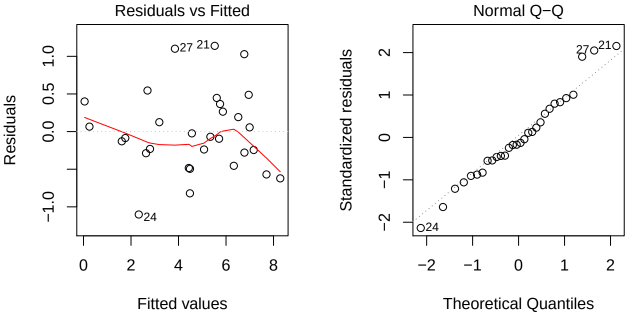
The regression equation is given by:

$$\widehat{\sqrt{Y}} = 5.593 + 1.565x_1 - 0.292x_2 + 0.013x_3$$

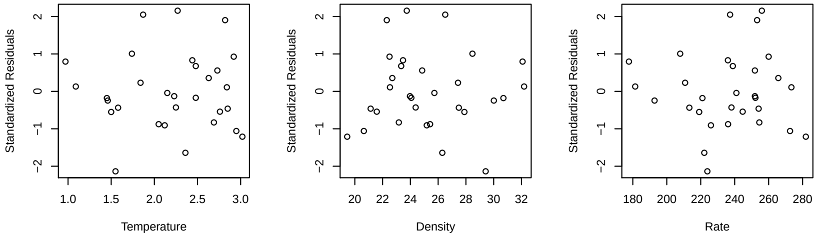
```
> lm2 <- lm(sqrt(Defective) ~ Temperature + Density + Rate,
            data=defects)
> summary(lm2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.59297    5.26401   1.062   0.2978
Temperature  1.56516    0.66226   2.363   0.0259 *
Density     -0.29166    0.11954  -2.440   0.0218 *
Rate         0.01290    0.01043   1.237   0.2273
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5677 on 26 degrees of freedom
Multiple R-squared:  0.943,Adjusted R-squared:  0.9365
F-statistic: 143.5 on 3 and 26 DF,  p-value: 2.713e-16
```

After making the transformation, the plots of the residuals versus fitted values, and residuals versus each predictor, are much improved. The points appear to be randomly scattered in each plot, and show no obvious patterns or nonconstant variance. Thus, the transformed model appears to fit the data reasonably well.

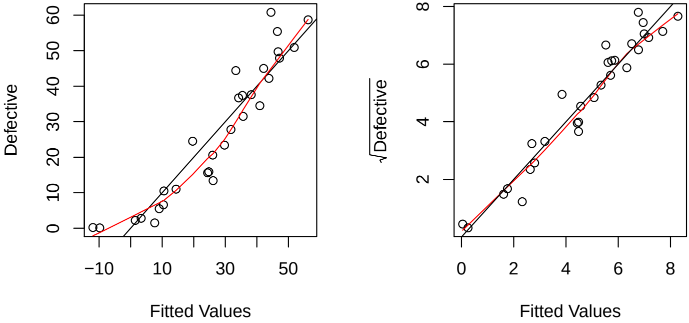
```
> par(mfrow=c(1,2), mar=c(4.5, 4.5, 2, 2))
> plot(lm2, 1:2)
```



```
> par(mfrow=c(1,3), mar=c(4.5, 4.5, 2, 2))
> plot(defects$Temperature, rstandard(lm2),
      xlab="Temperature", ylab="Standardized Residuals")
> plot(defects$Density, rstandard(lm2),
      xlab="Density", ylab="Standardized Residuals")
> plot(defects$Rate, rstandard(lm2),
      xlab="Rate", ylab="Standardized Residuals")
```



Another useful diagnostic is a plot of the observed versus fitted values. Ideally, we should see a straight line relationship (around the 1-1 line). In the plots below, we see that before transforming the response, the relationship is curved; while after transforming the response, the relationship is straight. The loess curve is shown in red.



The code used to generate the last plot:

```
> par(mfrow=c(1,2), mar=c(4.5, 4.5, 2, 2))
> plot(predict(lm1), defects$Defective,
       xlab = "Fitted Values", ylab = "Defective")
> abline(0,1)
> lines(lowess(predict(lm1), defects$Defective), col='red')

> plot(predict(lm2), sqrt(defects$Defective),
       xlab = "Fitted Values", ylab = expression(sqrt(Defective)))
> lines(lowess(predict(lm2), sqrt(defects$Defective)), col='red')
> abline(0,1)
```

Transforming the Predictors

- Inspection of the scatterplot matrix can sometimes indicate whether or not to log transform a predictor.
- Recall, that the log transformation is useful when the data are skewed and/or range over several orders of magnitude.
- Other types of transformations, such as the square root, can also be considered for the predictors.
- Once transformations have been chosen for the predictors, the Box-Cox method can be used to estimate a transformation for the response.

Summary and Remarks

- The Box-Cox method is a useful technique for transforming the response variable.
- Box-Cox transformations can help linearize the relationship between the response and the predictors, and also overcome problems due to nonconstant variability.
- Scatterplot matrices and residual plots should also be used when determining transformations. Do not just rely on automated procedures, also make sure to look at your data.

Summary and Remarks

- We can round λ to the nearest interpretable value. For example, if $\hat{\lambda} = 0.45$ then we can use $\sqrt{y}$ as the transformation. The confidence interval for λ can be used as an aid to determine the rounding.
- Transformations can make the model more difficult to interpret. So if $\hat{\lambda} \approx 1$ and the confidence interval for λ contains 1, a transformation is probably not necessary.
- For $\lambda > 0$, we can just use y^{λ} as the response transformation (no need to use the formal scaling $(y^{\lambda} - 1)/\lambda$).
- If some $y_i \leq 0$ we can add a small constant to make all the y data positive.
- There is usually little justification for making extreme transformations such as $\hat{\lambda} = 5$.

Your turn

Consider the New York City restaurant price data:

```
nyc <- read.csv("https://ericwfox.github.io/data/nyc.csv")
lm2 <- lm(Price ~ Food + Decor + East, data=nyc)
```

Use the `boxcox()` function to plot the log-likelihood and find a confidence interval for λ. Based on these results, does it seem necessary to transform the response?