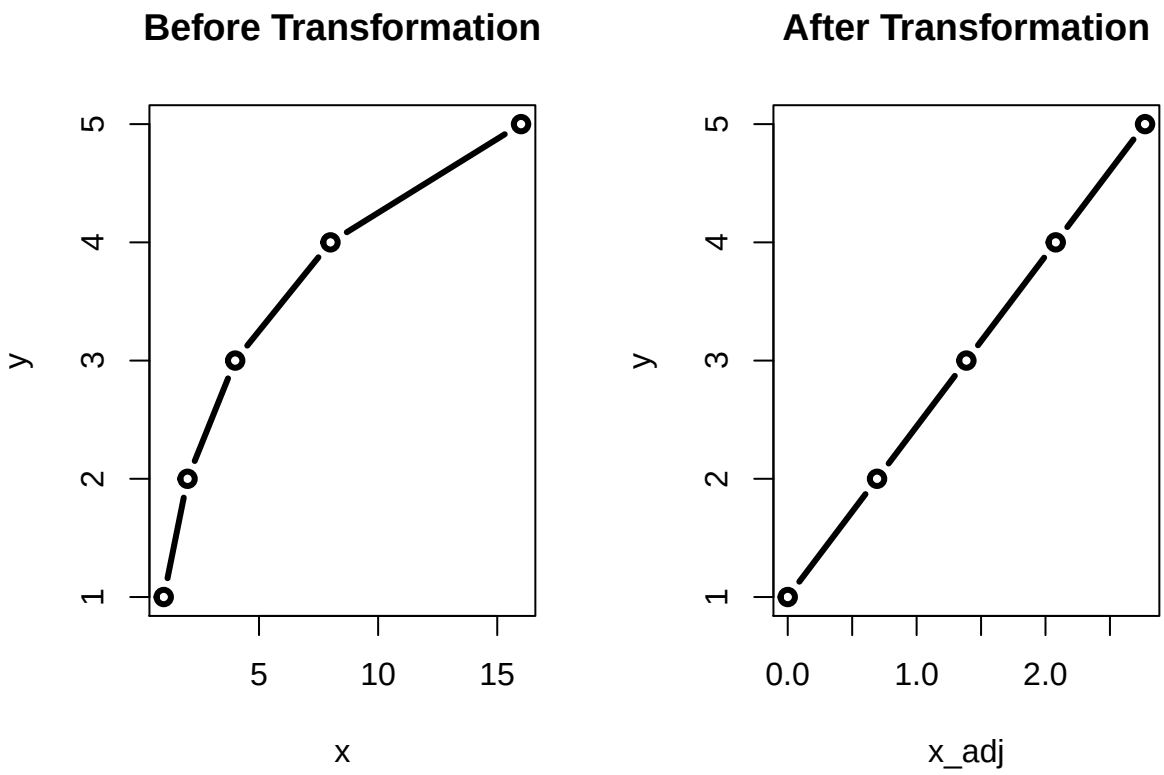
chapter04_lm_4assums

1. linearity

```
par(mfrow=c(1,2))
x = c(1,2,4,8,16)
y = c(1,2,3,4,5)
plot(x, y, type='b', lwd=3, main='Before Transformation')
x_adj = log(x)
plot(x_adj, y, type='b', lwd=3, main='After Transformation')
```



2. homogeneity

```
library(Lahman)
rec = subset(Teams, yearID == 2014)
rec$wp = rec$W / rec$G
a = lm(wp~R, rec)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(a)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  a
## BP = 3.5067, df = 1, p-value = 0.06112
```

It implies strong heteroscedasticity => Do Box Cox Transformation! => We can get a little improvement

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
b = BoxCoxTrans(rec$wp)
c = cbind(rec, wp_adj=predict(b, rec$wp))
d = lm(wp_adj~R, c)
bptest(d)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  d
## BP = 3.3017, df = 1, p-value = 0.06921
```

3. independence

1) Errors of dependent variables not correlated with errors of independent variables.
2) Errors of dependent variables are independent each other dependent variables.

4. normality

ERA and avg models for estimating winning rate

```
library(Lahman)
rec = subset(Teams, yearID == 2014)
rec$wp = rec$W/rec$G
rec$avg = rec$H / rec$AB
avg_model = lm(wp~avg, rec)
ERA_model = lm(wp~ERA, rec)
```

About Residual Standard Error

If we use small residual standard error, we can narrow confidence and prediction interval so that we can improve the decisions. Otherwise, it's not helpful for decision making.

avg_model

```
summary(avg_model)
```

```
##
## Call:
## lm(formula = wp ~ avg, data = rec)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.132044 -0.034014  0.007319  0.035945  0.092524
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1087     0.2457   0.443    0.661
## avg           1.5584     0.9778   1.594    0.122
##
## Residual standard error: 0.05774 on 28 degrees of freedom
## Multiple R-squared:  0.08318,    Adjusted R-squared:  0.05044
## F-statistic:  2.54 on 1 and 28 DF,  p-value: 0.1222
```

ERA model :

```
summary(ERA_model)
```

```
##
## Call:
## lm(formula = wp ~ ERA, data = rec)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.071104 -0.034146  0.002395  0.021711  0.089269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.87071    0.06623  13.148 1.68e-13 ***
## ERA         -0.09917    0.01760  -5.634 4.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04128 on 28 degrees of freedom
## Multiple R-squared:  0.5313, Adjusted R-squared:  0.5146
## F-statistic: 31.74 on 1 and 28 DF,  p-value: 4.918e-06
```