A2: Analysis to Forced Expiratory Volume data

Last name: Akad
First name: Dogan
Student ID: 1001386083
Course section: STA302H1F-L0101

Nov. 17, 2016

Q1: Fit a linear model to original data and looking for transformation using Box-Cox procedure

(a) Looking at the residual plot, residuals don't follow a pattern and they are randomly spread which indicates a linear relationship between age and FEV. Also the variance is not constant because residuals aren't spread across the fitted value. The points follow a line at the scatter plot which also indicates a linear relationship.

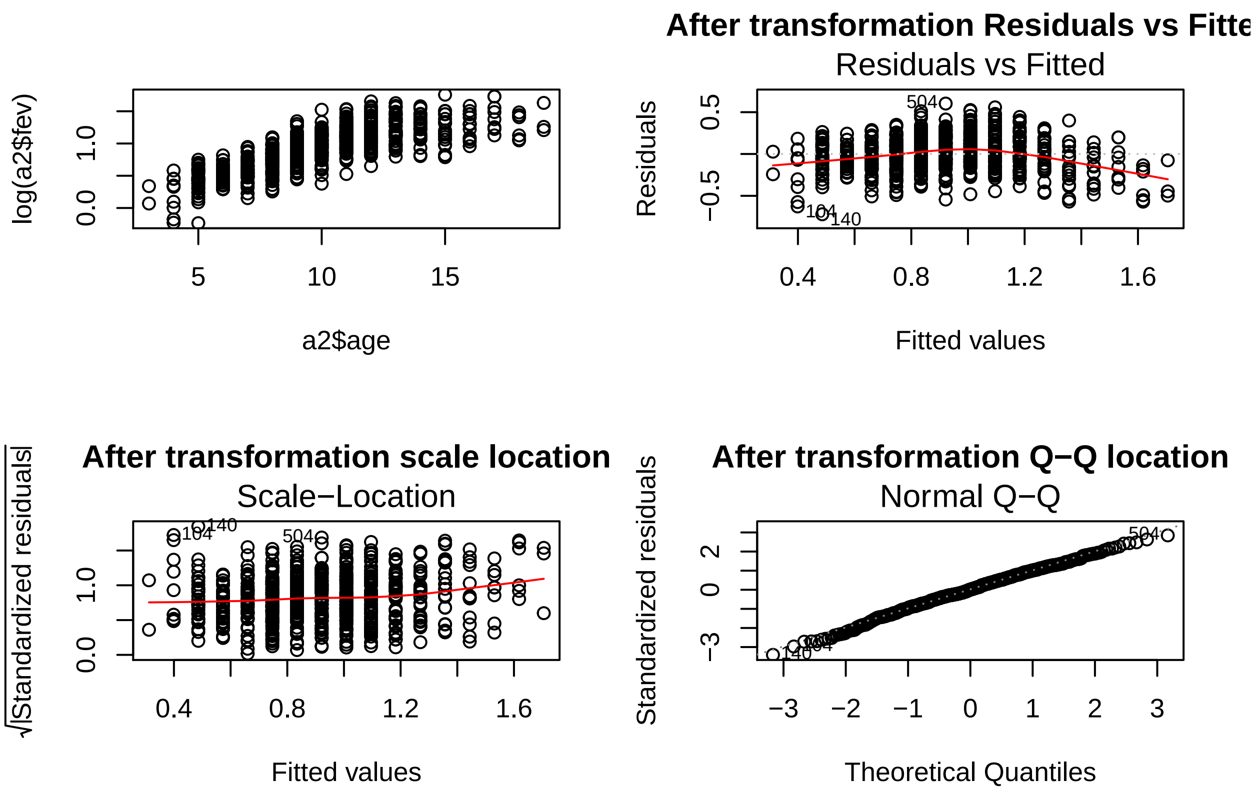
(b) For a simple transformation, the log of FEV(response variable) appears to be the most appropriate because the value of MLE is 0.18 which is close to 0.

Q2: Fit a linear model with transformed FEV and examine the residual plot of the fit.

(a) Estimated model (give the form of f(y) and replace the question mark with estimates)

$$\hat{f}(Y) = 0.0506 + 0.0871 * \text{age}$$

(b)



Comments on plot: Looking at the scale-location plot the residuals appear randomly spread. So, transformation improved adherence to the constant variance assumption. This is also confirmed by residuals vs fitted line plot because the plot is spread out evenly across x-axis.Looking at the residual vs fitted plot, the residuals follow a distinctive pattern. It shows that there isn't a linear relationship. Looking at the normal Q-Q plot, residuals are lined well on the straight dashed line which indicates residuals are normally distributed. The linear model is more acceptable than the former model because two of the assumptions(except linearity) are satisfied but it's not perfect.

(c) Since slope is greater than 0,if age is increased by 1, we expect FEV to increase by 9.1 percent.

(d) 95% Confidence Interval for mean response in untransformed scale when age=c(8,17,21):

$P(2.0705 < \text{FEV} < 2.1527) = 0.95$,

$P(4.4316 < \text{FEV} < 4.8224) = 0.95$,

$P(6.1482 < \text{FEV} < 6.9764) = 0.95$.

95% Prediction Interval for mean response in untransformed scale when age=c(8,17,21):

```
P(1.3916 < FEV < 3.2030) = 0.95,
```

```
P(3.0420 < FEV < 7.0253) = 0.95,

P(4.2982 < FEV < 9.9790) = 0.95
```

Q3

- Estimated model
$$\hat{f}(Y) = -0.98772 + 0.84615 * \log(\text{age})$$

- 95% Confidence Interval for Intercept:

  b0 = [-1.1008,-0.8747]

  95% Confidence Interval for slope:

  b1 = [0.7963,0.8959]

- If age is doubled, we would expect mean of FEV increase by 79.7697 percent.

- For the last model looking at the residual vs fitted plot, the residuals are equally spread around the horizontal line which indicate a linear relationship and the residuals are spread evenly across the fitted value which indicates constant variance. Also looking at the normal Q-Q plot since residuals are lined well on the straight dashed line we can say that residuals are normally distributed. Since for the last model all the assumptions are met and it has lower SSE(looking at the ANOVA analysis) which indicates a better fit than the model on question 2, I prefer using the model on q3.

Q4: Source R code

```
# ---------> complete and run the following code for this assignment  <-------
#
#
# R code for STA302 or STA1001H1F assignment 2
# copyright by YourName
# date: Oct. 26, 2016
#

## Load in the data set
a2 = read.table("/Users/doganakad/desktop/uoft/first semester/sta302/Assignments/A2/a2data.txt",header=T)

## Q1: fit a linear model to FEV on age
age = a2$age
FEV = a2$fev
mod1 = lm(FEV~age)

## ==> Q1(a) produce the scatter plot (FEV vs Age) and the residual plot with fitted value

par(mfrow=c(1,2))
plot(a2$age,a2$fev, type="p",col="blue",pch=21, main="FEV vs age")
abline(mod1,col="red",lty=2)

## Plot residual vs fitted value
plot(mod1,which=1)

##==> Q1(b): boxcox transformation
library(MASS)
bc = boxcox(mod1, lambda = seq(-2,2,0.01))

## boxCox MLE
MLE = bc$x[which.max(bc$y)]
MLE


## Q2

##(a) Estimated Regression Model
mod2 = lm(log(FEV)~age)
par(mfrow=c(2,2))

## Plots
plot(age, log(FEV))
abline(mod2,col="red",lty=2)
plot(mod2,which=1,main="After transformation")

##(c)Value of slope
summary(mod2)

##(d) Confidence and Prediction Interval for the mean response
newdata = data.frame("age" = c(8,17,21))
exp(predict.lm(mod2, newdata, interval="confidence",level=0.95))
exp(predict.lm(mod2, newdata, interval="prediction",level=0.95))

## Q3:

##(a) Estimated regression model
```

```r
mod3=lm(log(a2$fev)~log(a2$age))

##(b) 95% Confidence intervals for intercept and slope of the model in the transformed scale
confint(mod3, level=0.95)

##(c) Value of slope
summary(mod3)

##(d) Produce residual vs fitted plot
plot(mod3,which=1)
plot(mod3,which=2)

## ANOVA analysis
anova(mod3);anova(mod2)
```