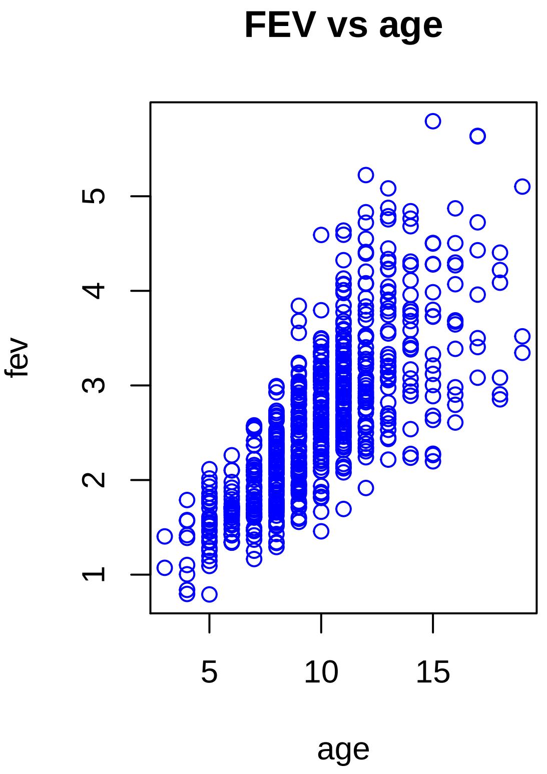
A2: Analysis to Forced Expiratory Volume data

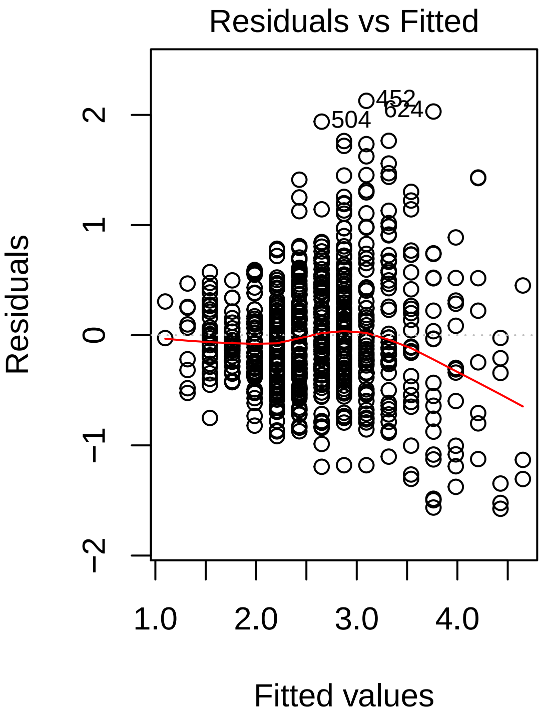
Last name: Zheng
First name: Siyuan
Student ID: 1000726814
Course section: STA302H1F-L0101

Nov. 10, 2016

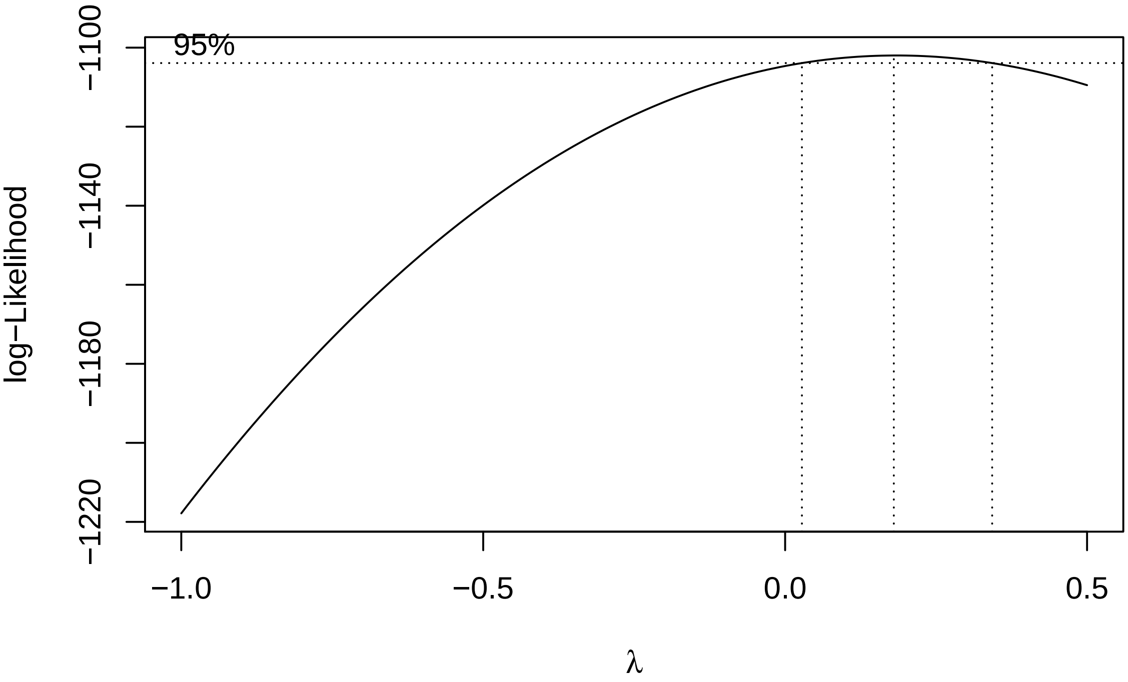
Q1: Fit a linear model to original data and looking for transformation using Box-Cox procedure

- (a) From scatter plots, although it seems that there could be some SLR line to fit the data, but there are many outliars, which are bad leverage points.Thus, fev and age do not have a linear relationship, although they have direct relationship. From residuals vs fitted plots, the points do not evenly spread out across fitted values, thus, again, fev and age do not have a linear relationship.





- (b) By boxcox(), lambda that maximizes log-Likelihood is 0.18, among {-1,0,0.5}, 0 is the closest number.Thus,log seems to be the best transformation, i.e., $Y' = log(Y)$


95%
log−Likelihood
−1100
−1140
−1180
−1220
−1.0
−0.5
0.0
0.5
λ

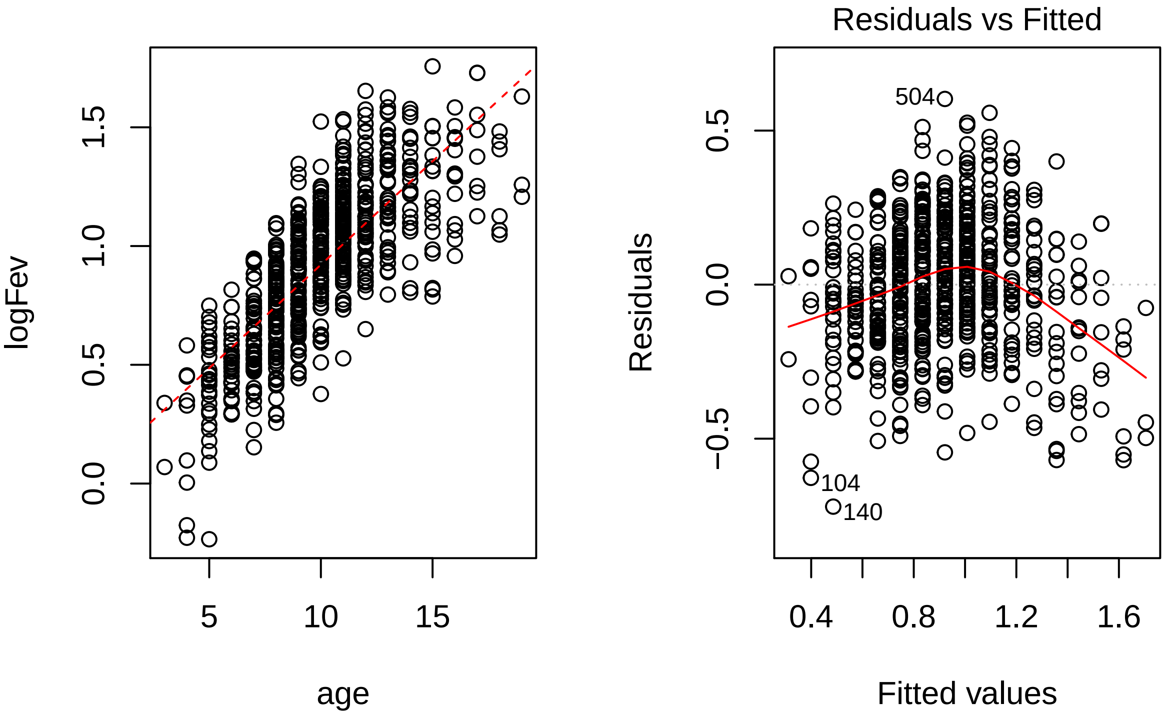
Q2: Fit a linear model wit transformed FEV and examine the residual plot of the fit.

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

$$\hat{f}(Y) = 0.0505960 + 0.0870833\text{age}$$

  ,where $\hat{f}(Y) = log(Y)$

- (b)



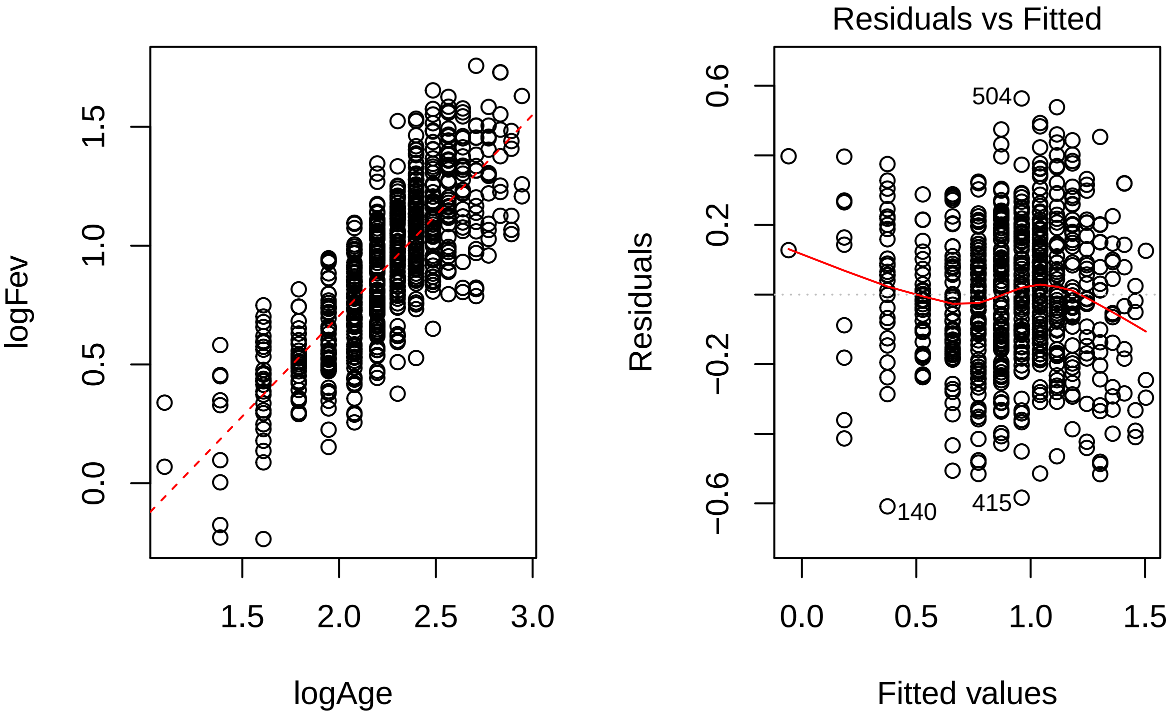
Comments on plot: It does improve adherence to the constant variance, since by residuals vs fitted plots, points tend to spread evenly across fitted values(compared to the one without any transformation). The linear model is acceptable, but not perfect, since by scatter plots, there are not many outliars and data points generally flow along the red line. It is not perfect because there are still some outliars as well as variance can not be strongly considered as constant.

- Q2(c): $Y = e^{log(Y)} = e^{0.0505960+0.0870833age}$
  $\frac{E(Y|X+1)}{E(Y|X)} = \frac{e^{0.0505960+0.0870833(age+1)}}{e^{0.0505960+0.0870833age}} = e^{0.0505960+0.0870833(age+1)-(0.0505960+0.0870833age)} = e^{0.0870833}$, which is e^{slope}, thus, the slope, 0.0870833, is the estimated 'per-age-increment' growth rate for the exponential growth model of fev.

- Q2(d):
  95% CI:
  age 8:[2.070532,2.152692]
  age 17:[4.431587,4.822374]
  age 21:[6.148179,6.976410]
  95 PI:
  age 8:[1.391573,3.203006]
  age 17:[3.041955,7.025340]
  age 21:[4.298236,9.979029]

Q3

- Q3(a)
  $\hat{f}(Y) = -0.9877223 + 0.8461529 log(age)$,
  where $\hat{f}(Y) == log(Y)$
  where Y stands for fev
- Q3(b)
  95% CI for β_0:
  [-1.100753,-0.8746918]
  95% CI for β_1:
  [0.7963774,0.8959283]
- Q3(c)
  Associated with each doubling of age, the mean of fev changes by the the multiplicative factor of $e^{0.8461529 log(2)}$, that is, 1.797700767. As age is doubled, the mean of fev increased by 79.77%.
- Q3(d)



```
## [1] 19
```

```
## [1] 3
```

```
## Analysis of Variance Table
##
## Response: logFev
##            Df Sum Sq Mean Sq F value    Pr(>F)
## logAge      1 45.753  45.753  1114.2 < 2.2e-16 ***
## Residuals 652 26.773   0.041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Response: logFev
##            Df Sum Sq Mean Sq F value    Pr(>F)
## age         1 43.210  43.210  961.01 < 2.2e-16 ***
## Residuals 652 29.316   0.045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on these two plots, I prefer this model. Compared the one of Q2, scatter plot indicates that the the points tend to more flow along the SLR line when age is a little higher(ln(age) and ln(fev) have direct relationship, and ln(X) is a monotonic increasing function). Also. the variance given by residuals vs fitted plot indicates that the tents to be constant when fev is higher.

It means that the model of Q2 fits the younger people within the age range of [3,19] better while the model of Q3 fits the older people within the age range of [3,19] better. I think the model fitting the higher age people is relatively more meaningful.

Also, MSE(Q3's model) == 0.041 < 0.045 == MSE(Q2's model), thus, I prefer this model.

Q4: Source R code

```r
# ---------> 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("a2data.txt",sep="",header=T)

## Q1: fit a linear model to FEV on age
# creates variable fev
fev <- a2$fev
# creates variable age
age <- a2$age
# let fit be the SLR without any transformation
fit = lm(fev~age)

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

par(mfrow=c(1,2))
# generates scatter plot
plot(age,fev, type="p",col="blue",pch=21, main="FEV vs age")
# generates residuals vs fitted plots
plot(fit,which=1)

##==> Q1(b): boxcox transformation
# import relevant library
library(MASS)
# generates boxcox to find lambda
bc= boxcox(fev~age, lambda=seq(-1,0.5,0.01))
# find the lambda that maximizes log-Likelihood
bc$x[which.max(bc$y)]


## Q2(a)
#  take log of fev
logFev <- log(fev)
# generates a new SLR after fev's log transformation
logFevFit <- lm(logFev~age)
# gets the coefficient of
logFevFit$coef
#(b)
#read data
a2 = read.table("a2data.txt",sep="",header=T)
# take log of fev
logFev <- log(a2$fev)
# age
age <- a2$age
# fit data with a SLR model
logFevFit <- lm(logFev~age)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,2))
# scatter plots
plot(age,logFev)
```

```r
# red line to indicate SLR
abline(logFevFit,col="red",lty=2)
# residuals vs fitted values plots
plot(logFevFit,which=1) # which=2: for the Normal QQ-plot
#(d)
# 95% CI
CI <- predict.lm(logFevFit,newdata=data.frame(age=c(8,17,21)),interval='confidence',level=0.95)
# untransformed
exp(CI)
# 95% PI
PI <- predict.lm(logFevFit,newdata=data.frame(age=c(8,17,21)),interval='prediction',level=0.95)
# untransformed
exp(PI)

## Q3:
# take log of age
logAge <- log(age)
# generates SLR after both variables' transformation
logFevLogAgeFit <- lm(logFev~logAge)
# get the coefficients of SLR
logFevLogAgeFit$coef
# find the 95% CI for coefficients
confint(logFevLogAgeFit, '(Intercept)', level=0.95)
confint(logFevLogAgeFit, 'logAge', level=0.95)
# (d)
# read data
a2 = read.table("a2data.txt",sep="",header=T)
# take log over age
logAge <- log(a2$age)
# take log over fev
logFev <- log(a2$fev)
# get SLR
logLogFit <- lm(logFev~logAge)
# 1 by 2 plots in one window
par(mfrow=c(1,2))
# scatter plots
plot(logAge,logFev)
# red line to indicate SLR
abline(logLogFit,col="red",lty=2)
# residuals vs fitted values plots
plot(logLogFit,which=1)
# get max age of data.txt
max(age)
# get min age of data.txt
min(age)
# to get the MSE of Q3
anova(logLogFit)
# to get the MSE of Q2
anova(logFevFit)
```