A2: Analysis to Forced Expiratory Volume data

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

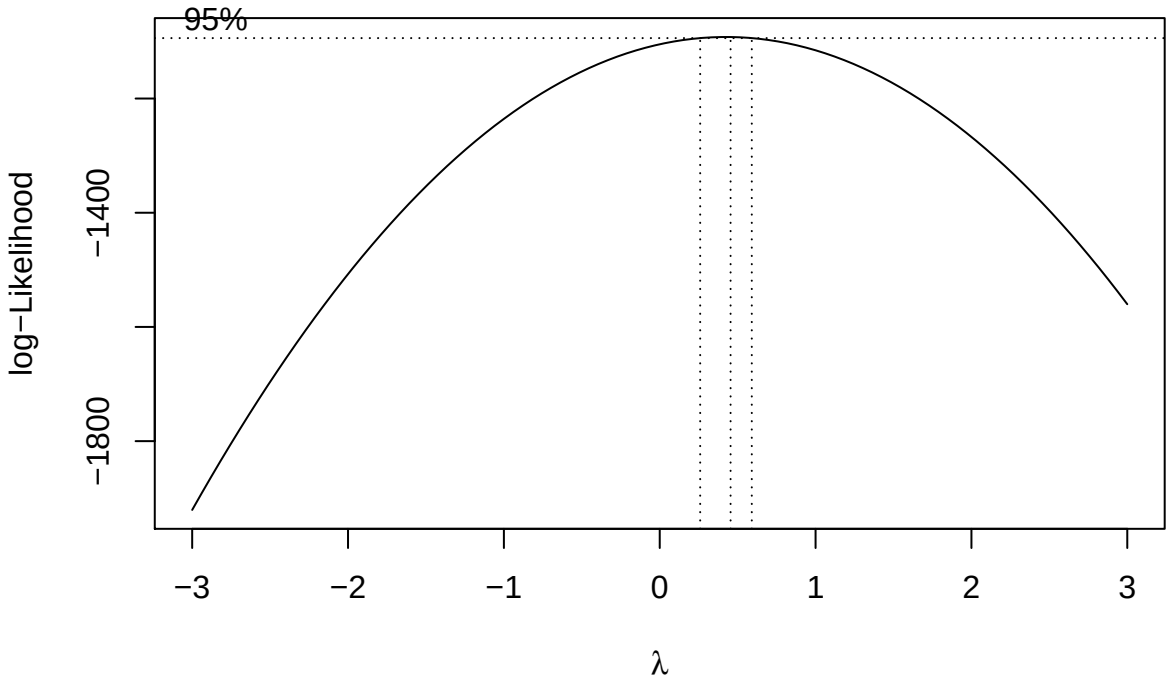
Nov. 09, 2016

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

(a)These two plots give us the information that the data generally have direct relationship, i.e., the older a person is, the better lung capacity he/she would have.But there is no linear relationship between a person's lung capacity and age, since the residual plot does not follow the rule that "randomly spread without a clearly pattern".

(b)The $\hat{\lambda}$ that maximizes the log-likelihood is 0.4545455, thus, choose $\hat{\lambda}$ as $0.5.Y' = \sqrt{Y}$ is the best transformation.

```
a2 = read.table("a2data.txt",header=T)
fev <- a2$fev
age <- a2$age
library(MASS)
#adjust the panel back to 1 by 1
par(mfrow=c(1,1))
# generate boxcox graph to see different lambda's impacts
bc= boxcox(fev~age,lambda=c(-3, 3,by=0.01))
```



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)

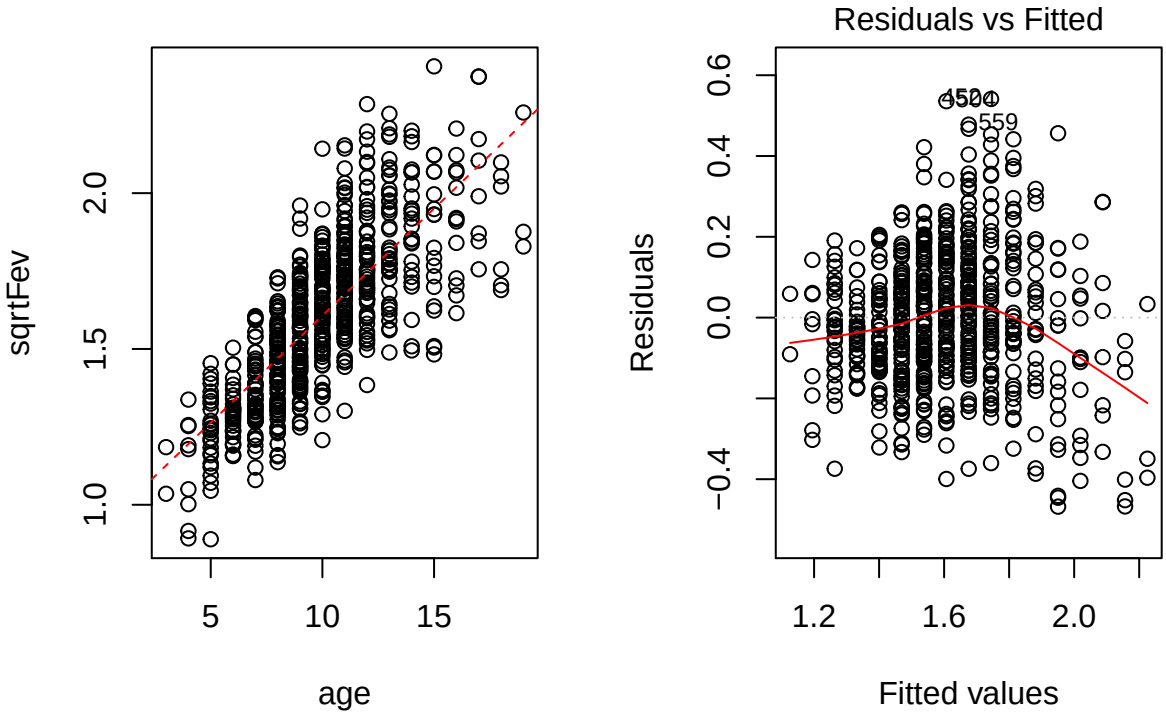
$$\hat{f}(Y) = 0.91999354 + 0.06869976\text{age}$$

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

- (b)

```
# put echo=FALSE will only give the plot wihout code in output
#
# for Rmarkdown, R chunks are independent.
# so every time you want to produced a plot, you need load in the data and run a complete
# code for each question or part of a question.

# sqrtFev is the square-rooted fev value
sqrtFev <- sqrt(fev)
# fit data with a SLR model
sqrtFit <- lm(sqrtFev~age)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,2))
plot(age,sqrtFev)
abline(sqrtFit,col="red",lty=2)
plot(sqrtFit,which=1) # which=2: for the Normal QQ-plot
```



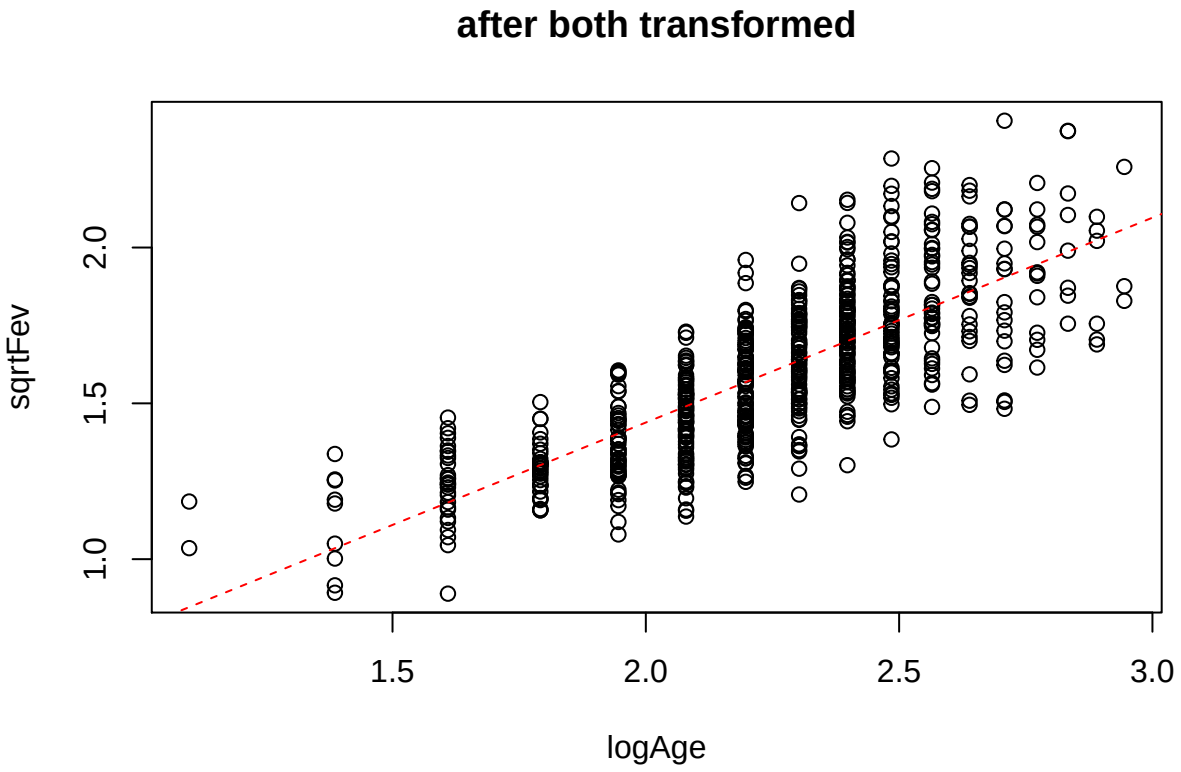
Comments on plot: It has not improved and this linear model is not acceptable. Since by the residual plots,the data does not spread evenly across fitted values as well as the red line indicates that the pattern of these points is not falling along horizontal straight line, it gives us the information that the variance is not constant and the relationship is not linear.

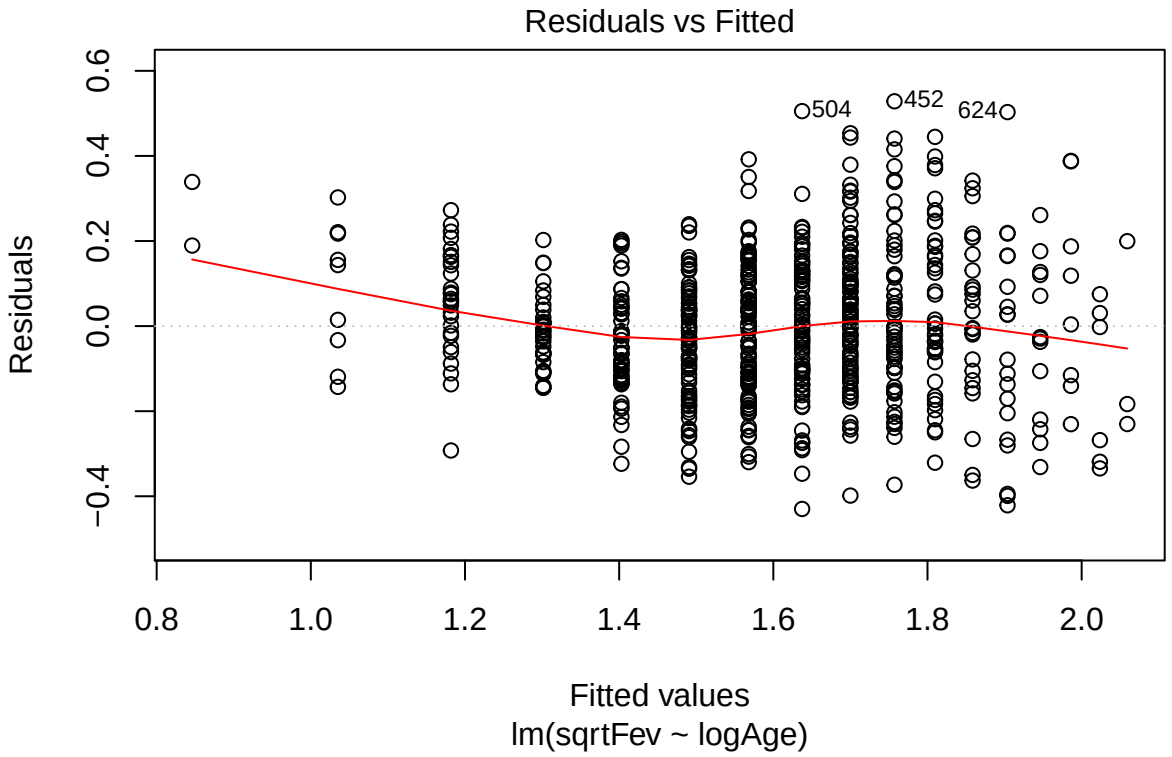
- Q2(c), Interpretation of slope:When age increases by 1, the square root of fev will increases by 0.06869976.Interpretation of SLR: $E(\sqrt{Y}|X) = 0.91999354 + 0.06869976X$, where X is age and Y is fev ,thus $E(Y|X) = (0.91999354 + 0.06869976X)^2$.$E(Y|X+1) - E(Y|X) = (0.91999354 + 0.06869976(X+1))^2 - (0.91999354 + 0.06869976X)^2) = 0.1337 + 0.0098X$.When age increases by 1 from age to age + 1, the fev will increases by 0.1337 + 0.0098 times the former age(X).
- Q2(d), 95% CI: age 8: [2.155901 2.260051] age 17:[4.093253 4.319436] age 21: [4.925386 5.263631]

95% PI: age 8: [1.092359,3.323593] age 17:[3.086220,5.326470] age 21: [3.967347,6.221670]

Q3

- Q3(a), $\hat{f}(Y) = 0.1240271 + 0.6572306 \times log(age)$, where $\hat{f}(Y) = \sqrt{Y}$
- Q3(b), 95% CI for β_0 [0.03168476,0.2163695] 95% CI for β_1 [0.6165656,0.6978956]
- Q3(c), Interpretation of slope:In a simpler way, associated with each e-fold increase of age, the square root of fev will increase by 0.0098.Interpretation of SLR:since $log(eX/X) = log(e) = 1$, thus, combined with the interpretation of question 2(c). We conclude that associated with each e-fold increase (where e is the natural base number) of age, the fev will increase by 0.1337 + 0.0098 times the ln(age).
- Q3(d)





Comment: Based on the presented result, i.e, scatter plots and the residuals vs fitted plots, I consider the log-transformed age plus square-root-transformed fev is a better model to indicate the linear relationship and build the SLR. Since from the scatter plots, data tends to more follow along a straight line and they are more close to the 'fitted' line , as indicated by the red line,compared to the square-root fev transformed model. Also, the variance tends to be more stable than q2's model, as the fitted values increases, which is opposite to what happen for q2's residuals vs fitted plots.

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
#


# To distinguish the output generated by the codes written on Q1-Q3 above..
rm(list = ls())
dev.off()

## Load in the data set
a2 = read.table("a2data.txt",header=T)

## Q1: fit a linear model to FEV on age

mod1 = lm(a2$fev~a2$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")
plot(mod1,which=1)

##==> Q1(b): boxcox transformation
fev <- a2$fev
age <- a2$age
library(MASS)
#adjust the panel back to 1 by 1
par(mfrow=c(1,1))
# generate boxcox graph to see different lambda's impacts
bc= boxcox(fev~age,lambda=c(-3, 3,by=0.01))
# lambda value that maximizes the log-likelihood function
bc$x[which.max(bc$y)]


## Q2
# create the square-rooted fev data set
sqrtFev <- sqrt(fev)
# use SLR to generate a model for transformed fev
sqrtFit <- lm(sqrtFev~age)
# get the coefficient of SLR
sqrtFit$coef
# create the SLR untransformation model
fit <- lm(fev~age)
# get the 95% CI
predict.lm(fit,newdata=data.frame(age=c(8,17,21)),interval='confidence',level=0.95)
# get the 95% PI
predict.lm(fit,newdata=data.frame(age=c(8,17,21)),interval='prediction',level=0.95)

## Q3:
# get the log age values...
logAge <- log(age)
# get the SLR after transformation of age and fev
```

```r
sqrtLogFit <- lm(sqrtFev~logAge)
# get the coefficients
sqrtLogFit$coef
# find the 95% CI for coefficients
confint(sqrtLogFit, '(Intercept)', level=0.95)
confint(sqrtLogFit, 'logAge', level=0.95)
# find the scatter plot and the residuals vs Fitted plots
# scatter plots
plot(logAge,sqrtFev)
# add the red line to the scatter plot
abline(sqrtLogFit,col="red",lty=2)
# residuals vs fitted plots
plot(sqrtLogFit,which=1)
```