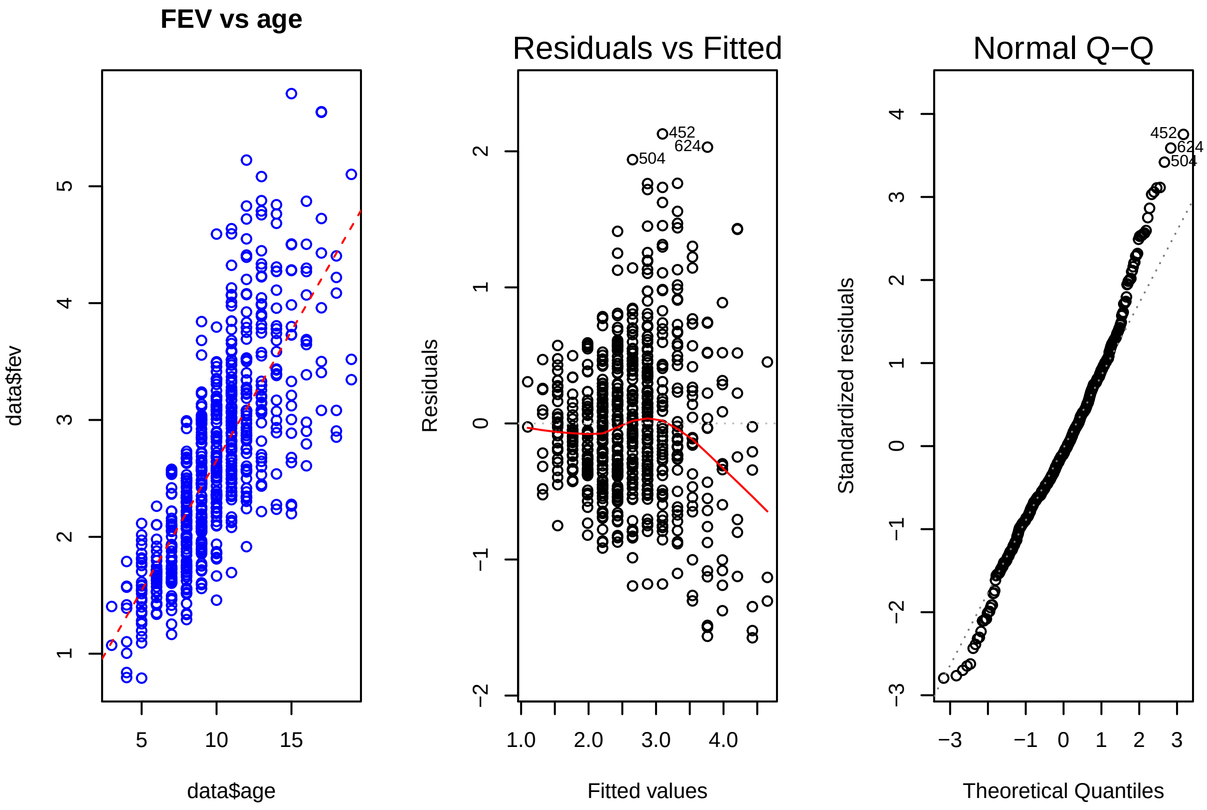
A2: Analysis to Forced Expiratory Volume data

Last name: Tao
First name: Tianwen
Student ID: 999726541
Course section: STA302H1F-L0101

Oct. 29, 2016

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

```
# Loading data from txt file
data = read.table("/Users/leotao/Downloads/DataA2.txt",sep=" ",header=T)
x = data[,1]
y = data[,2]
# fit data with a SLR model
m = lm(y~x)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,3))
plot(data$age,data$fev, type="p",col="blue",pch=21, main="FEV vs age")#scatter
abline(m,col="red",lty=2)
plot(m,which=1) # Residual plot
plot(m,2)
```



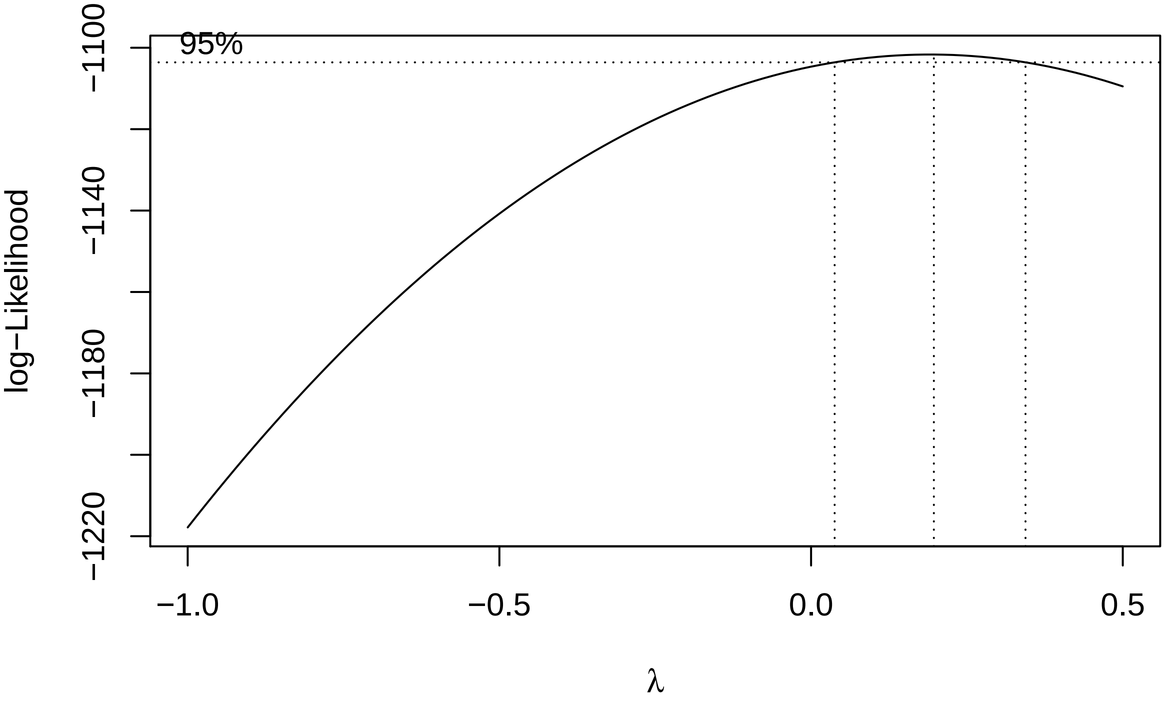
```r
# get R-squared value of the data
summary.lm(m)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57539 -0.34567 -0.04989  0.32124  2.12786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
## x           0.222041   0.007518  29.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5675 on 652 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5716
## F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

1) As you can tell from the graph, the points are not near the line and the residual VS fitted plot massive scattered. The R-squared value is only 0.5722 which indicates it is poorly explained allmost half of the data points. However, since the P value is quite small, we can tell that there is linear relationship with FEV and age. Furthermore, the variance is quite large, so even in the same age, there are a lots children have different FEV value. The result is make ssense in real life since most people from different area would have different FEV even though they are in same age. But the linear relationship tell us that with age increasing FEV will also increasing. As the age increase the variance's amplitude is keep increasing from the residuals plot, so we use Y-transformation to improve the data set.

Q1/b)

```
# Loading data from txt file
data = read.table("/Users/leotao/Downloads/DataA2.txt",sep=" ",header=T)
library(MASS)
bc=boxcox(y~x, lambda = c(-1,0,0.5))
```



```
lambdahat= bc$x[which.max(bc$y)]
lambdahat
```

```
## [1] 0.1969697
```

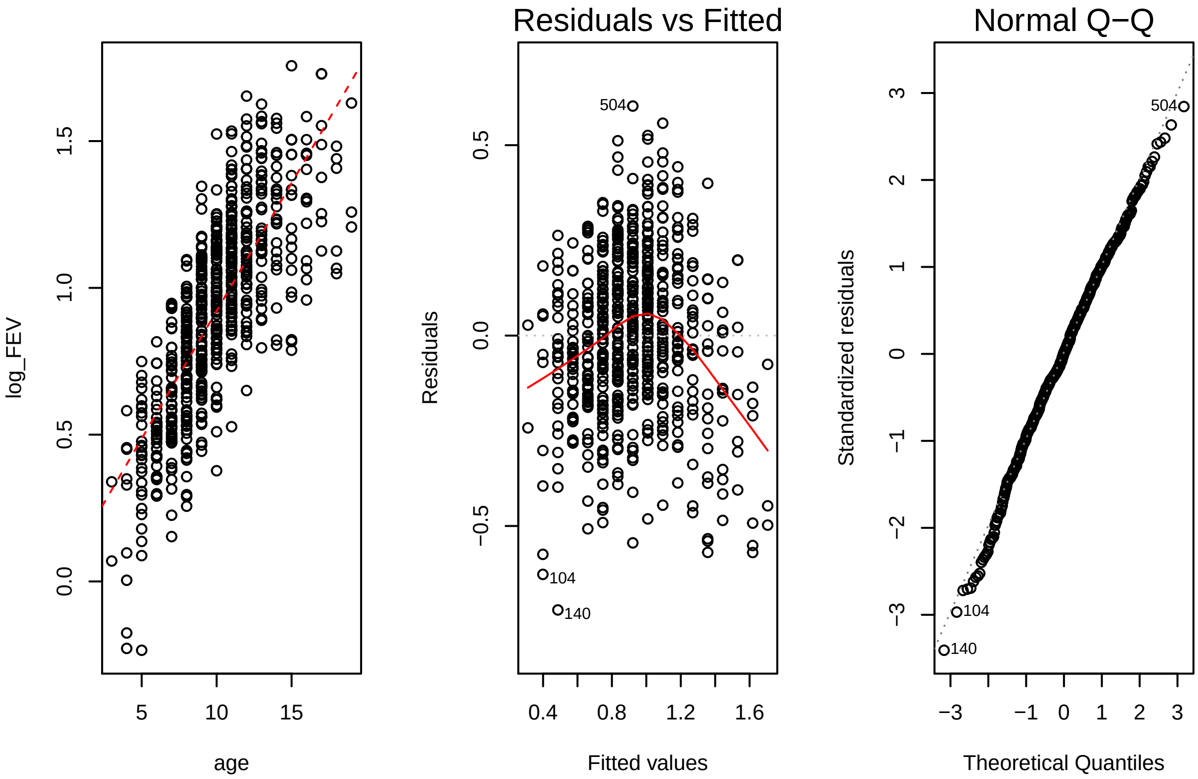
2) From the above graph, we can tell the maximum of the likelihood is reached when transformation power equals 0.197, since 0 is more near to the peak compare with 0.5, so it is better to use log-transformation.

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

Type your concise and clear answer here.

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

$$\hat{f}(Y) = \mathrm{e}^{0.051+0.087\mathrm{age}}$$



```
##
## Call:
## lm(formula = log_FEV ~ age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72047 -0.13752  0.00302  0.14681  0.60267
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.050596   0.029104   1.738   0.0826 .
## age         0.087083   0.002809  31.000   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.212 on 652 degrees of freedom
## Multiple R-squared:  0.5958, Adjusted R-squared:  0.5952
## F-statistic:   961 on 1 and 652 DF,  p-value: < 2.2e-16
```

- your answer to Q2(b).

Comments on plot: As the plot shows, although the result is still not vary convincing, it did improve a little bit from 0.5722 to 0.5958 compare with previous data set. Still there is linear relationship between age and FEV. Further more the amplitude of Residual no longer increasing as the age increase. So the model is relatively acceptable compare with previous one.

- your answer to Q2(c)

It is Log-Level model, with the : As age increases by 1, the mean of FEV increases by

$$\mathrm{e}^{0.051} - 1$$

With age increasing the FEV will increase as well.

- your answer to Q2(d)

Cofidence Interval

Age	fit	lwr	upr
8	2.111212	2.070532	2.152692
17	4.622853	4.431587	4.822374
21	6.549215	6.148179	6.976410

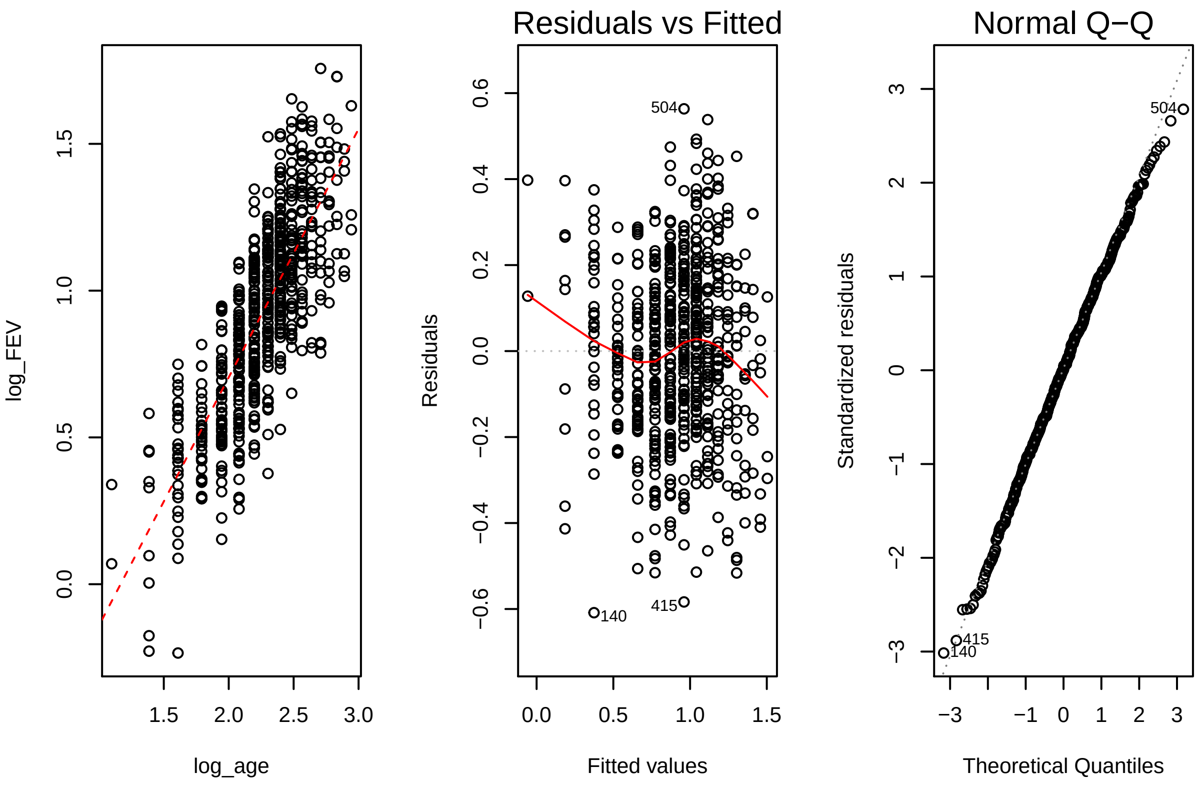
Prediction intervals

Age	fit	lwr	upr
8	2.111212	1.391573	3.203006
17	4.622853	3.041955	7.025340
21	6.549215	4.298236	9.979029

Q3

- your answer to Q3(a)

$$\hat{f}(log(Y)) = -0.988 + 0.846log(\text{age})$$



```
##
## Call:
## lm(formula = log_FEV ~ log_age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60857 -0.13532  0.00227  0.14329  0.56348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.98772    0.05756  -17.16   <2e-16 ***
## log_age      0.84615    0.02535   33.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2026 on 652 degrees of freedom
## Multiple R-squared:  0.6309, Adjusted R-squared:  0.6303
## F-statistic:  1114 on 1 and 652 DF,  p-value: < 2.2e-16
```

- your answer to Q3(b)

	2.5 %	97.5 %
(Intercept)	-1.1007528	-0.8746918
log_age	0.7963774	0.8959283

- your answer to Q3(c)

It is Log-Level model, with the : As X increases by 1, the mean of Y increases by

$$\mathrm{e}^{0.846*log(2)} - 1$$

With age increasing the FEV will increase as well.

- your answer to Q3(d)

I would prefer this model compare with model in Q2(a):

-1. Higher R-squared value, which indicate this model explain more data points compare with previous model

-2. The SSE for model in Q2 is 294, and the SSE for model in Q3 is 280 which is better.

So choose model in 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
data = read.table(data = read.table("/Users/leotao/Downloads/DataA2.txt",sep=" ",header=T)

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

## ==> Q1(a) produce the scatter plot (FEV vs Age) and the residual plot with fitted value
x = data[,1]
y = data[,2]
# fit data with a SLR model
m = lm(y~x)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,2))
plot(data$age,data$fev, type="p",col="blue",pch=21, main="FEV vs age")#scatter
abline(m,col="red",lty=2)
plot(m,which=1) # Residual plot
# get R-squared value of the data
summary.lm(m)


##==> Q1(b): boxcox transformation
# Loading data from txt file
library(MASS)
bc=boxcox(y~x, lambda = c(-1,0,0.5))
lambdahat= bc$x[which.max(bc$y)]
lambdahat


## Q2
#===>(a)
# fit data with a SLR model
# Loading data from txt file
data = read.table("/Users/leotao/Downloads/DataA2.txt",sep=" ",header=T)
age = data[,1]
FEV = data[,2]
log_FEV = log(y)
m2 = lm(log_FEV~age)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,3))
plot(age,log_FEV)
abline(m2,col="red",lty=2)
plot(m2,which=1)
plot(m2,2)
#Analysis with R-square
summary.lm(m2)

#===>(b)
ages=c(8,17,21)
value = predict.lm(m2,newdata=data.frame(age = ages), interval="confidence")
```

```r
#Change back
exp(value)
value = predict.lm(m2,newdata=data.frame(age = ages), interval="prediction")
#Change back
exp(value)


## Q3:
#===>(a)
# fit data with a SLR model
# Loading data from txt file
data = read.table("/Users/leotao/Downloads/DataA2.txt",sep=" ",header=T)
age = data[,1]
FEV = data[,2]
#Log both variables
log_FEV = log(y)
log_age = log(age)
m3 = lm(log_FEV~log_age)
# plot the scatter plot and residual plot in one panel
par(mfrow=c(1,3))
plot(log_age,log_FEV)
abline(m3,col="red",lty=2)
plot(m3,which=1)
plot(m3,2)
#Analysis with R-square
summary.lm(m3)

#===>(b)
confint(m3,level=1-0.05)

#===>(d)
# calculate SSE for model in Q3
data3 = predict(m3)
#build vactor to store value
sse = rep(NA,length(FEV))
for(i in 1:length(FEV)){
  #calculate model error
  sse[i] =  abs(exp(data3[i])-FEV[i])
}
#all error after untransformtion for model 3 which is 280
sum(sse)

# calculate SSE for model in Q2
data2 = predict(m2)
sse2 = rep(NA,length(FEV))
for(i in 1:length(FEV)){
  #calculate model error for model 2
  sse2[i] =  abs(exp(data2[i])-FEV[i])
}
#all error after untransformtion for model 2 which is 294
#larger than model 3, so model 3 is better
sum(sse2)
```