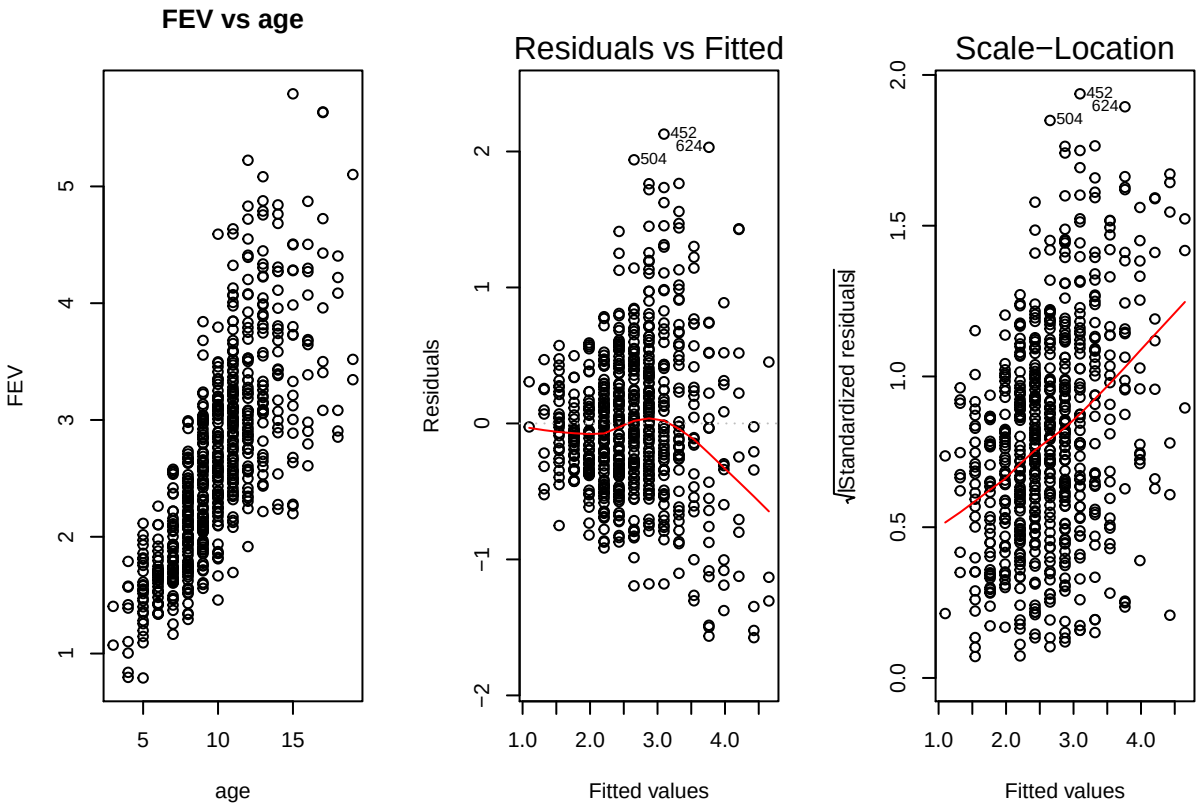
A2: Analysis to Forced Expiratory Volume data

Last name: LastName
First name: FirstName
Student ID: 00000000
Course section: STA302H1F-L0101

Due: 11pm, NOV 17, 2016

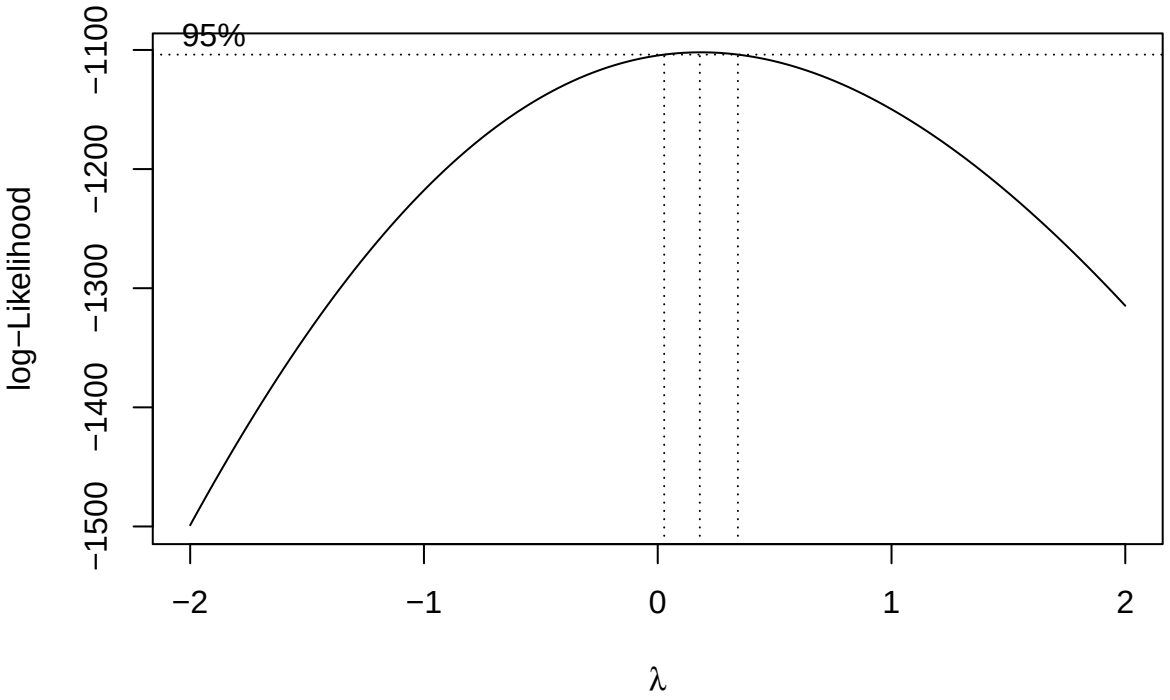
Q1: (5+5=10 pts) Fit a linear model to original data

Q1-a: Scatter plot and residual plot



Comments: according to the scatter plot, it indicates that there is a linear association between FEV and age. But we also observed that as the age variable increases, so does the variance in the response variable (FEV). The non-constant variance of FEV can be further confirmed in the plot of residuals (before and after standardized) versus fitted values. Furthermore, from the residual plot (either before or after standardized), there are three observations (452, 504, and 624) have larger residuals (their standarized residuals are close to 2).

Q1(b): use R function boxcox() to find a simple power transformation




```
## [1] 0.18
```

Comments: using the Box-Cox method, the MLE of the power parameter (λ) is 0.18, and the 95% confidence interval for λ is closer to 0 than 1, therefore the log of the response variable (FEV) appears to be the most appropriate.

On why we chose log transformation, the following argument is also acceptable

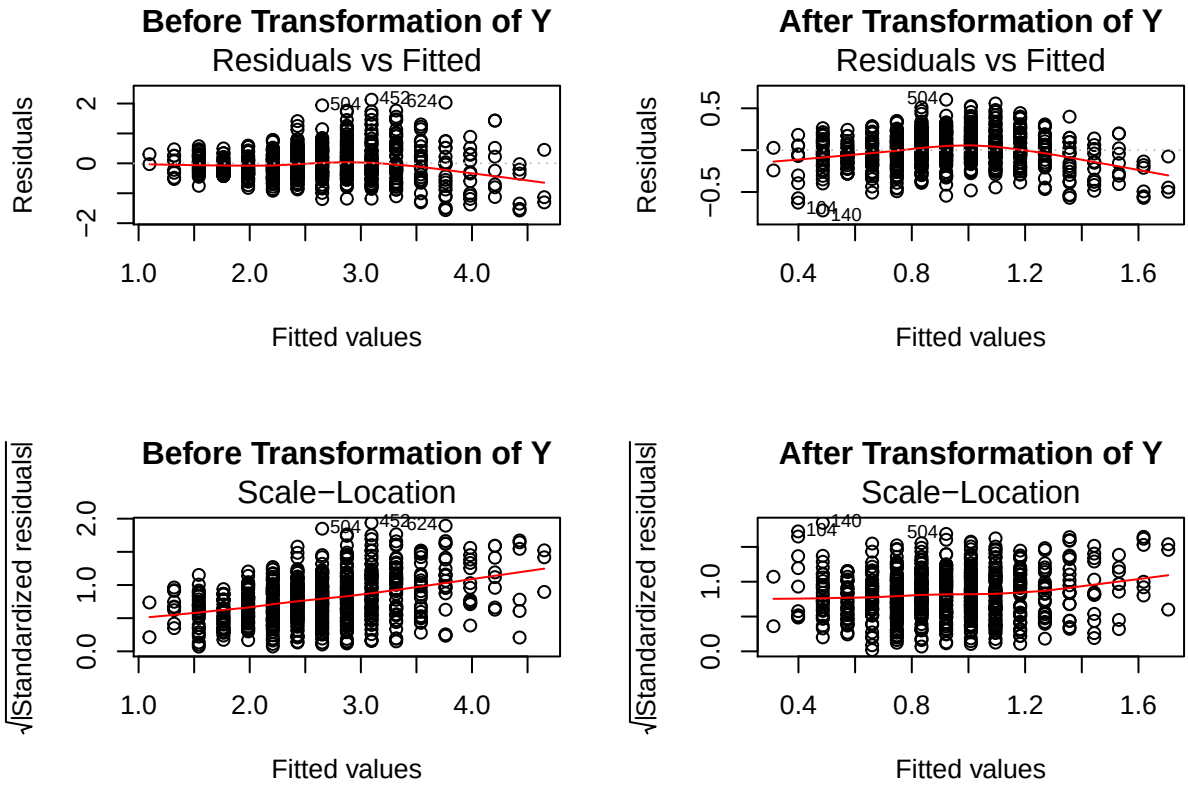
For Box-Cox procedure, the SSE is fairly stable in a neighbourhood of $\hat{\lambda}$. From given data, the estimate of the power parameter is 0.18 which is closer to 0 than 1. Therefore, a logarithmic transformation on the response variable (FEV) was chosen.

Q2: Q2: (3+2+3+3=11 pts) Fit a linear model with log(FEV) predicted by age

- Q2-a: Estimated model using log(FEV) as the response.

$$\widehat{\log(FEV)} = 0.050596 + 0.087083 \text{ age}$$

- Q2-b: Examine the model with transformed FEV



Comments: according the scale-location plot, the standardized residuals vs fitted plot before and after the logarithmic transformation of FEV, we observed that before transfomation the standardized residual increases as the fitted value increases but it was stablized a lot after log transformation of FEV. The constant variance assumption is better met under the log transformation. However, we observed a curvilinear in the raw residual (before standardization) diagram that was not there in data fitting before the log transformation. But overall, the transformation has improved adherence to the constant variance assumption.

We further examine the scatter plot and the Normal QQ-plot after the log transformation, the linearity and and normality assumptions look fine. As we dicussed, the constant variance assumption is improved but not well satisfied after the log transformation. Although I believe this linear model is acceptable, it may not be optimal.

- Q2-c: How do you interpret the slope?

Assume this model is acceptable. It is estimated that, on average, FEV increases (slope estimate is positive) by 9.098755% (i.e. $e^{0.087083} - 1$) if age increases by one year.

- Q2-d: Find the 95% CI for $E(Y_h)$ and 95% PI for Y_h when age=c(8, 17,21)

Age	Fitted value $\hat{Y}_h$	95% CI for $E(Y_h)$	95% PI for Y_h
08	2.111212	(2.070532, 2.152692)	(1.391573, 3.203006)
17	4.622853	(4.431587, 4.822374)	(3.041955, 7.025340)
21	6.549215	(6.148179, 6.976410)	(4.298236, 9.979029)

For given data set, the range of variable **age** is: [3,19], the confidence interval and prediction interval at **age**=21 is outside of the range, we can't be sure that the linear model is still appropriate at this age level, we should be cautious when we use the CI and PI at this level of age.

Q3: (3+4+3+4=14 pts) Fit SLR with log(FEV) predicted by log(age)

- Q3-a: the estimated model in terms of transformed data

$$\widehat{\log(FEV)} = -0.98772 + 0.84615 \log(age)$$

- Q3-b: Find 95% CIs for intercept and slope in the transformed scale

Parameters	95% CI
β_0	(-1.1007528, -0.8746918)
β_1	(0.7963774, 0.8959283)

- Q3-c: Interpret the slope of model in Q3-a

This is the log-log regression model. The following two interpretations are acceptable:

(1) Associated with each doubling of age, it is estimated that FEV will change by the multiplicative factor of 1.7977 (i.e. $e^{0.84615*\log(2)}$).

(2) Associated with each doubling of age, it is estimated that there is a 79.77% (i.e. $e^{0.84615*\log(2)} - 1$) increases in the mean value of FEV.

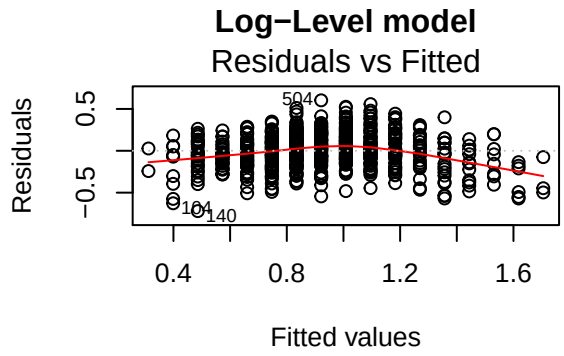
- Q3-d: log-level vs log-log, which model is better?

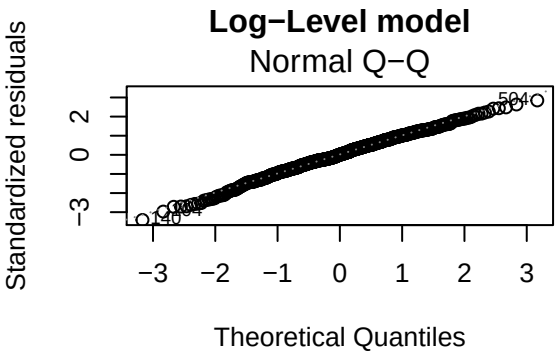
model	R^2	SSE (in original scale)	MSE (in log scale)	AIC
Log-level (Q2-a)	59.58%	241.2044	0.044939	-168.6845
Log-log (Q3-a)	63.09%	210.0379	0.041104	-228.0325

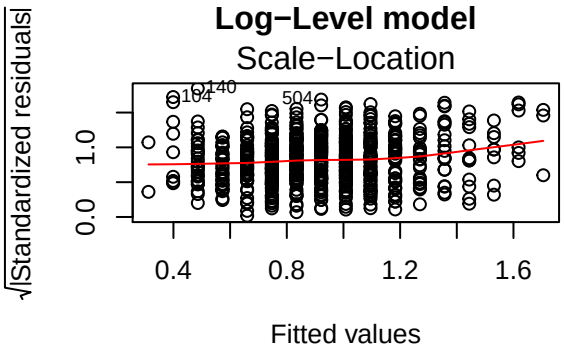
- where $SSE = \sum_1^n (Y_i - \hat{Y}_i)^2$, $\hat{Y}_i = \exp\{\hat{Y}_i'\}$ and $Y_i' = log(Y_i)$
- $MSE = SSE/df(SSE)$
- $R^2 = 1 - \frac{SSE}{SSTO}$ but this is calculated under the log-transformed scale.
- Akaike information criterion (AIC): is a measure of the relative quality of statistical models for a given set of data. The lower AIC value the better.
- Side note: for the comparison of these two models, we do the same transformation on FEV, so a direct comparison of SSE, MSE or R^2 are acceptable and they should lead to the same conclusion.

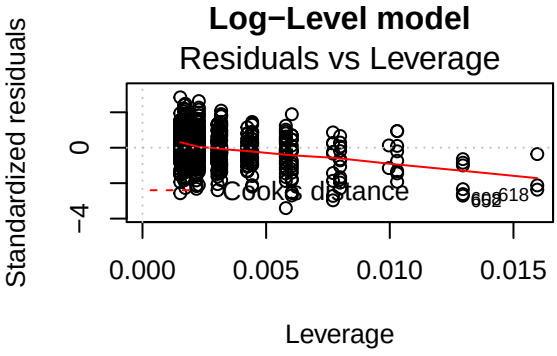
Diagnostic plots of both models suggest that he linearity and normality assumptions look fine for both models. But it seems that the constant vaiance is better met in the log-log model.

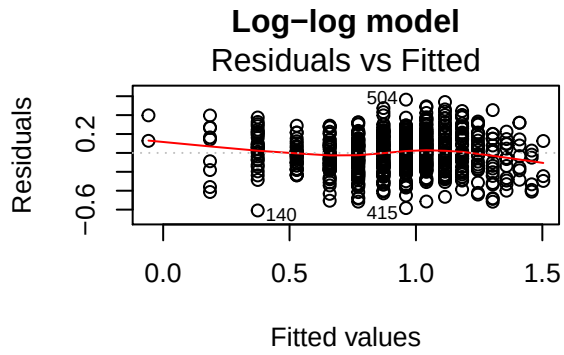
Further, we use R^2 (in log scale), or SSE (in original scale), MSE or AIC as criteria, we observed that the log-log model has higher R^2, smaller SSE and AIC. Therefore the log-log model is chosen as a better model because it satisfies the model assumptions and appears to be linear with constant variance over most of the interval.

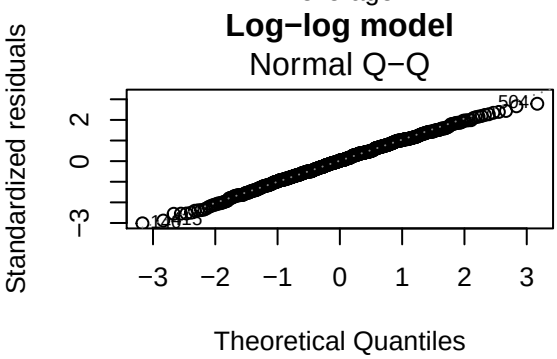

Log−Level model
Residuals vs Fitted
Residuals
Fitted values
504
104
140

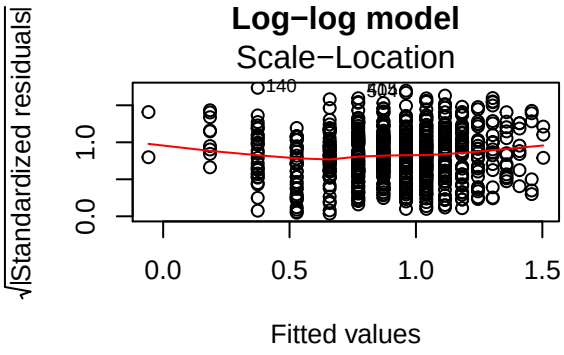

Log−Level model
Normal Q−Q
Standardized residuals
Theoretical Quantiles
504

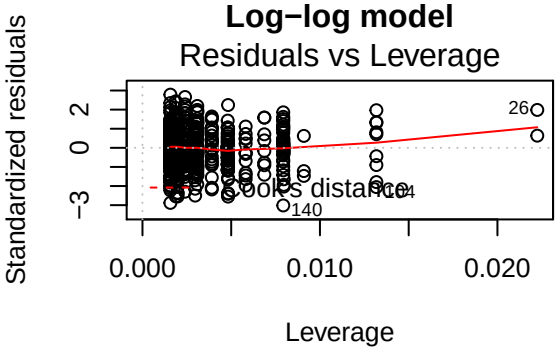

Log−Level model
Scale−Location
√|Standardized residuals|
Fitted values
104
140
504


Log−Level model
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Log−log model
Residuals vs Fitted
Residuals
Fitted values
504
140
415


Log−log model
Normal Q−Q
Standardized residuals
Theoretical Quantiles
504


Log−log model
Scale−Location
√|Standardized residuals|
Fitted values
140


Log−log model
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
26
140

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/Wei/TA/Teaching/0-STA302-2016F/HW/A2/DataA2.txt",header=T)
dim(a2)          # see the data size and number of variables
str(a2)          # structure of an object
class(a2)        # class or type of an object
names(a2)        # names

## Q1: fit a linear model to FEV on age
a2 = read.table("/Users/Wei/TA/Teaching/0-STA302-2016F/HW/A2/DataA2.txt",header=T)
mod1 = lm(fev~age,data=a2)

## ==> 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="black",pch=21, main="FEV vs age")
plot(mod1,which=1)

##==> Q1(b): boxcox transformation
library(MASS)
bc=boxcox(mod1,lambda=seq(-2,2,by=0.01))
bc$x[which.max(bc$y)]  # MLE: estimation of lambda

## Q2: log(FEV)~age

# === Q2-(a): estimated model
mod2 = lm(log(fev)~age,data=a2)
summary(mod2)

# === Q2-(b): examine the scatter plot before and after transformation
par(mfrow=c(1,2))
plot(fev~age,data=a2,xlab="age",ylab="FEV",main="Raw data")
plot(log(fev)~age,data=a2,xlab="age",ylab="log(FEV)",main="log(FEV)")

par(mfrow=c(1,2))
plot(mod1,which=1,main="Raw data")
plot(mod2,which=1,main="log(FEV)")
# investigate the residual plots before and after
par(mfrow=c(2,2))
plot(mod1,which=3,main="Raw data")
plot(mod2,which=3,main="log(FEV)")
plot(mod1,which=2,main="Raw data")
plot(mod2,which=2,main="log(FEV)")

# === Q2-(c): how to interprete the slope if assume this model is acceptable
exp(mod2$coef[2])-1

# === Q2-(d): find CI and PI for age=c(8, 17,21)
# newd=list(age=c(8,17,21))
```

```r
# exp( predict(mod2,newdata=newd, interval=c("confidence")) )
range(a2$age)
exp( predict.lm(mod2,newdata=data.frame(age=c(8, 17,21)), interval=c("confidence")) )
exp( predict.lm(mod2,newdata=data.frame(age=c(8, 17,21)), interval=c("prediction")) )

## Q3: log(FEV)~log(age)

#====> Q3-a: find the estimated model
mod3 = lm(log(fev)~log(age),data=a2)
summary(mod3)

#====> Q3-b: find 96% CI for intercept and slope
confint(mod3)

#====> Q3-c: interpreation of the slope
exp(mod3$coef[2]*log(2))

#====> Q3-d: compare models in Q2-a and Q3-a, choose a better model
summary(mod2)
summary(mod3)
par(mfrow=c(2,2)); plot(mod2)
par(mfrow=c(2,2)); plot(mod3)


# model assumption checking
par(mfrow=c(2,2))
plot(mod2,which=1:2,main="Log-Level model")
plot(mod3,which=1:2,main="Log-log model")

# model selection
SSE2a = sum( (a2$fev-exp(mod2$fitted))^2 )
SSE3a = sum( (a2$fev-exp(mod3$fitted))^2 )
c(SSE2a,SSE3a)
c(AIC(mod2),AIC(mod3)) # the smaller, the better
```