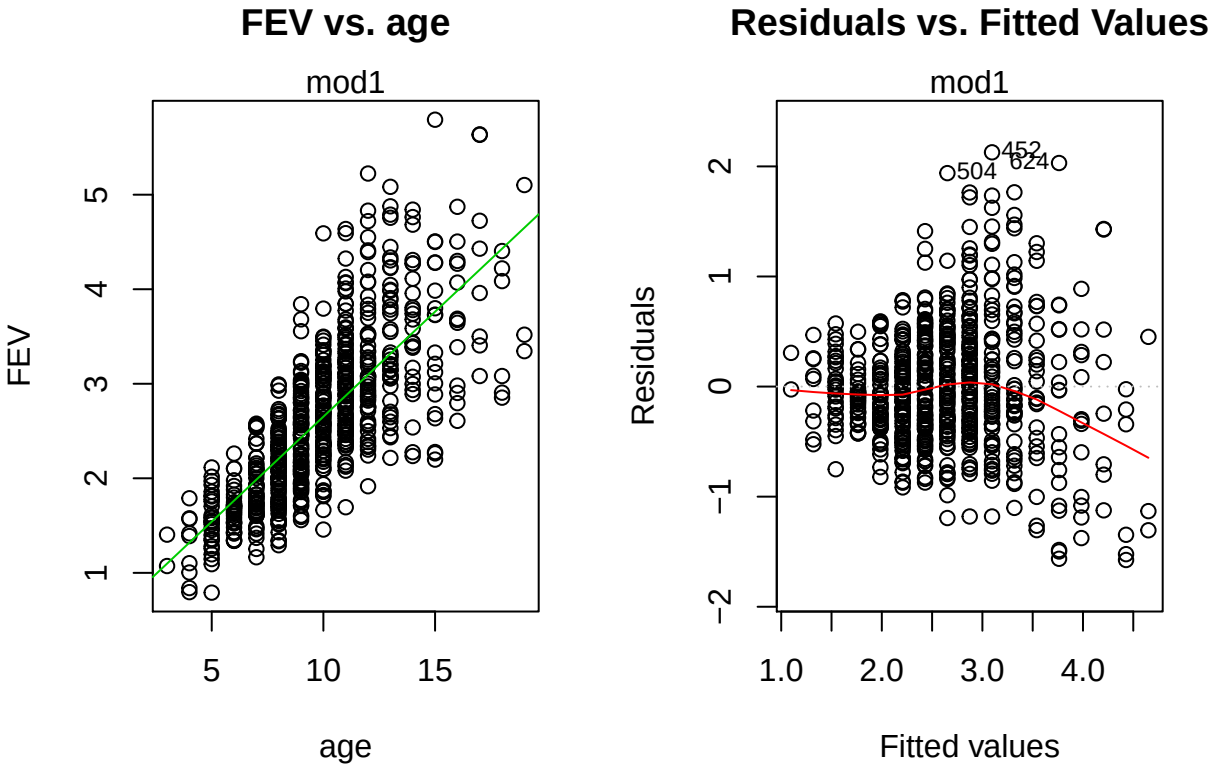
A2: Analysis to Forced Expiratory Volume data

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA302H1F-L0101

Nov. 7, 2016

Q1: Fit a linear model to predict FEV from age.

(a): Scatter plot of FEV versus age.



- Concise comments:

By observing the scatter plot and the residual plot, I can conclude that:

- **Variance is not constant.**

The general rule is that if the observations we have are of equal variance, values should be spread out evenly across x or across the regression line (for scatter plot). However, by first observing the scatter plot we notice that there is clearly slightly more variation under the regression line as it heads towards the end of the line, suggesting that the assumption of equal error variances is ureasonable. Heading to the residual plots it's clear to see that there is some sort of pattern, slightly in the beginning but much more clear to the end as our residual plot has a slight fan shaped pattern outwards, the residuals do not form a "horizontal band" around the 0. This suggests that the variances of the error terms are not equal.

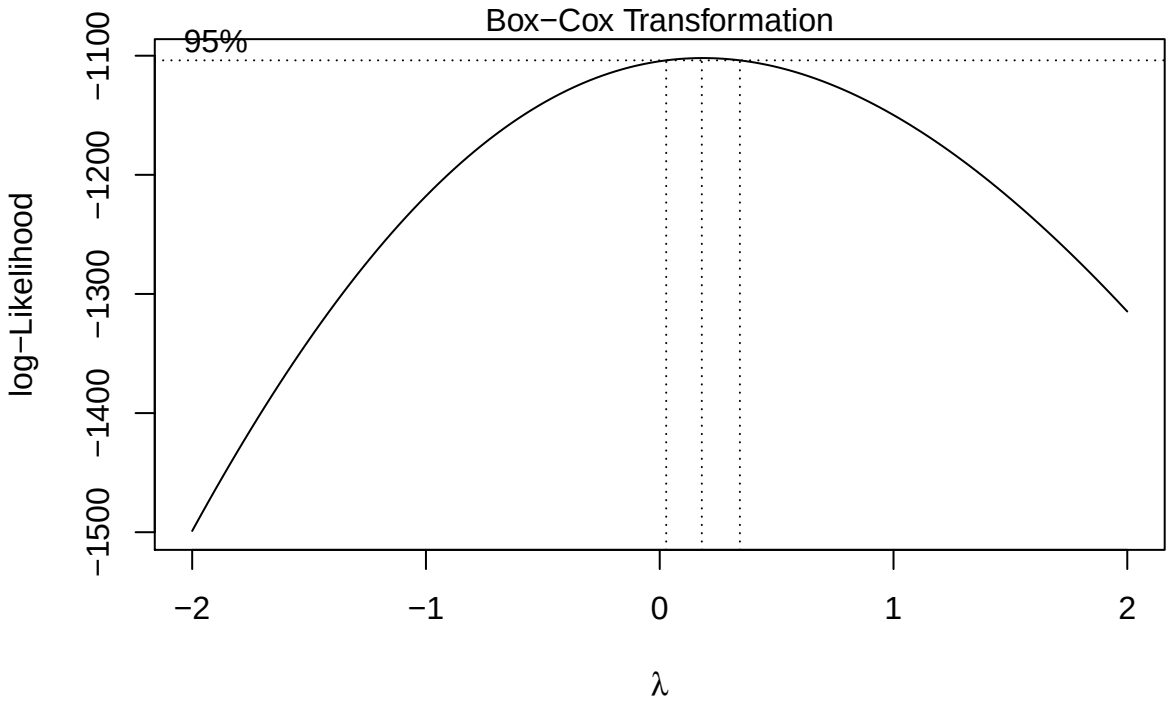
- **Non-linear realtionship.**

The general rule is that in order to conclude that there is in fact a linear relationship there should be no clear pattern in our residual plot. However, observing the residual plot it's clear to see that there is indeed a pattern, slightly in the beginning and more clear in towards the end as the plot forms a fan shaped pattern outwards. We also notice this in the scatter plot as there seems to be a curve downwards which is more clear to see towards the end of the regression line as more points fall under it. This suggests that there is a non-linear relationship for FEV against age.

- **Model we have is unacceptable.**

From these conclusions we just found of non-linearity and unequal variance, the assumptions of the linear regression model are not satisfied which means that the given fit model we currently have is **not** acceptable.

(b): Use boxcox() to find a simple power transformation.



- From this plot, which simple transformation seems best?

Solution:

For a simple power transformation, I believe that the log of the response variable (FEV) appears to be the most appropriate choice. If we were to analyze the box-cox transformation diagram, we notice that the 95% confidence interval for λ is at the maximum when it's at the value 0.18, which is closer to 0 than $\frac{1}{2}$ which is my reasoning in choosing the log transformation.

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

(a): Estimated regression model:

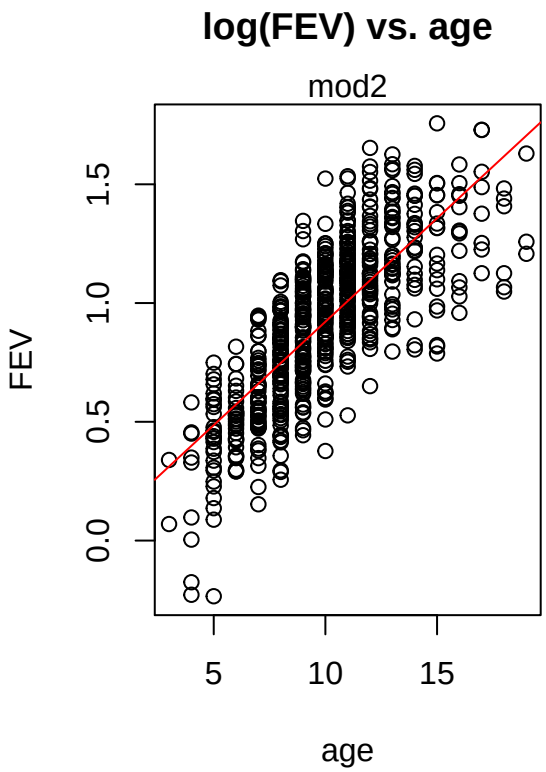
$$\hat{l}og(FEV) = 0.050596 + 0.087083\text{age}$$

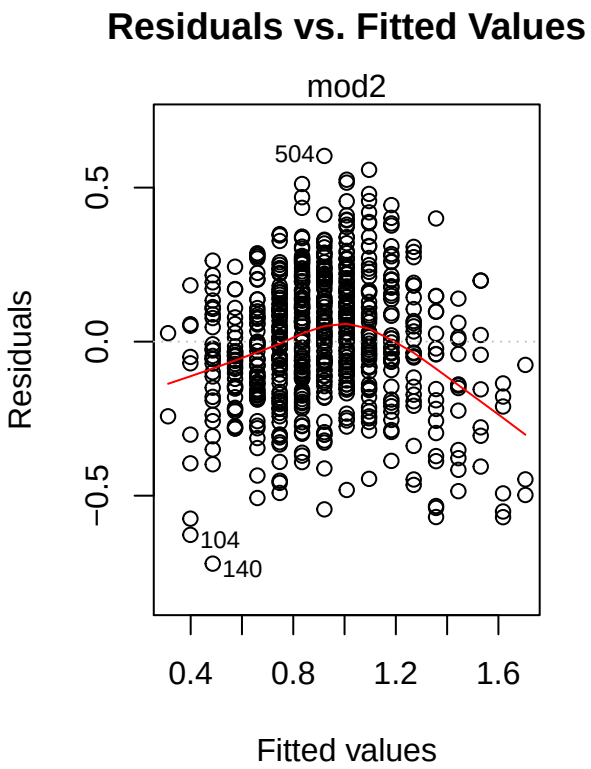
(b): Give comments on the plot.

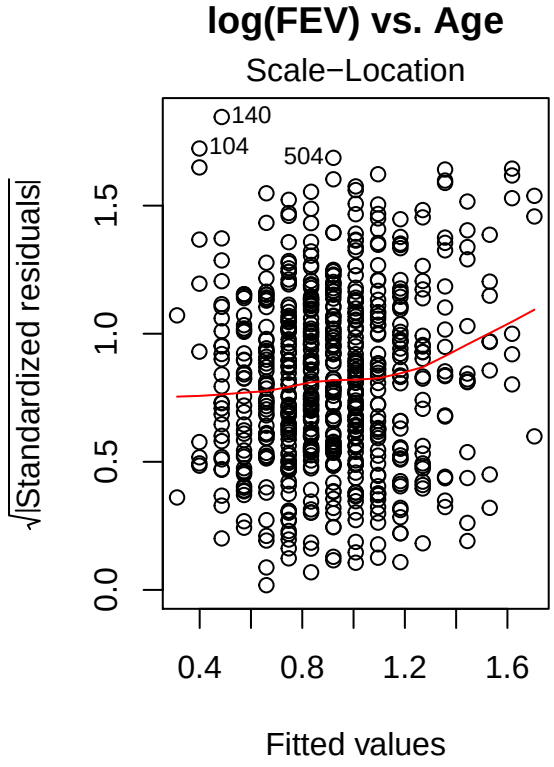
Solution:

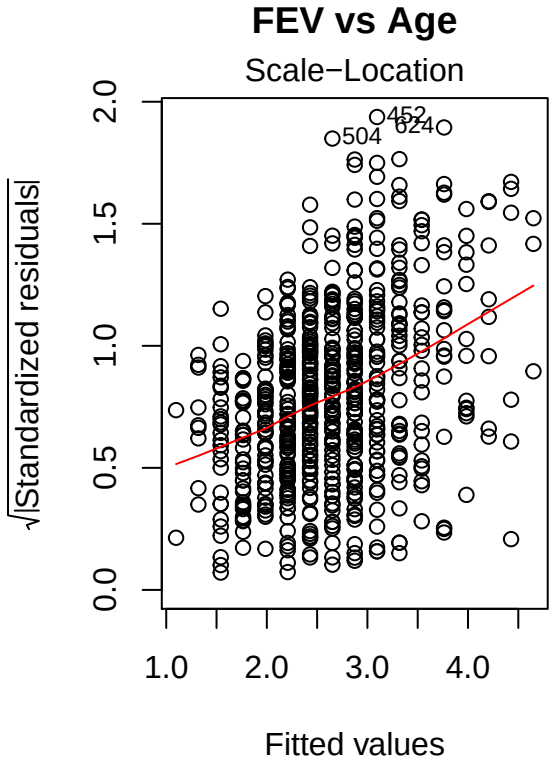
We can observe that the constant variance assumption is better satisfied under the log transformation of the response variable (FEV). If we are to compare the original scale-location model vs. transformed scale location model to compare variance, we notice that there is more of a horizontal line with equally randomly spread points as opposed to the original scale-location model.
However, if we further observe the scatter diagram and the residual plot, we notice that there seems to be a slight curvature formed after this transformation, which was not initially there. This brings another issue with linearity.
We can conclude that this model can be acceptable especially if compared with the previous but it may not be flawless or optimal in general.

PS: Please see the graphs on next page, couldn't all fit on one page


log(FEV) vs. age
mod2
FEV
0.0
0.5
1.0
1.5
5
10
15
age


Residuals vs. Fitted Values
mod2
Residuals
−0.5
0.0
0.5
504
104
140
0.4
0.8
1.2
1.6
Fitted values


log(FEV) vs. Age
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
140
104
504
0.4
0.8
1.2
1.6
Fitted values


FEV vs Age
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
452
624
504
1.0
2.0
3.0
4.0
Fitted values

(c): Assume this model is acceptable, how do you interpret the slope?

Solution:

Given the assumption that this model is acceptable, we note that the current model can be classified as a Log-Level model with the following estimated model:

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

As age were to increase by 1 unit, the mean of FEV should be changing by the multiplicative factor of 1.091 ($e^{0.087083}$)

We also note that $\beta_1 = 0.087083$ which is greater than 0. This means that, as age were to increase by 1 unit, the mean of FEV increases by 9.099%

It is also estimated that, on average, FEV increases by 9.099% with each one unit increase in the age variable.

(d): Find 95% confindence intervals for mean response and 95% prediction intervals for FEV when age is 8, 17 and 21.

Solution:

- 95% confidence intervals:

Age (8): (2.070532, 2.152692)
Age (17): (4.431587, 4.822374)
Age (21): (6.148179, 6.976410)

- 95% prediction intervals:

Age (8): (1.391573, 3.203006)
Age (17): (3.041955, 7.025340)
Age (21): (4.298236, 9.979029)

Q3: Use the simple transformation in Q1(b) on the response variable (FEV), but use log(age) as the predictor variable.

(a): Write down the estimated regression model.

Estimated regression model:

$$\hat{l}og(FEV) = -0.98772 + 0.84615log(\text{age})$$

(b): Find 95% confindence intervals for each model parameter (intercept and slope) in the (possibly) transformed scale.

Solution:

- 95% confidence intervals:

Intercept: (-1.1007528, -0.8746918)
Age: (0.7963774, 0.8959283)

(c): Assume this model is acceptable, how do you interpret the slope?

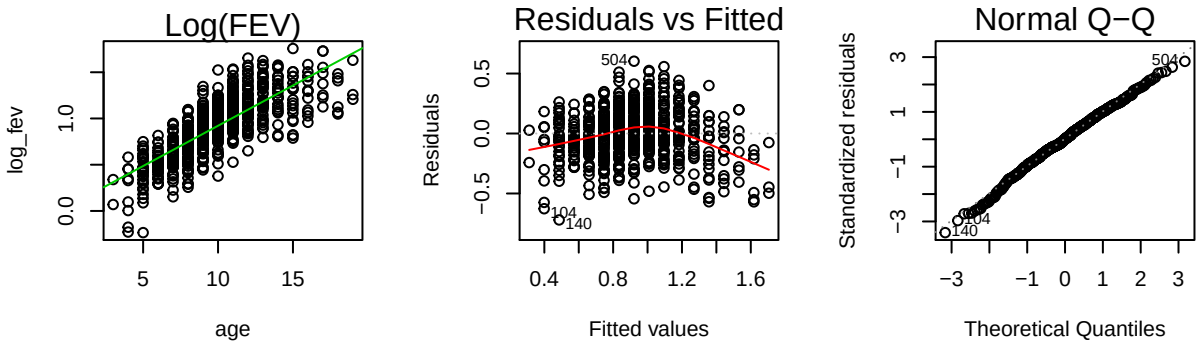
Given the assumption that this model is acceptable, we note that the current model can be classified as a Log-Log model with the following estimate:

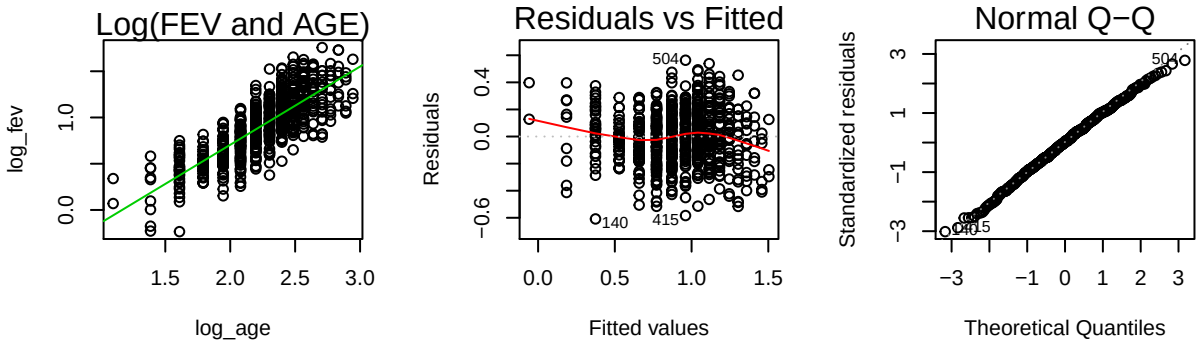
$$\hat{l}og(FEV) = -0.98772 + 0.84615log(\text{age})$$

The interpretation is associated with each doubling of age, the mean of the FEV will change by the multiplicative factor of 1.2901 ($e^{(0.84615)\times log(2)}$)

We also note that $\beta_1 = 0.84615$ which is greater than 0. This means that, as age were to double in unit, the mean of FEV increases by 29.01%

(d): Compare the current model with the model of Q2(a). Which model do you prefer? What criteria do you use to choose a better model between them? Briefly present your result and give a concise explanation.





Solution:

After careful comparison between the log-level model and the log-log model. I believe that the assumptions are better met with the response variable and predictor variable both being log-transformed i.e. (log-log model).

We notice that the log of the response transform (log-level model) also meets the criteria of normality and constant variance based off its Normal Q-Q plot and Residual plot, but one of the main issues with the log-level model is that there seems to be some sort of asymptotic curve as the age variable increases.

Comparing the two R^2 values, the log-log model is at 63.09%, while the log-level model is at 59.58%. By default, the higher the R^2, the better the model fits our data since the more variance that would be accounted for by the regression model, the closer the data points will fall to the fitted regression line. This makes the log-log model the better fit if we use R^2 as our criteria.

If we compare the SSE, log-level model = 241.2044. Log-log model = 210.0379.
The smaller SSE, the more reliable the predictions obtained from the model. In our situation, we note that the log-log model shows to have lowere SSE, hence the better choice in this criteria.

With these reasonings, I believe that the final model (log-log model) should be chosen as the best since it seems to satisfy the model assumptions and also portrays linearity and constant variance across most of the interval as well as have a lower SSE and a higher R^2.

Q4: Source R code

```
# R code for STA302 or STA1001H1F assignment 2
# copyright by Sunil M Khambaita
# date: November 7th, 2016
#

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


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

# Fitting a linear model mod1 which predicts FEV from age.
fev = a2data$fev
age = a2data$age
mod1 = lm(fev ~ age)

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

# Combining 2 figures arranged in 1 row and 2 columns.
par(mfrow=c(1,2))

# Plotting the scatter plot of FEV vs. Age
plot(age, fev, main = "FEV vs. age", xlab = "age", ylab = "FEV", type = "p", col = 9, pch = 1)

# Plotting a regression line to aid in analysis
abline(mod1, col = 3)

# Plotting residuals vs fitted values
plot(mod1, which = 1, caption = list(""), main = "Residuals vs. Fitted Values", sub.caption = "")

##==> Q1(b): boxcox transformation

# Performing a box-cox transformation of mod1
bx = boxcox(mod1, seq(-2, 2, 0.01))

# Creating a variable lambdahat which finds the maximum y value at the corresponding x position
lambdahat = bx$x[which.max(bx$y)]

# Labelling the box-cox transformation graph
mtext("Box-Cox Transformation")

# Returning the value of lambda
lambdahat


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

# Fitting a linear model mod2 which predicts log(FEV) from age.
log_fev = log(a2data$fev)
age = a2data$age
mod2 = lm(log_fev ~ age)
```

```r
##==> Q2(a): estimated model

# Finding a summary of the fitted linear model (mod2) in order to determine the
# estimated regression model (mod2).
summary(mod2)

##==> Q2(b): transformation analysis

# Combining 2 figures arranged in 1 row and 2 columns.
par(mfrow=c(1,2))

## Plotting the scatter plot of log(FEV) vs. Age
plot(age, log_fev, main = "log(FEV) vs. age", xlab = "age", ylab = "FEV", type = "p", col = 9, pch = 1)

## Plotting a regression line to aid in analysis
abline(mod2, col = 2)

## Plotting log (residuals vs fitted values)
plot(mod2, which = 1, caption = list(""), main = "Residuals vs. Fitted Values", sub.caption = "")

## Plotting scale-location plot for mod1
plot(mod1, which = 3)

## Plotting scale-location plot for mod2
plot(mod2, which = 3)

## ==> Q2(d) 95% CI and PI for age = 8, 17, 21.

# 95% confidence intervals for mean response in untransformed scale FEV when age=c(8, 17,21)
exp(predict(mod2, newdata = data.frame(age = c(8, 17, 21)), interval = "confidence", level = 0.95))

# 95% prediction intervals in untransformed scale for FEV when age=c(8, 17,21)
exp(predict(mod2, newdata = data.frame(age = c(8, 17, 21)), interval = "prediction", level = 0.95))


## Q3: use the simple transformation in Q1(b) on the response variable (FEV),
## but use log(age) as the predictor variable.

# Fitting a linear model mod3 which predicts log(FEV) from log(age).
log_fev = log(a2data$fev)
log_age = log(a2data$age)
mod3 = lm(log_fev ~ log_age)

##==> Q3(a): estimated model

# Finding a summary of the fitted linear model (mod3) in order to determine the
# estimated regression model (mod3).
summary(mod3)

## ==> Q3(b) 95% CI for each model parameter.
confint(mod3)

## ==> Q3(d)

# Finding the R squared values for mod2
summary(mod2)
```

```r
# Finding the R squared values for mod3
summary(mod3)

# SSE for mod2
mod2sse= sum((fev-exp(mod2$fitted))^2)
mod2sse

# SSE for mod3
mod3sse= sum( (fev-exp(mod3$fitted))^2)
mod3sse
```