STA302/1001HF: Methods of Data Analysis I **Fall 2016**

Assignment 2 : Forced Expiratory Volume

Out: Oct. 26,2016 *Due: Nov. 17, 2016*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 11 :00pm, Nov. 17, 2016. Submit your solution as instructed by Crowdmark, namely, one pdf file for each question.

Most problems on this assignment require using R. Your turned in solutions should not include all of the R output and graphs that you will produce. Write your solutions and include only sparingly R output or graphs when necessary to support a point you are making in response to the problem question. If a problem asks for a graph, provide it. If the problem asks for you to comment about a graph, you do not need to include the graph in your solution.

Late assignments will be subject to a deduction of 4% of the total marks for the assignment for each day late. Any late assignment after the day I post the solution will get zero mark.

Presentation of solutions is very important. A Rmarkdown template for the solution is provided. I highly recommend you use it. But if it is too much for you, you could also write your solution in a Word document, covert it to PDF and then upload it on Crowdmark. Please make sure the source R code at the end is complete. Marks will be deducted if the instructions herein are not followed.

Data

This data set consists of 654 observations on children aged 3 to 19. Forced Expiratory Volume (FEV), which is a measure of lung capacity, is the variable in interest. Age and height are two continuous predictors. Sex and smoke are two categorical predictors.

The variables in the dataset are :

- age : age of the 654 children (years).
- FEV : forced expiratory volume (liters), a measure of lung capacity.
- height : height (inches).
- sex : female is 0. Male is 1.
- smoke : nonsmoker is 0. Smoker is 1.

More detail about this data can be found in the following two articles :

1. Rosner, B. (1999) Fundamentals of Biostatistics, 5th Ed., Pacific Grove, CA : Duxbur.
2. Michael J. Kahn (2005) An Exhalent Problem for Teaching Statistics Journal of Statistics Education Volume 13, Number 2.

In this assignment, we will consider models that use age (age of children), and fev (forced expiratory volume, in liters).

Questions

1. (10 points) Fit a linear model to predict FEV from age.
   (a) (5 points) Produce the FEV against age scatter plot and the residual plot versus fitted values. Give concise comments.
   (b) (5 points) Use the R function **boxcox()** to find a simple power transformation (-1 = reciprocal, 0 = log, 0.5 = square root) close to the Box-Cox maximum likelihood estimate. Show the plot produced from **boxcox()**. From this plot, which simple transformation seems best ?

2. (11 points) Fit a linear model to predict the best simple transformed response from **age** and examine the residual plot of the fit.
   (a) (2 points) Write down the estimated regression model.
   (b) (2 points) Has the transformation improved adherence to the constant variance assumption ? Is this linear model acceptable ? Briefly explain why or why not ?
   (c) (3 points) Assume this model is acceptable, how do you interpret the slope ?
   (d) (4 points) Find the 95% confidence intervals for mean response in untransformed scale and 95% prediction intervals for FEV when age=c(8, 17,21) ?

3. (14 points) Use the simple transformation in Q1(b) on the response variable (FEV), but use $\log(age)$ as the predictor variable.
   (a) (3 points) What is the estimated model in terms of transformed data ?
   (b) (4 points) Find 95% confidence intervals for each model parameter (intercept and slope) in the (possibly) transformed scale.
   (c) (3 points) Assume the model you obtain in Q3(a) is acceptable, how do you interpret the slope ?
   (d) (4 points) Compare this model with the model that you have in 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.

4. (5 points) Clear R source code for this assignment. (Write brief and clear comment in the between of your source code to ensure your R code is readable. Refer to the sample comment I provide in Q1.)

Incomplete R code

```
#
# R code for STA302 or STA1001H1F assignment 2
# copyright by YourName
# date: Oct. 26, 2016
#

## Load in the data set
a2 = read.table("???/A1/a2data.text",sep=" ",header=T)
## or
A2 = read.table("DataA2.txt",header=T)

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

mod1 = lm(.. )

## ==> 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
boxcox()


## Q2

## Q3:
```

Listing 1 – Incomplete code for the data analysis

Extra : A brief summary of R on SLR

In this section, I am using a simple data to give you a summary of the R relevant function and code regarding the simple linear regression (SLR) model.

```r
# make up data
x=c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
y=c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)
```

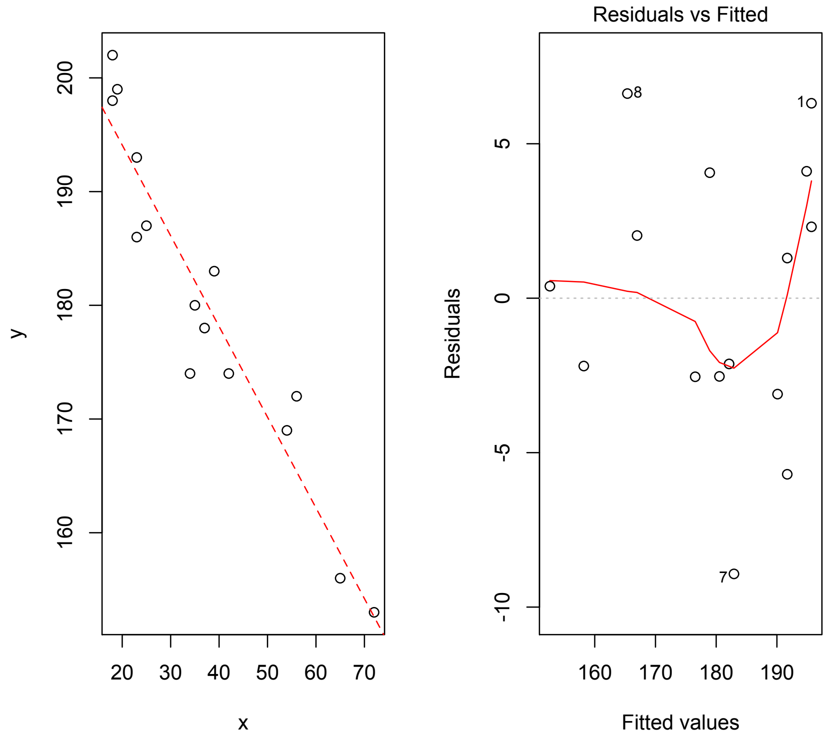
Listing 2 – R example

A scatter plot of the data and the regression line are shown in the next graph. The graph is obtained using the commands :

```r
#make a scatter plot of the data
plot(x,y)

#adding the regression line (with red colour and line type is 2) in the scatter plot
m = lm(y~x)
abline(m,col="red",lty=2)

# want to get the residual plot vs fitted value
plot(m,which=1) # which=2: for the Normal QQ-plot

# Put two plots in one panel
par(mfrow=c(1,2))
plot(x,y)
abline(m,col="red",lty=2)
plot(m,which=1) # which=2: for the Normal QQ-plot
```

Listing 3 – R example

Here is the plot produced from running the last 4 lines



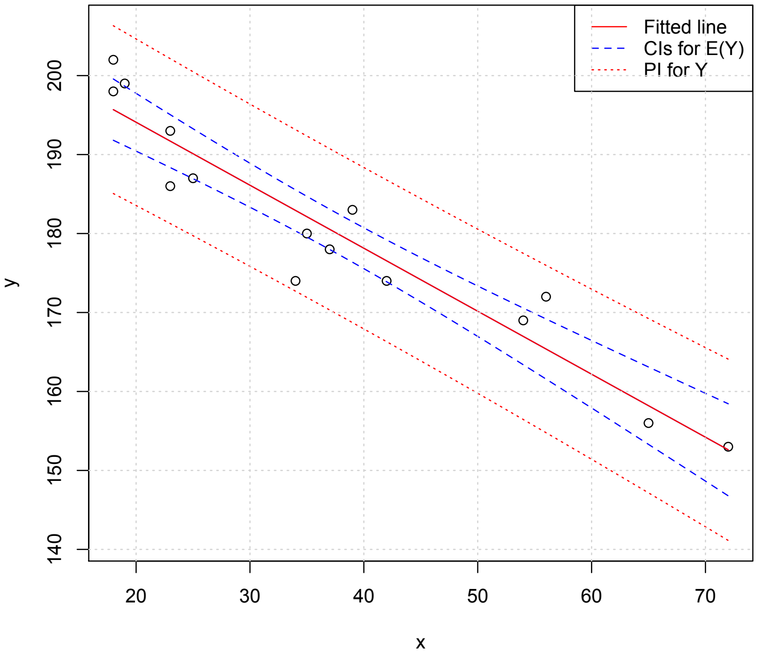
Functions of **lm()**, I summarize here a few functions of interest that you can explore. Each of the following functions acts on **m (m=lm(y∼ x))**, defined by

- **coefficients(m)** - model coefficients.
- **coef(m)** - the same as **coefficients(m)**.
- **confint(m,level=0.99)** - confidence intervals for the regression coefficients.
- **deviance(m)** - residual sum of squares (SSE).
- **fitted(m)**- vector of fitted y values.
- **residuals(m)** -vector of model residuals.
- **resid(m)** - the same as **residuals(m)**.
- **summary(m)** - the summary function already described.
- **vcov(m)** - variance-covariance matrix of the main parameters.
- **df.residual(m)** - the degree of freedom of MSE.
- **plot(m,which=1)** - diagnostic plot.
  which=1 : residuals vs fitted.
  which=2 : Normal QQ-plot.
  which=3 : scale-location.
  which=4 : Cook's distance.
  which=5 : residuals vs leverage.
  which=6 : Cook's distance vs leverage

Plotting the two bands requires some care. Here is a way to do it :

```r
pred.xframe = data.frame(x=18:72)
pi = predict(m,interval="prediction",newdata=pred.xframe)
pci = predict(m,interval="confidence",newdata=pred.xframe)
plot(x,y,ylim=range(y,pi,na.rm=TRUE))
pred.x=pred.xframe$x
matlines(pred.x,pci,lty=c(1,2,2),col="blue")
matlines(pred.x,pi,lty=c(1,3,3),col="red")
legend("topright", c("Fitted line", "CIs for E(Y)", "PI for Y"), lty=c(1,2,3), col=c("red","blue","red"))
grid()
```

Listing 4 – R example


Fitted line
CIs for E(Y)
PI for Y
y
x