STA302/1001 Autumn 2017, Homework #2

28 October. With thanks to Becky Lin

Instructions: These aren't for credit, so don't hand them in. They relate to Chapter 3 of the textbook, and you should be comfortable with them all using the lectures we've already had.

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 of interest. Age and height are two continuous predictors, and there are two categorical predictors: sex and smoking status. In summary, the variables in the dataset are:

- `age`: Age (years)
- `fev`: FEV (litres)
- `ht`: Height (inches)
- `sex`: Sex (female is 0; male is 1)
- `smoke`: Smoking status (nonsmoker is 0; smoker is 1)

In this assignment, we'll only consider models that use `age` and `fev`. Example R code to load the data etc is below.

1. Fit a linear model to predict FEV from age.

(a) Produce the FEV against age scatter plot. Show the scale-location plot, and comment.

(b) Plot square root-, logarathmic-, and reciprocal transformations of the data (transform both variables similarly). From these plots, which transformation seems best?

2. Investigate the transformed response

(a) Write down the estimated regression model for your selection in 1(b).

(b) Has the transformation improved adherence to the constant variance assumption? Is this linear model acceptable? Briefly explain why or why not.

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

(d) For `age=c(8,17,21)`, find the 95% confidence intervals for mean response in the untransformed scale. Find the corresponding 95% prediction intervals. Hint: you may use the way we have done so far, or for faster results you may wish to experiment with the `predict.lm` command in R.

Incomplete R code

```
# R code for STA302 or STA1001 Assignment 2

a2 = read.table("DataA2.txt",sep=" ",header=T) # Load the data set
fev <- a2$fev; age <- a2$age

# Q1: plot FEV vs age and the scale-location plot
mod1 = lm(...)
par(mfrow=c(1,2))
plot(age,fev,type="p",col="blue",pch=21, main="FEV vs age")
plot(mod1,which=3) # 'which' selects from among the 4 lm plots
```

Extra reading

For those interested in doing additional reading (not needed for the homework), details about the published study can be found in the following two articles:

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