550.413 Assignment 2

Fall 2016

Instruction: This assignment consists of 4 problems. The assignment is due on **Friday, October 14, 2016**, in section. If you cannot make it to section, please leave the assignment under the door at Whitehead 306E and email the course instructor. If possible, please type up your assignments, preferably using LaTeX. For problems with R programming, please also attach a print-out of the **R** code. When asked to perform hypothesis testing, you are free to use any (reasonable) choice of significance level α.

Problem 1: (20pts)

This problem uses the dataset `cars` from the `faraway` library. The `cars` data records the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s. Using this dataset, first perform a simple linear regression with `dist` as the response variable and `speed` as the predictor variable. Then using this model, answer the following questions

1. Check the constant variance assumption for the errors.
2. Check the normality assumption for the errors.
3. Check for outliers and nominate potential outliers, if they exist.
4. Without using the `anova` command, check the goodness-of-fit of this simple linear regression model.

Problem 2: (10pts)

Table 1 displays the male death (per millions) due to lung cancer vs per capita cigarette consumption for several countries.

(a) Using the above data, fit a simple linear regression. Do any of the points look particularly influential ? Delete the United States, refit a simple linear regression and note how the regression line changes. Now, put the U.S.A back and delete Great Britain and note how the regression line changes.[1].

[1] If you have better things to do than to manually enter the above data into your **R** session, you can load the dataset as dataset `E8.12` in the package `SenSrivastava`

Country	Y	X
Ireland	58	220
Sweden	115	310
Denmark	165	380
United States	190	1280
Switzerland	250	530
Great Britain	465	1145
Norway	90	250
Canada	150	510
Australia	170	455
Holland	245	460
Finland	350	1115

Table 1: Male death per millions (Y) vs per capita ciragette consumption (X) in 1930 for several selected countries.

(b) What is a plausible explanation for the outliers or influential points from (a) ?

Problem 3: (10pts)

Install the package `aprean3` to get access to data from the book "Applied Regression Analysis" by Draper and Smith. The following problem uses the dataset `dse13e` which records the growth rate and densities of ice crystals. Ice crystals are introduced into a chamber where the interior is maintained at a fixed temperature (-5 degree centigrade) and humidity. The growth of the crystals with time is observed. The variables here are `t` and `m` denoting the times in seconds from the introduction of the crystals and the mass of the crystals in nanograms, respectively.

Using the above dataset, answer the following question

1. Perform a simple linear regression with `m` as the response variable and `t` as a predictor variable. Discuss whether or not this simple linear regression model is appropriate.

2. Try to see if you can "improve" the simple linear regression model in part (1) by transforming either (or both) the predictor or the response variable.

Problem 4: (20pts)

Install the package `aprean3` to get access to data from the book "Applied Regression Analysis" by Draper and Smith. The following problem uses the dataset `dse07c` which records the number of motor vehicle deaths and the number of drivers for 50 states in the US in 1966. We augment the data for the 50 state with the statistics for Washington DC and clean up the variable names as follows.

```r
library("aprean3")
data(dse07c)
data(state)
y <- c(dse07c$y, 115)
x1 <- c(dse07c$x1, 35)
x2 <- c(dse07c$x2, 12524)
x3 <- c(dse07c$x3, NA)
x4 <- c(dse07c$x4, "No")
x5 <- c(dse07c$x5, 44)
x6 <- c(dse07c$x6, 23)
df <- data.frame(deaths = y, drivers = x1, density = x2, rural.mileage = x3, more.male = x4,
    january.temp = x5, fuel.consumption = x6, state.name = c(state.name, "DC"))
```

Here the variable `drivers` is in units of 10^4, the variable `density` is number of persons per square mile, `rural.mileage` is the total length of rural road in the state (the unit is in thousand of miles), `january.temp` is the average high temperature in January in the state, and `fuel.consumption` is the total amount of gallons consumed per year (the unit is in 10^7 gallons).

Given this augmented data, answer the following question.

1. Plot the number of deaths against the number of drivers. Argue why a simple linear regression with `log(deaths)` as the response variable against `log(drivers)` as the predictor variable is to be preferred over a simple linear regression with `deaths` as the response variable and `drivers` as the predictor variable.

2. Perform a simple linear regression with `log(deaths)` as the response variable against `log(drivers)` as the predictor variable. Plot the residuals against the fitted value for this regression model. Which of the residuals are potential outliers ?

3. It is claimed that by adding another predictor variable, the new model will no longer exhibit any potential outliers, thereby yielding a model with better fit. Looking at the list of remaining variables, which would be the most logical candidate to be added to the current model ? You are not required to fit a new regression model.