Coursework 1

*** IMPORTANT ***

Please submit ONE .doc/.html/.pdf file to the dropbox 1 on moodle. Please include any R code, plots or results obtained by running the R code to your solution file, if required.

1. Basis Statistics (1%)

a. Suppose you ask a group of 10 students at Birkbeck College how many brothers and sisters they have. The number obtained are as follows:

2 3 0 5 2 1 1 0 3 3

Find the following measures of central tendency:

(i) the mean,
(ii) the median and
(iii) the mode.

Find the following measures of spread:

(iv) the variance and
(v) the standard deviation

(b) Suppose these 10 students have the following age:

23 25 18 45 30 21 22 19 29 35

(i) Find the covariance and correlation between the number of siblings and their age.
(ii) Is there a positive or negative or no correlation between the two?
(iii) Is there causation between the two? Justify your answers.

2. Getting familiar with R (2%) [Textbook 2.10]

This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

```
> library (MASS)
```

Now the data set is contained in the object Boston. Use the following command to read about the data set:

```
> Boston
```

You could get more info by the following command:

```
> ?Boston
```

How many rows are in this data set? How many columns? What do the rows and columns represent?

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

(e) How many of the suburbs in this data set bound the Charles river?

(f) What is the median pupil-teacher ratio among the towns in this data set?

(g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

3. Linear Regression (2%) [Textbook 3.13 (a-f)]

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

(a) Using the `rnorm()` function, create a vector, x , containing 100 observations drawn from a N (0, 1) distribution, i.e., a normal distribution with mean 0 and variance 1. This represents a feature, X.

(b) Using the `rnorm()` function, create a vector, *eps*, containing 100 observations drawn from a N (0, 0. 25) distribution i.e. a normal distribution with mean 0 and variance 0. 25.

(c) Using x and *eps*, generate a vector y according to the model

$$Y = -1 + 0.5X + \varepsilon.$$

What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.