MATH 423/533 - ASSIGNMENT 1

To be handed in not later than 11:59pm, 9th October 2018.
Please submit your solutions with relevant R code included as a pdf file via myCourses

---

QUESTION 1

Data stored in three data files on the course website contain x and y variables that are to be used for simple linear regression. The data files are `a1-1.txt`, `a1-2.txt` and `a1-3.txt`.

Code for Question 1

```
#Read in data set 1
file1<-"http://www.math.mcgill.ca/yyang/regression/data/a1-1.txt"
data1<-read.table(file1,header=TRUE)
plot(data1$x,data1$y,pch=18)
x1<-data1$x
y<-data1$y
```

(a) Perform a least squares fit of a simple linear regression model (including the intercept) in `R` for each of the three data sets. In particular, for each data set

   (i) report the parameter estimates arising from a least squares fit;

   (ii) produce a plot of the data with the line of best fit superimposed;

   (iii) plot (against the x values) the residuals $e_i, i = 1, \ldots, n$, from the fit;

   (iv) comment on the adequacy of the straight line model, based on the residuals plot – that is, comment on whether the assumptions of least squares fitting and how they relate to the residual errors ϵ_i are met by the observed data.

   Note: the `R` functions `lm`, `coef` and `residuals` will be useful.

(b) Demonstrate what happens to the least squares estimates if the predictor is

   (i) Theoretically, compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the location shift data and the rescaled data,

      (1) Location shift: $x_{i1} \longrightarrow x_{i1} - m$ for some m;

      (2) Rescaled: $x_{i1} \longrightarrow lx_{i1}$ for some $l > 0$; Compute $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the rescaled data, and compare them with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the original data.

   and compare them with $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for the original data. Also in `R` use the dataset `a1-1.txt` and choose the values for m and l, repeat the computation of the parameter estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ arising from the new least squares fits and numerically verify the above theoretical results.

   (ii) Theoretically compute $\mathbb{E}[\widehat{\beta}_0|\mathbf{X}]$, $\mathbb{E}[\widehat{\beta}_1|\mathbf{X}]$ and $\text{Var}[\widehat{\beta}_0|\mathbf{X}]$, $\text{Var}[\widehat{\beta}_1|\mathbf{X}]$ for the location shift data and the rescaled data respectively. Describe also how the properties of these corresponding estimators change, compared with the original data case.
   10 Mark

QUESTION 2

Let X and Y be random variables. Show that

$$\text{Cov}(a + bX, c + dY) = bd\text{Cov}(X, Y).$$

QUESTION 3

Let

$$Y = 5X + \epsilon$$

where $\epsilon \sim N(0,1)$ and $X \sim \text{Unif}(-1,1)$. Assume that X and $epsilon$ are independent.

(a) Find the mean and variance of Y.
(b) Find $\mathbb{E}[Y^2]$.
(c) Find $\mathbb{E}[Y|X=x]$

QUESTION 4

Suppose that

$$Y_i = \beta_1 X_i + \epsilon_i, i = 1, \ldots, n$$

where $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$. In this model, there is no intercept.

(a) Find the least squares estimate $\widehat{\beta}_1$ of Y.
(b) Show that $\widehat{\beta}_1$ is unbiased.
(c) Suppose that you use the estimator from (a) but, in fact, the true model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n$$

Show that the estimator from part (a) is biased and find an expression for the bias.

QUESTION 5

Suppose that the data are $(x_1, y_1), \ldots, (x_n, y_n)$. If we fit least squares to get $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Let $\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. Prove that

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{y_i} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

QUESTION 6

Simulation problem.

(a) First generate $n = 100$ data points as follows. Take $X_i \sim \text{Uniform}(-1,1)$. Then set

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n$$

where $\beta_0 = 5$ and $\beta_1 = 3$ and $\epsilon_i \sim N(0,1)$. Plot the data and fit the regression line. Add the fitted line to the plot.

(b) Repeat the experiment in part (a) 1,000 times. Each time you will get a different value of $\widehat{\beta}_1$. Denote then by $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. Compute the sample mean of these values, and compare it with the value $\beta_1 = 3$. Plot a histogram of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$.

(c) Repeat (b) but now take ϵ_i to have a Cauchy-distribution. How does the histogram change?

(d) Now we will investigate what happens when the X_i's are measured with error. Generate $n = 100$ data points as follows:

$$X_i \sim \text{Uniform}(-1,1)$$
$$W_i = X_i + \delta_i$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n$$

where $\beta_0 = 5$ and $\beta_1 = 3$, $\epsilon_i \sim N(0,1)$ and $\delta_i \sim N(0,2)$. Suppose we only observe $(Y_1, W_1), \ldots, (Y_n, W_n)$. Plot the data and fit the regression line. Add the fitted line to the plot. Now repeat this 1000 times and find the sample mean of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. Also, plot a histogram of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$. Based on this experiment, discuss that what is the effect of having errors in the X_i's.

EXTRA QUESTION FOR MATH533

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.
(b) What is our prediction with $K = 1$? Why?
(c) What is our prediction with $K = 3$? Why?
(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?