Homework 1

BANA 7052: Applied Linear Regression

Section 1: Wednesday 6:00 – 9:50

Group 5

Tess Newkold

Aabhaas Sethi

Anirudh Chekuri

James Brand

Question 1

```
url <- "https://bgreenwell.github.io/uc-bana7052/data/alumni.csv"
alumni <- read.csv(url)
Y <- alumni$alumni_giving_rate
X <- alumni$percent_of_classes_under_20
```

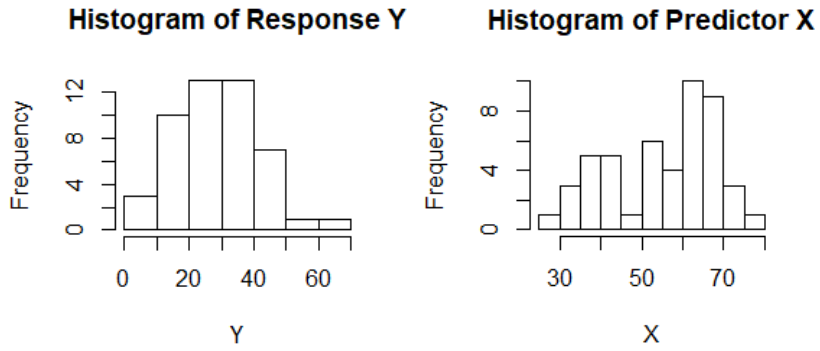
- **Start with a basic exploratory data analysis. Show summary statistics of the response variable and predictor variable.**

```
summary(Y)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   18.75   29.00   29.27   38.50   67.00

summary(X)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   44.75   59.50   55.73   66.25   77.00
```

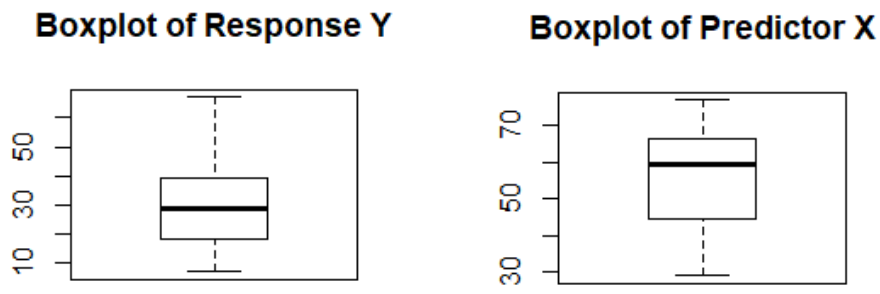
For Y, the median and mean are very close together, so the distribution of Y is probably not very skewed. For X, the median is almost 4% higher than the mean, so the distribution of X may be slightly skewed. We will find out in the next section by exploring their distributions.

- **What is the nature of the variables X and Y? Are there outliers? What is the correlation coefficient? Draw a scatter plot. Any major comments about the data?**

```
hist(Y, main = "Histogram of Response Y")
hist(X, main = "Histogram of Predictor X")
```




```
boxplot(Y, main = "Boxplot of Response Y")
boxplot(X, main = "Boxplot of Predictor X")
```

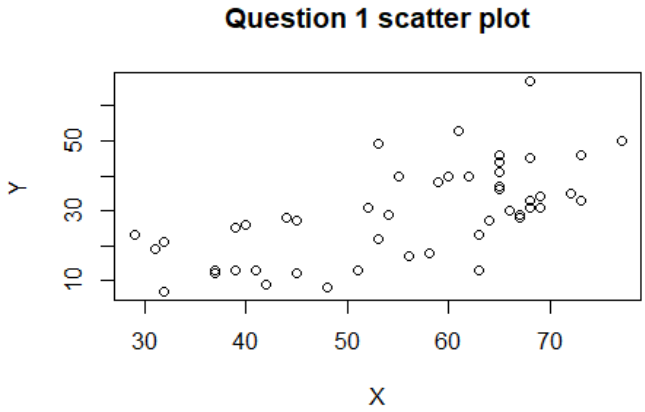


There certainly do not appear to be any outliers in the distribution of Y. Although the distribution of X is left-skewed, none of the values are extreme enough to qualify as outliers.

```
cor(X, Y)

## [1] 0.6456504
```

The correlation between X and Y is about 0.65, which indicates that there may be a moderate linear relationship between X and Y. The positive value of r indicates that increases in the percentage of classes with fewer than 20 students are associated with increases in the alumni giving rate.

```
plot(X,Y, main = "Question 1 scatter plot")
```



The scatterplot further shows that there appears to be some linear association between X and Y.

- **Fit a simple linear regression to the data. What is your estimated regression equation?**

```
fit1 <- lm(Y ~ X)
fit1$coefficients

## (Intercept)           X
##  -7.3860676   0.6577687
```

Our estimated linear regression equation is $\hat{Y} = -7.386 + 0.658X$.

- **Interpret your results**.

The value of $\hat{\beta}_1 = 0.658$ represents the slope of our regression line. We predict that a 1% increase in percentage of classes with fewer than 20 students will result in a 0.658% increase in alumni giving rate. The value of $\hat{\beta}_0 = -7.386$ represents the Y-intercept of our regression line. It would indicate the level of alumni giving expected when 0% of classes have fewer than 20 students, but in this case, it is mostly meaningless since we cannot have a negative value for alumni giving rate.

Question 2

A Simulation Study (Simple Linear Regression). Assuming the mean response is E(Y|X) = 10 +5x:

- Generate data with X ~ N (μ = 2, σ = 0.1) , sample size=100, and error term ε~N(μ = 0, σ = 0.5).

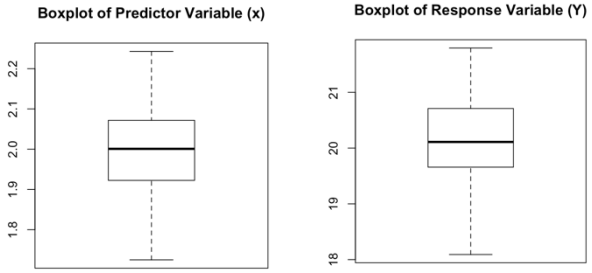
```
1  set.seed(7052)
2  x <- rnorm(100, 2, 0.1)
3  error <- rnorm(100, 0, 0.5)
4  Y <- (10 + (5*x) + error)
```

- Show summary statistics of the response variable and predictor variable.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.725   1.923   2.001   2.004   2.070   2.243
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.09   19.67   20.11   20.17   20.70   21.80
```

a. Are there outliers?
There are no outliers in either variable x or Y.

```
8  boxplot(x, main = "Boxplot of Predictor Variable (x)")
9  boxplot(Y, main = "Boxplot of Response Variable (Y)")
```

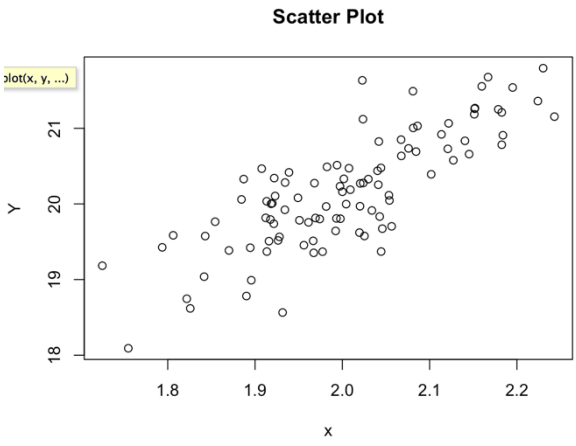


b. What is the correlation coefficient?

```
> cor(Y,x)
[1] 0.8042198
```

c. Draw a scatter plot.

```
12  plot(x, Y, main = "Scatter Plot")
```



- Fit a simple linear regression.

```
15  fit <- lm(Y~x)
16  summary(fit)
17  abline(fit)
```

  a. What is the estimated model?
  Estimated model is $\hat{Y} = 9.0218 + 5.5652x$.

  b. Report the estimated coefficients.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0218     0.8336   10.82   <2e-16 ***
x             5.5652     0.4155   13.39   <2e-16 ***
```

  c. What is the model mean squared error (MSE)?
  MSE = 0.2033

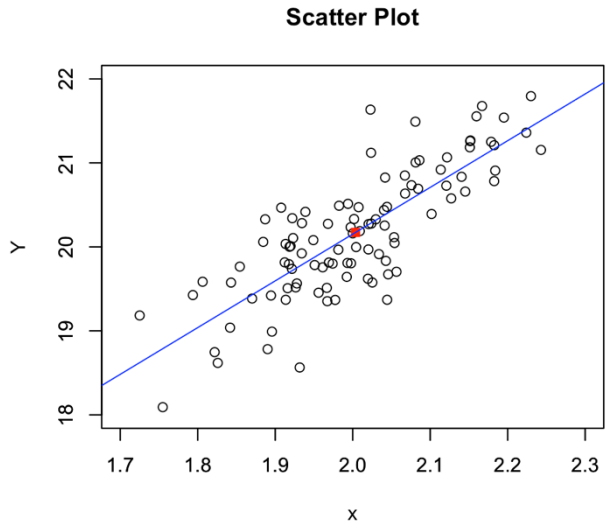
- What is the sample mean of both X and Y?

  $\bar{X} = 2.003677$

  $\bar{Y} = 20.17258$

  a. Plot the fitted regression line and the point. What do you find?

  We find that the sample mean point lies on the fitted regression line.



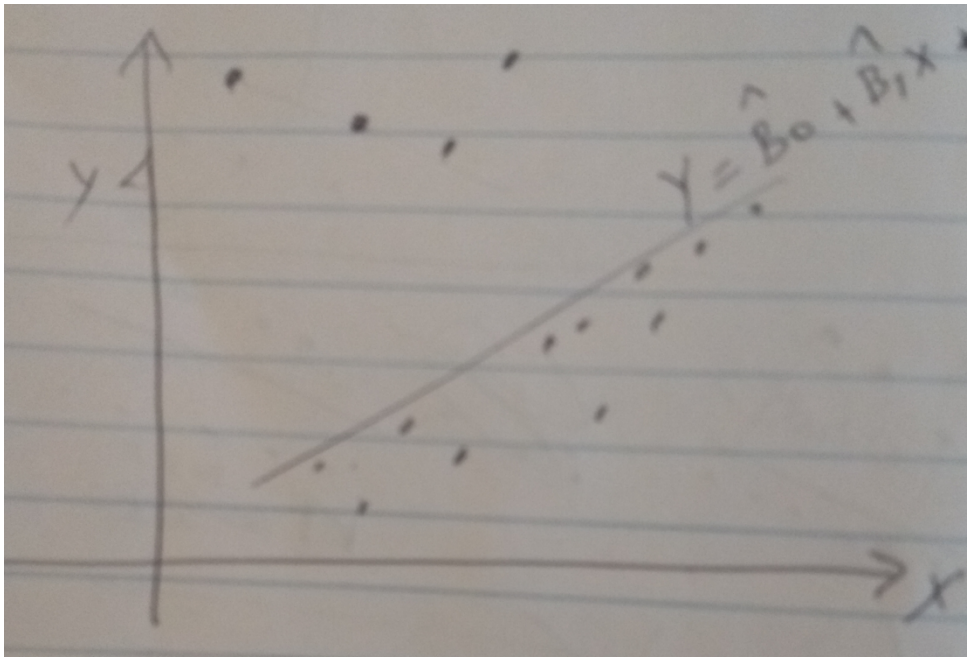
Question 3.

Ordinary least squares (OLS) is typically used to estimate the regression coefficients β0β0 and β1β1 in the simple linear regression model by minimizing the residual sum of squares (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{1}^{n} (Y_i - \beta_0 - \beta_1 * X_i)^2 = \sum_{1}^{n} \epsilon^2$$

a. How about minimizing $\sum_1^n \epsilon$?

Ans.



Cons

I. There can be scenarios such as above where the magnitude of the sum of errors might be 0. However, the regression line is skewed towards a large number of points with less error and away from a few points contributing a lot of error (the magnitude error is large but these points are not outliers).
In real life scenarios, such kind of model will be impractical. For example, seasonal businesses predicting demand will miss their targets in the peak season by a lot of margin, thus resulting in significant opportunity loss.
One of the answers that I read head summarised this point as follows: " Missing by a little lots of times is better than missing by a lot a few times"

II. Mathematically, it would be difficult to find solutions compared with the OLS method since the positive and negative values are cancelling each other out.

Pros

I.For very simple cases with less number of points and not much variation, this method can be used easily. In some cases, just by visual inspection.

b) How about minimizing $\sum_1^n|\epsilon|$?

I. Mathematically, it is hard to differentiate absolute equations. It will take a lot of computational power to arrive at the desired result.

II. There can be cases in which there will be multiple solutions for intercept and coefficients for a given set of data points. In such cases, we would not have any mathematical support to find the best solution. Visual inspection may be required to see which solutions is the best. So, more than 1 cases may be possible.

Pros

I.Just like the previous method, for very simple cases with less number of points and not much variation, this method can be used easily.

c) Why is OLS a popular choice for estimating β_0 and β_1?.

Pros.

I. Mathematically, it is easier to apply calculus on the sum of errors and then differentiate them to get the intercept and coefficient. We are bound to get a single result as only 1 point will correspond to the minima of the parabola (sum of errors)

II. OLS method is more sensitive to variation among points. For the problem mentioned in the first case related to the seasonal demand, OLS method will be better able to reduce the error for the peak demand as it takes into account the square of the error, hence will be pushed more towards the point with larger deviation.

III. The OLS estimates for beta 0 and beta 1 are the same as the maximum likelihood estimates.

Cons.

I. Continuing from point II above, if the data has not been cleaned properly and there are significant outliers that should have been removed, our regression line will be heavily skewed towards those points compared with the other methods. This is because the OLS measure tries to minimize the square of errors and is more sensitive to the points of large deviation.

Question 4

Ans 4) a) The fitted line passes through $(\bar{x}, \bar{y})$.

We know,

$$\hat{Y}_i = \hat{B}_0 + \hat{B}_1 X_i$$

So, if the line passes through $(\bar{x}, \bar{y})$ then when we put value of $\bar{x}$, we should get $\hat{Y}_i = \bar{Y}$

$$\hat{B}_0 = \bar{Y}_i - \hat{B}_1 \bar{X} \quad \text{.... after partial differentiation}$$

$$\Rightarrow \hat{Y}_i = \hat{B}_0 + \hat{B}_1 \bar{x}$$

$$\hat{Y}_i = \hat{Y}_i - \hat{B}_1 \bar{x} + \hat{B}_1 \bar{x}$$

$$\Rightarrow \hat{Y}_i = \hat{Y}_i$$

Hence, the fitted line passes through $(\bar{x}, \bar{y})$

(b) $\sum_{i=1}^{n} e_i = 0$

$$e_i = y_i - \hat{B}_0 - \hat{B}_1 X_i$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{B}_0 - \hat{B}_1 \sum_{i=1}^{n} X_i$$

$$= n\bar{y} - n\hat{B}_0 - \hat{B}_1 n\bar{x}$$

$$= n\bar{y} - n(\bar{y} - \hat{B}_1 \bar{x}) - \hat{B}_1 n\bar{x} = \boxed{0}$$

(c) $\boxed{\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i}$

$$\varepsilon_i = Y_i - \hat{Y}_i$$

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i$$

but from (b) we know $\sum_{i=1}^{n} \varepsilon_i = 0$

$$\Rightarrow \boxed{\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i}$$

(d) $\boxed{\sum_{i=1}^{n} X_i e_i = 0}$

$$RSS = Q = \sum_{i=1}^{n} (Y_i - B_0 - B_1 X_i)^2$$

By differentiating Q wrt B_1 and equating to 0, we get the equation

$$-2 \sum_{i=1}^{n} X_i (Y_i - B_0 - B_1 X_i) = 0$$

substituting $B_0 = \hat{B_0}$ and $B_1 = \hat{B_1}$

$$-2 \sum X_i (Y_i - (\hat{B_0} + \hat{B_1} X_i)) = 0$$

$$\Rightarrow \sum x_i (y_i - \hat{y}_i) = 0$$

$$\Rightarrow \sum e_i x_i = 0$$

e) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$$

$$\sum e_i \hat{y}_i = \sum e_i (\hat{B}_0 + \hat{B}_1 x_i)$$

$$= \sum e_i x_i + \hat{B}_1 \sum e_i$$

$$= 0 + 0 = 0$$

f) $\sum_{i=1}^{n} e_i^2$ is minimized

~~To minimize $\sum_{i=1}^{n} e_i^2$,~~

To check if $\sum_{i=1}^{n} e_i^2$ is minimized, we need to check if the second partial derivatives wrt B_0 and B_1 are positive or not.

$$Q = \sum_{i=1}^{n} (y_i - B_0 - B_i x_i)^2$$

$$\frac{\partial Q}{\partial B_0} = -2 \sum (y_i - B_0 - B_i x_i)$$

$$\frac{\partial^2 Q}{\partial B_0^2} = 2n$$

$$\Rightarrow \frac{\partial^2 Q}{\partial B_0^2} > 0$$

Now $$\frac{\partial Q}{\partial B_1} = -2 \sum x_i (Y_i - B_0 - B_1 x_i)$$

$$\frac{\partial^2 Q}{\partial B_1^2} = -2 \sum x_i (-x_i) = 2 \sum (x_i)^2$$

$$\Rightarrow \frac{\partial^2 Q}{\partial B_1^2} > 0$$

$$\Rightarrow \sum_{i=1}^{n} e_i^2 \text{ is minimized}$$