Bonus Question

Assignment 1 - Question 6

For the Bonus Question, I'll be working on modifying my answer of the above question to incorporate the feedback provided by the marker. My feedback included 5 main points that I've tried to improve in the following solution (I have described how I have improved my solution using the feedback next to each point):

- LaTeX inconsistencies - tried to ensure that there are no LaTeX inconsistencies. Knowledge of LaTex has improved significantly over the course, and I feel much more confident writing LaTeX code. Tried to ensure to stick to the preferred symbols and using the right mathematical notations.
- Writing issues - Tried to improve grammar and better construction of complete sentences.
- Notation issues - Improved upon. For example, for a population notation earlier I was using P but now modified it by using $\mathcal{P}$. Or using u instead of simply 'u' for units.
- Ordered vs. unordered indices - this issue has been resolved
- Describing the example more clearly - tried to go into more detail and use of graphs to provide a better understanding of the example. Provided 2 graphs: Δ vs Index and Δ vs y. Interpretation of graph for a better understanding of the dataset and influence and also stating which units in the population have the most influence in the given data set.

What is Influence?

Consider a population $\mathcal{P}$ and an attribute a such that

$$a(\mathcal{P}) = a(y_1, ..., y_u, ..., y_n)$$

Now the first thing one would wonder is how each unit affects the population? This is exactly what **influence** helps us to understand. We use influence to understand how and which units affect a particular variate. We can know a unit's impact on a particular variate using influence.

How do we calculate influence?

The influence of a unit u, on a variate a, for population $\mathcal{P}$ could be calculated by subtracting that particular unit from the population and calculating the difference between the attribute for the population without the unit, u and with the unit, u.

In other words, influence can be defined as the absolute value of the change in the attribute after removing the element.

$$\Delta(a, u) = a(y_1, ..., y_{u-1}, y_u, y_{u+1}, ..., y_n) - a(y_1, ..., y_{u-1}, y_{u+1}, ..., y_n)$$

We use the Δ sign to denote influence.

Example

In this example we will use a dataset (murders.csv) and plot the influence of the total number of murders in a state on the average of the total numbers.

Average without unit u,

$$a(y_1, ..., y_{u-1}, y_{u+1}, ..., y_n)) = ((N * \bar{y}) - y_u)/N - 1$$

Using the above understanding of influence, the influence for a given u for the variate average is,

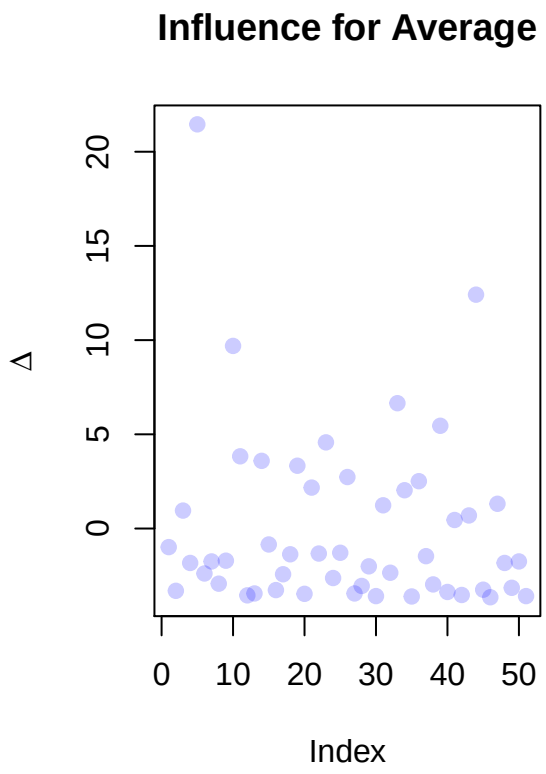
$$\Delta(a, u) = (y_u - \bar{y})/(N - 1)$$

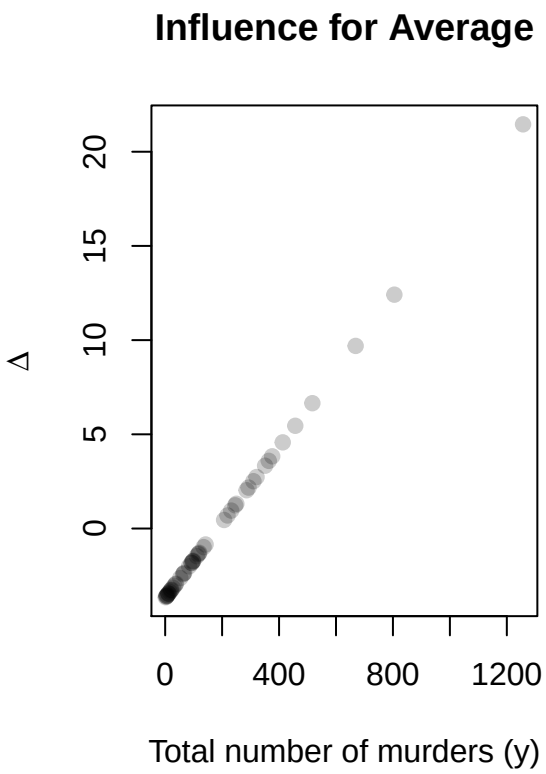
Let us first calculate delta. Here y is the column of murders in each state. We will use y to get delta.

```
data <- read.csv("murders.csv", header = TRUE)
par(mfcol = c(1, 1))
y = data$total
N = length(y)
delta = sum(y)/N-(sum(y)-y)/(N-1)
```

Now, we will plot the influence for every unit u i.e the influence of total murders in each state on the average of total murders by y.

```
par(mfrow = c(1, 2))
plot(delta, main = "Influence for Average", pch = 19, col = adjustcolor("blue",
alpha = 0.2), xlab = "Index", ylab = bquote(Delta))
plot(y, delta, main = "Influence for Average", pch = 19, col = adjustcolor("black",
alpha = 0.2), xlab = "Total number of murders (y)", ylab = bquote(Delta))
```





There is, at least, one (if not a few) states whose number of murders seem to be more influential on the population average compared to other states.

We can also see that the murders in some states have a larger impact/influence on the average number of murders. This could be due to a number of factors such as - gun laws in that particular state, quality of policing, state population etc.

```
# Indexes which have deltas greater than 10
which(delta>10)
```

```
## [1]  5 44
```

Texas and *California* with murder counts *805* and *1257* respectively have the most influence on the average murders. Δ greater than 10 for both of them.

Using influence in a real life scenario

Imagine that you actively invest in stocks. Now you purchase equal number of stocks of different companies but of equal worth, hence you invest the same money in all companies. After some years you want to know which company impacted your earnings the most (in a positive or negative way).

This is where influence helps us. If you were to calculate the influence of stocks of different companies on the average returns receieved over the years you'll be able to understand the stocks of which company benefitted your portfolio the most and which company had no impact on your average earnings.