Stat 243 – Homework 02

Richmond Yevudza

January 27, 2022

1.80 This study is an **observational study**. The three different groups are world-class sprinters, world-class marathon runners and athletes who were examined for the variant gene (ACTN3). There is no treatment given to each of these 3 groups meaning these groups weren't manipulated.

1.84 In this study, the confounding variable is **the amount of snow and ice on the roads** and the variables that are of interest are amount of salt spreads on the roads and the number of accidents. More salt is used to spread on the roads when there is more snow fall and ice on the roads. When the amount of salt increases, the amount of accidents also increase (directly proportional). This makes it obvious that the confounding variable (the amount of snow and ice on the roads) has an association between the two variable of interests (amount of salt spreads on the roads and the number of accidents).

1.88 **Yes**, there is enough evidence that louder music causes people to drink more beer. The reason is because in this study, the sound level of the music was randomly manipulated. An association between the music sound level and the amount of beer consumed by the subjects was noticed.

1.92a The researchers have studied 50 participants who suffer from depression. The two treatment groups are participants who actually get the fluoxetine and participants who get the placebo. 25 participants are then randomly assigned to each group. Therefore, in this case, the explanatory is whether or not the participants get fluoxetine and the response variable measuring level of depression is measured. At the end of the study, results are produced for both groups and compared.

1.92b The placebo pills will be exactly the same as the fluoxetine pills. However, it doesn't contain any active component. This will be given to participants the same way as the fluoxetine creating a placebo in the research.

1.92c The participants won't know whether they are treated with fluoxetine or the placebo. In addition, the people who treat the participants and who carry out the questionnaire won't know which of the groups the participants belong to. This will make the experiment double blind.

1.94a The researchers have studied 50 athletes. The two treatment groups are athletes who consume more carbohydrates and athletes who consume fewer. 25 participants are then randomly assigned to each group. Therefore, in this case, the explanatory is whether or not the consume more carbohydrates or not and the response variable is the athletic performance the next day. At the end of the study, results are produced for both groups and compared.

1.94b For a matched pairs experiment, each athlete wou;ld have a lot of carbohydrates and fewer carbohydrates on different weeks. The order here is randomly determined. The athletic performance would be measured for both cases on the next day. The difference in the performance for each person could then be compared from the two treatment groups to obtain results.

1.94c The design of **matched pairs** is better for this situation. This is because the result would be more accurate when the athletic performance of the same athlete is measured for the two treatment groups than different people's athletic performance for the two treatment groups.

2.132a The **Action** movies appear to have the largest budget because the data approximately lies from 0 to 250.

The **Horror** and **Drama** movies have the smallest budget because the data approximately lies from 0 to 50.

2.132b The **Action** genre has the biggest spread while the **Drama** genre has the smallest spread.

2.132c **Yes**, there appears to be an association between genre of a movie and size of budget. The genre of a movie and size of budget with Action movie is greater as compared to other movies.

2.136a The middle line of each boxplot determines the median. Therefore, looking at the diagram, the estimated median number of hits for American League is about 1455 and the estimated median number of hits for National League is about 1410.

This means that the estimated difference in median hits between the two leagues is 1455 - 1410 = **45 hits**.

Hence, the American League appears to get more hits.

2.136b The American League of the five number summary (minimum, first quartile, median, third quartile, maximum) is greater when compared to the five number summary for National League.

2.144a The **traffic lights are on a fixed or flexible system** is the explanatory variable because it causes an effect on the observed outcomes. This is shown as the **categorical variable** since the variable can't be either measured numerically.

The **delay time** is the response variable because it shows the effect of the observed outcomes. This is also shown as the **quantitative variable** since the variable can be measured numerically.

2.144b Mean (Timed)

```
mean(~Timed, data=TrafficFlow)
```

```
## [1] 105
```

Standard Deviation (Timed)

```
sd(~Timed, data=TrafficFlow)
```

```
## [1] 14.05579
```

Mean (Flexible)

```
mean(~Flexible, data=TrafficFlow)
```

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

Standard Deviation (Flexible)

```
sd(~Flexible, data=TrafficFlow)
```

```
## [1] 3.375101
```

2.144c Mean (Difference)

```
mean(~Difference, data=TrafficFlow)
```

```
## [1] 61
```

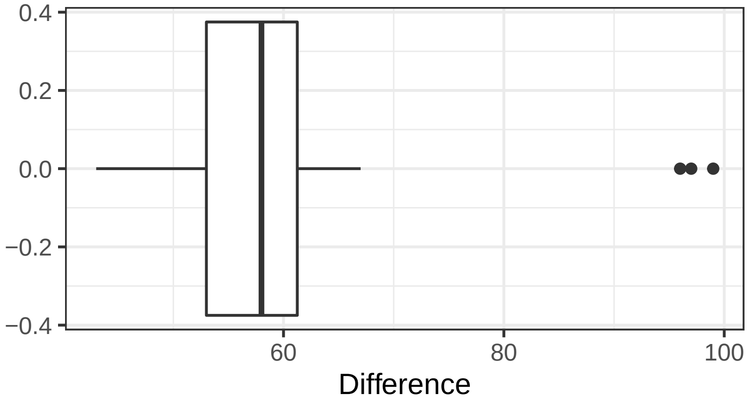
Standard Deviation (Difference)

```
sd(~Difference, data=TrafficFlow)
```

```
## [1] 15.18867
```

2.144d Boxplot of Differences

```
gf_boxplot(~ Difference, data=TrafficFlow)
```



This distribution is skewed left. In addition, there appears to be three outliers at the upper end of the boxplot.

2.158 I expect **negative association** between distance driven and the amount of gas left in the tank. This is because if the amount of gas left in the tank is decreased, then the corresponding variable distance driven is increased.

2.160 I expect **positive association** between number of text messages sent on a cell phone and number of text messages received on the phone. This is because the more you text, the more likely you are to receive messages.

2.168a A **positive association** between Height and Weight means that if height is increased, then the corresponding weight is also increased. Therefore the larger value of height is related to the larger value of weight, and the smaller value of height is related to the smaller value of weight.

A **negative association** between Height and Weight means that if height is increased, then the corresponding weight is decreased. Therefore the larger value of height is related to the smaller value of weight, and the smaller value of height is related to the larger value of weight.

The association i expect to be more likely is the **positive association** between the height and the weight because taller people are usually associated with more weight.

2.168b There appears to be a **positive relationship** between height and weight. This is because, as seen on the scatterplot, it is clear that there is an upward trend between the two variables.

2.168c Looking at the scatterplot, the person represented by the outlier in the lower right corner with height and weight is around 83 inches and 135 pounds. This point is shown as the outlier which shows that taller people are associated with less weight.

2.170a There is a **positive association** between the more amount of maternal nurturing one receives as a child and hippocampus size because the larger the amount of maternal nurturing one receives as a child, the larger the hippocampus size.

2.170b There is a **positive association** between the hippocampus size and the resilliency because the larger the hippocampus size, the larger the resilliency.

2.170c The randomized experiment is designed to test the effect described in part (a) in humans this way:

Some of the children are randomly assigned to take more nurturing while some children are randomly assigned to take less nurturing. The size of the hippocampus in their brains is measured, after long years. This experiment won't be ethical because it isn't fair to assign some of the children to not take nurturing food.

2.170d Because there are more possible surprising variables and also the experiment can't be obtained, the maternal nurturing in humans causes the hippocampus to grow larger can't be concluded.

In addition, it can be said that maternal nurturing in animals causes the hippocampus to grow larger because the animal inference come from the experiments. This shows that the maternal nurturing in animal causes the hippocampus to grow larger; there is a causation result in humans.

2.180a The positive association between these two variables shows that online time increases with the rise in connection speed of the internet.

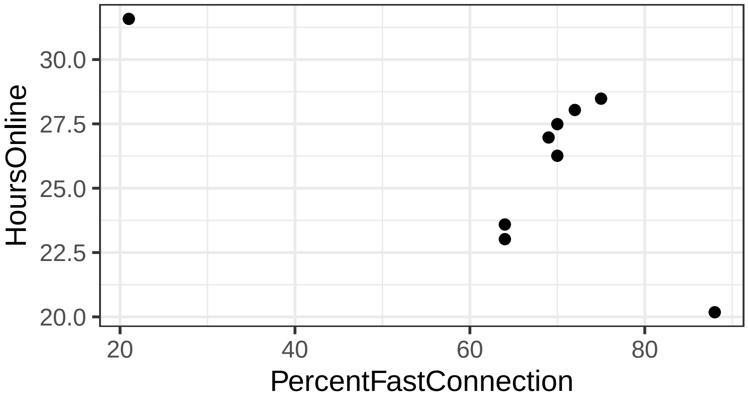
The positive relationship might make sense in this context because people will enjoy more being online with high-speed internet connection than when the internet connection speed is low. Therefore, a conclusion can be made that people spend more time online with high-speed internet connection.

2.180b The negative association between these two variables shows that online time decreases with the rise in connection speed of the internet.

The negative relationship might make sense in this context because people can finish their work with less time online with high-speed internet connection. Therefore, a conclusion can be made that people spend less time online with high-speed internet connection.

2.180c Scatterplot of Hours Online vs Percent Fast Connection

```r
ggplot(data=GlobalInternet, aes(PercentFastConnection,HoursOnline)) +
  geom_point()
```



Looking at the scatterplot, there is a negative linear relationship between the two variables. In addition, there are two outliers in the scatterplot. The first outlier appears at the top left of the graph which corresponds to Brazil and the second outlier appears at the bottom right of the graph which corresponds to Switzerland.

2.180d Eliminating two outliers from the scatterplot will show that the remaining countries have a **positive relationship** between these two variables.

2.180e Correlations: Percent Fast Connection, Hours Online

```
cor(HoursOnline ~ PercentFastConnection, data = GlobalInternet)
```

```
## [1] -0.6492027
```

2.180f **No**, it can't be concluded that a faster connection speed causes people to spend more time online because the data mainly comes from an observational study.

2.196a The explanatory variable is the **duration of the waggle dance** because it causes an effect of the observed outcomes. This is also the categorical variable since the variable can't be measured numerically.

The response variable is the **distance to the source** because it reflects the effect of the observed outcomes. This is also the quantitative variable since the variable can be measured numerically.

2.196b Looking at the scatterplot, there is a clear upward trend between the duration of the waggle dance and distance to the source. This means that there is a **strong positive linear relationship** between the duration of the waggle dance and distance to the source.

2.196c Correlations: Duration, Distance

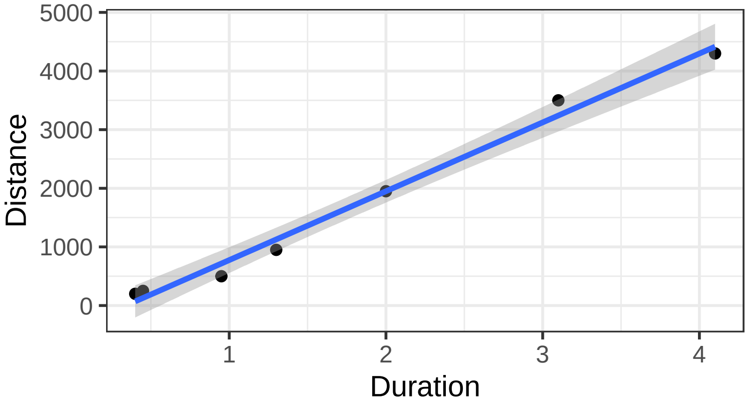
```
cor(Distance ~ Duration, data = HoneybeeWaggle)
```

```
## [1] 0.9941942
```

2.196d Scatterplot of Distance vs Duration

```
ggplot(data=HoneybeeWaggle, aes(Duration,Distance)) +
  geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



2.196e The value of the slope is 1174.26. This means that the estimated value of distance increased by 1174.26 meters for every one meter increase in duration.

2.196f Distance = -399.1 + 1174.26(1) = -399.1 + 1174.26 = 775.16

Therefore, the predicted value for dance lasting 1 second is **775.16 meters**.

2.202a The scatterplot containing data with the largest information is the is **Using abdomen circumference to predict percent body fat** because there is an upward trend with strong positive linear relation between the variables.

2.202b Looking at the scatterplot of "Using abdomen circumference to predict percent body fat", the perso estimated body fat of the person that has a very large abdomen circumference of about 127cm is **34%** whereas the predicted body fat percent for this same person is **40%**.

2.202c The abdomen circumference for this person is around **113cm** and the predicted body fat percent is around **32%**.

$Residual = y - \hat{y}$

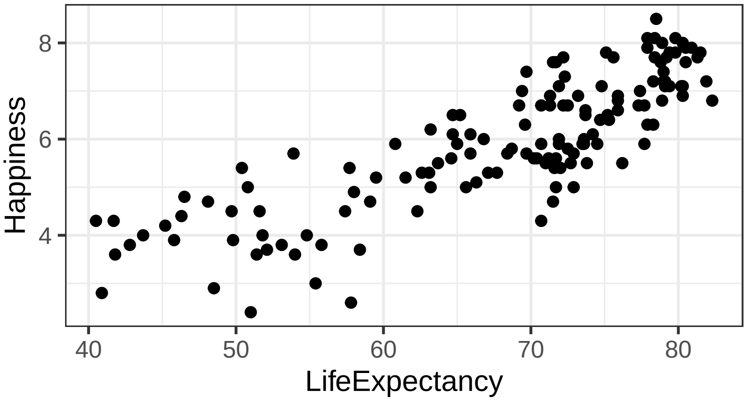
where y = observed value, and $\hat{y}$ = predicted value

Residual = 40 - 32 = 8

Therefore, the residual for the actual body fat percent is **8**, which is greater than the predicted value.

2.208a Scatterplot of Happiness vs LifeExpectancy

```
ggplot(data=HappyPlanetIndex, aes(LifeExpectancy,Happiness)) +
  geom_point()
```

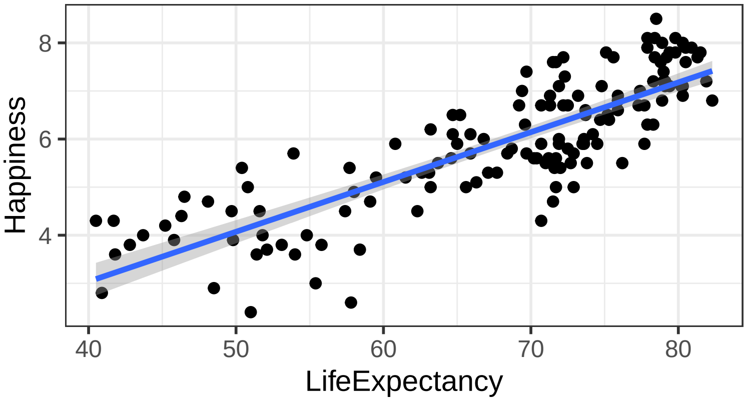


2.208b The formula for regression line is:

Happiness = -1.1 + 0.104 (LifeExpectancy)

```
ggplot(data=HappyPlanetIndex, aes(LifeExpectancy,Happiness)) +
  geom_point() + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



2.208c Looking at the regression line, the value of the slope is 0.104. This means that the estimated value of happiness is increased by 0.104 units for every one unit increase in life expectancy.