# ASSIGNMENT 5

Jahedur Rahman

1/27/2022

## Student Survey Analysis Questions

### i.

Use R to calculate the covariance of the Survey variables and provide an explanation of why you would use this calculation and what the results indicate.

```
student_survey_df <- read.csv("C:/Users/jahed/OneDrive/Documents/GitHub/dsc520/data/student-survey.csv"
cov(student_survey_df)
```

```
##              TimeReading       TimeTV  Happiness      Gender
## TimeReading   3.05454545 -20.36363636 -10.350091 -0.08181818
## TimeTV      -20.36363636 174.09090909 114.377273  0.04545455
## Happiness   -10.35009091 114.37727273 185.451422  1.11663636
## Gender       -0.08181818   0.04545455   1.116636  0.27272727
```

> A covariance matrix helps us see relationships between variables. A high covariance between two variables mean that there could be a correlation between variables. Negative covariance means that as one variable moves toward the mean, the other moves away. Positive coovariance means that both variables are moving the same direction.

### ii.

Examine the Survey data variables. What measurement is being used for the variables? Explain what effect changing the measurement being used for the variables would have on the covariance calculation.Would this be a problem? Explain and provide a better alternative if needed.

```
head(student_survey_df)
```

```
##   TimeReading TimeTV Happiness Gender
## 1           1     90     86.20      1
## 2           2     95     88.70      0
## 3           2     85     70.17      0
## 4           2     80     61.31      1
## 5           3     75     89.52      1
## 6           4     70     60.50      1
```

> The levels of measurement for each variable are, discrete for Time TV, discrete for Time Reading, ordinal for Happiness, and binary for Gender. If the measurement was effected to have larger numbers and a larger range than the covariance will be effected. A better alternative for this would be to use a standard measure for the correlation coefficient. This is because it is based on standard deviation. So if the variables have differing levels of measurement there will be no issue.

**iii.**
Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

> The type of correlation test that I will perform is Kendall’s Tau Correlation test. This would be between Time TV and Happiness. I chose this test because the sample size is 11 students and I am not assuming normality. In addition, the variables Time TV and Happiness were chosen because of their covariance score. Using this the prediction is that the test will yield a positive correlation.

```
nrow(student_survey_df)
```

```
## [1] 11
```

```
cor(student_survey_df$TimeTV, student_survey_df$Happiness, method= "kendall")
```

```
## [1] 0.4630424
```

**iv.**
Perform a correlation analysis of:

1.All variables

```
cor(student_survey_df, method="kendall")
```

```
##             TimeReading      TimeTV   Happiness      Gender
## TimeReading  1.00000000 -0.80454045 -0.28894280 -0.07824608
## TimeTV      -0.80454045  1.00000000  0.46304237 -0.02507849
## Happiness   -0.28894280  0.46304237  1.00000000  0.09847319
## Gender      -0.07824608 -0.02507849  0.09847319  1.00000000
```

2.A single correlation between two a pair of the variables

```
cor(student_survey_df$TimeTV, student_survey_df$Happiness, method="kendall")
```

```
## [1] 0.4630424
```

3.Repeat your correlation test in step 2 but set the confidence interval at 99%

```
cor.test(student_survey_df$TimeTV, student_survey_df$Happiness,
         method = "kendall", conf.level = .99, exact = FALSE )
```

```
##
##  Kendall’s rank correlation tau
##
## data:  student_survey_df$TimeTV and student_survey_df$Happiness
## z = 1.9582, p-value = 0.05021
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.4630424
```

4.Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

> The calculations in the correlation matrix suggests that there is a strong negative correlation between Time TV and Time Reading. It was at -0.80 out of (+/- 1). This suggests that students spending more time watching TV spent less time reading. In addition, the correlation matrix suggests that there is a weak negative correlation between Time Reading and Happines which is at -0.29. There is a very weak negative correlation between Time Reading and Gender. Additionally, there is a positive correlation between Time TV and Happiness which is 0.46. This means that students who watched TV had more Happiness. The values that are 1.000 on the matrix are just the variables correlated with themselves.

**v.**

Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```
cor(student_survey_df$TimeReading, student_survey_df$TimeTV, method="kendall")^2 * 100
```

```
## [1] 64.72853
```

> I calculated the coefficient of determination for Time Reading and Time Tv by squaring the correlation. Then I multiplied by 100 to change the score to a percentage. I conclude that the results mean that Time Reading shares 64.73% of the variability found in Time Tv.

**vi.**

Based on your analysis can you say that watching more TV caused students to read less? Explain.

> Based on my analysis there is a strong correlation between Time TV and Time Reading. However, correlation does not imply causation. The strong correlation indicates that there is a relationship, but the is no evidence that can tell us if other factors effected the findings. To find evidence of causation there must be more experimentation done.

**vii.**

Pick three variables and perform a partial correlation, documenting which variable you are "controlling". Explain how this changes your interpretation and explanation of the results.

```
student_partial<-pcor(c("TimeReading", "TimeTV", "Happiness"), var(student_survey_df))
student_partial
```

```
## [1] -0.872945
```

```
student_partial ^2 * 100
```

```
## [1] 76.2033
```

```
pcor.test(student_partial, 1, 11)
```

```
## $tval
## [1] -5.061434
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.0009753126
```

```
cor(student_survey_df$TimeReading,student_survey_df$TimeTV)
```

```
## [1] -0.8830677
```

The three variables I have picked to perform a partial correlation are Happiness, Time Reading, and Time TV. The variable I am “controlling” is Happiness. The result of the partial correlation suggests that if there is no influence from the variable Happiness than the correlation changes from -0.80 to -0.87. After squaring we see that Time Reading shares 76.2% of variance with Time TV. The p-value is 0.00098 which means that the correlation is not due to chance per sampling. In addition, I did the initial tests using Kendall’s Tau. Since Kendall’s Tau is more rigorous in scoring the correlation compared to Pearson’s Correlation, I did the initial test again with Pearson’s Correlation. I found the score changes to -0.88 which means that there is a very small change when Happiness is controlled.