Assignment 5

Holly Figueroa

4/27/2021

Student Survey Analysis Questions

Question a.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.

```
#load student survey data
student_survey_df <- read.csv("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
```

> Covariance is a measure that is used to see how two random variables vary together in relation to the mean. To have a high covariance between two variables suggests there may be a correlation between variables. When viewing a covariance matrix of each variable in the student survey, we can see evidence of some relationships between variables. Those with negative values indicate that as one variable moves away from the mean, the other moves in the opposite direction. Those with positive values suggest that as one variable moves away from the mean, the other moves in the same direction.

Question a.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
```

Levels of measurement included in the data are as follows:

- Time TV (discrete)
- Time Reading (discrete)
- Happiness (ordinal)
- Gender (binary)

The variance of a variable is relative to it's own scale so the level of measurement is an important consideration when interpreting covariance. If you choose a unit of measurement that involves larger numbers by default and a larger range, your covariance score will reflect that. For example, a covariance score of 5 might be very significant if you are measuring time in hours, but negligible if you measured the same variable in minutes. To get around this, a standard measure for the correlation coefficient would be better. This measure is based on the standard deviation. As such, variables with different levels of measurement can be compared without issue.

Question a.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?

> Here I have decided to perform a Kendall's Tau Correlation test between Time TV and Happiness. I have decided to use this test because I do not want to assume normality and the sample size is only 11 students. I have also chosen the variables Time TV and Happiness based on their covariance score. While I cannot tell if the covariance result is significant because they do not share scale, it suggests some positive relationship may exist. Therefore, I predict that a positive correlation exists between Time TV and Happiness, where as a person reports watching more tv, happiness scores are higher.

```
nrow(student_survey_df)
```

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

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

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

Question a.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

```r
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%

```r
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.

> It suggests that Time Reading and Time TV have a strong negative correlation, or tau score, at -.80 out of (+/- 1). This means that the more time a student reported doing one activity, they reported doing the other activity less. The matrix suggests Time Reading had a weaker, negative relationship to Happiness at -.28. Time Reading had a negligible, negative correlation to gender at -.07. Time TV had a moderate, positive correlation with Happiness at 0.46, where people who reported watching more tv, had higher scores for Happiness. Time Reading had a negligible, negative correlation with gender. Happiness had an negligible, positive correlation with gender. Values on the matrix show as 1.000 as the variables are correlated with themselves, giving the appearance of a perfect correlation. The two most significant results were Time Reading vs. Time TV and Time TV vs. Happiness.

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

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

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

> Here we have the coefficient of determination for Time Reading and Time Tv by squaring the correlation. Once multiplied by 100, we have a score in the form of a percentage. The result means that Time Reading shares 64.72% of the variability found in Time Tv.

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

While the analysis suggests that these variables are significantly correlated, we cannot conclude there is a causal relationship. Correlations indicate a relationship exists, but however significant, there is no evidence to prove other factors are not actually causing the findings. Causation must be evidenced by other methods, such as experimentation, where variables can be controlled.

Question 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.

For a partial correlation I chose to control for the variable Happiness in it's effect on Time Reading to get a more pure measure of correlation between Time Reading and Time Tv. The partial correlation result suggests that, without the influence of Happiness scores, there is a slightly higher correlation. The initial score was -0.80, now up to -0.87.

```
#create variable for partial correlation and display result
student_partial<-pcor(c("TimeReading", "TimeTV", "Happiness"), var(student_survey_df))
student_partial
```

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

When sqaured, we find this means that Time Reading shares 76.02% of variance with Time Tv. The p-value is far under 0.05, at 0.0009, meaning the correlation is extremely unlikely to be due to chance per sampling.

```
#transform result to variance percentage
student_partial ^2 * 100
```

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

```
#get p value
pcor.test(student_partial, 1, 11)
```

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

However, my previous tests were conducted using Kendall's Tau. Kendall's Tau is more rigorous in it's scoring for correlation than Pearson's Correlation so, comparing the partial correlation score to the initial score would be misleading. When I rerun my initial test under the default, Pearson's Correlation, my starting correlation changes to -.88. This means my partial correlation actually resulted in a smaller, however significant, correlation once Happiness was controlled.

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

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