5.1 Assignment - Student Survey

Brandon Sams

1/11/2020

As a data science intern with newly learned knowledge and skills in statistical correlation and R programming, you will analyze the results of a survey recently given to college students. You learn that the research question being investigated is: "Is there a significant relationship between the amount of time spent reading and the time spent watching television?" You are also interested if there are other significant relationships that can be discovered? The survey data is located in this StudentSurvey.csv file.

1. 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.

```
cov(StudentSurvey)
```

```
##              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
```

The covariance matrix is a measure of the strength of the correlation between two or more sets of random variates. In this case, there are just 4 sets of variates, and thus, there are 16 ways covariance values to observe. One may use this calculation to determine which variables are correlate with which, if the correlation is negative or positive, and the size of the correlation. These results indicate that TimeReading has a negative correlation with TimeTV, for example, although the size of the correlation is not necessarily significant at this point.

2. 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(StudentSurvey)
```

```
##   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
```

Based on the size of the values that are given for each of the variables, it appears that:

```
TimeReading is in hours
TimeTV is in minutes
Happiness is a scale from 0-100
Gender is an encoded value. 1=Male and 0=Female or vice versa.
```

Thus, it is worth converting TimeReading from hours to minutes, and seeing what the covariance is for the resulting dataframe.

```
StudentSurvey$TimeReading <- StudentSurvey$TimeReading * 60
cov(StudentSurvey)
```

```
##              TimeReading        TimeTV   Happiness      Gender
## TimeReading 10996.363636 -1.221818e+03 -621.005455 -4.90909091
## TimeTV      -1221.818182  1.740909e+02  114.377273  0.04545455
## Happiness    -621.005455  1.143773e+02  185.451422  1.11663636
## Gender         -4.909091  4.545455e-02    1.116636  0.27272727
```

It makes the resulting covariance for TimeTV and TimeReading be represented in minutes^2. It isnt direcly applicable in this dataset, but if the covariance of these two variables was compared to the covariance of, say, TimeReading and TimeOutdoors, it would make it easier to compare the two covariances.

3. 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?

This data set is relatively small, and a large amount of the data appears repeatedly in the dataset. Therefore, I would prefer to utilize Kendall's Tau as a correlation test. I predict a negative correlation to exist between the two, not specifically because of TV and Reading uniquely, but because as a person spends an increasing amount of time doing one activity, they have less time for other activities. Let's take a look.

```
cor.test(StudentSurvey$TimeReading, StudentSurvey$TimeTV, alternative = "less", method = "kendall")
```

```
## Warning in cor.test.default(StudentSurvey$TimeReading, StudentSurvey$TimeTV, :
## Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  StudentSurvey$TimeReading and StudentSurvey$TimeTV
## z = -3.2768, p-value = 0.0005249
## alternative hypothesis: true tau is less than 0
## sample estimates:
##       tau
## -0.8045404
```

4. Perform a correlation analysis of:

```
- All variables
- A single correlation between two of the variables
- Repeat your correlation test in step 2 but set the confidence interval at 99%
- Describe what the calculations in the correlation matrix suggest about the relationship between the va
```

```
cor(StudentSurvey)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

```
cor.test(StudentSurvey$TimeReading, StudentSurvey$TimeTV)
```

```
##
##  Pearson's product-moment correlation
##
## data:  StudentSurvey$TimeReading and StudentSurvey$TimeTV
```

```
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9694145 -0.6021920
## sample estimates:
##       cor
## -0.8830677
```

```
cor.test(StudentSurvey$TimeReading, StudentSurvey$TimeTV,conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  StudentSurvey$TimeReading and StudentSurvey$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9801052 -0.4453124
## sample estimates:
##       cor
## -0.8830677
```

This correlation analysis suggests that TimeReading is negatively correlated with TimeTV. It also suggests that TimeReading is negatively correlated with Happiness. This is different from the correlation between TimeTV and Happiness, which is a positive value. The correlation values for anything with gender are very close to zero, which indicates that gender does not play a role in in the amount of time a person spends reading, the time they spend watching tv, or how happy they are.

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

```
cov(StudentSurvey$TimeReading,StudentSurvey$TimeTV)/(sd(StudentSurvey$TimeReading) * sd(StudentSurvey$T
```

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

```
summary(lm(TimeTV ~ TimeReading, data = StudentSurvey))$r.squared
```

```
## [1] 0.7798085
```

The coefficient of determination is the square of the correlation coefficient. It seems like there is a particularly strong relationship that is present between the two variables. I anticipate that my earlier intuition is correct, in that as time is spent doing one activity it limits a person's ability to spend it doing a different activity.

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

It is very tempting to say that watching TV causes students to read less, based on the correlation values that were computed. However, correlation does not imply causation. And even if watching tv causes students to read less, the computed correlation values do not imply that. This analysis does not show if the correlation happended due to random chance, if there is a hidden common factor, or if reading causes students to watch less tv. Correlation tells us to what degree things appear to be related, but speaks not of causality.

7. Use TV Time and Happiness while controlling for Gender and perform a partial correlation. Explain how this changes your interpretation and explanation of the results.

```
library(ppcor)
```

```
## Loading required package: MASS
```

```r
pcor.test(StudentSurvey$TimeTV, StudentSurvey$Happiness, StudentSurvey$Gender)
```

```
##    estimate    p.value statistic  n gp  Method
## 1 0.6435158 0.04469059  2.377919 11  1 pearson
```

The estimate shows that, even when controlling for gender, time spent watching TV correlates in the positive direction strongly with happiness. Additionally, the p value being less than 0.05 shows that this finding is, in some sense, statistically significant.