# Exercise 9 : Student Survey

Daniel Angel

January 17, 2021

*As a data science intern with newly learned knowledge in 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.*

Table 1: Student Survey Responses

| TimeReading | TimeTV | Happiness | Gender |
|---|---|---|---|
| 1 | 90 | 86.20 | 1 |
| 2 | 95 | 88.70 | 0 |
| 2 | 85 | 70.17 | 0 |
| 2 | 80 | 61.31 | 1 |
| 3 | 75 | 89.52 | 1 |
| 4 | 70 | 60.50 | 1 |
| 4 | 75 | 81.46 | 0 |
| 5 | 60 | 75.92 | 1 |
| 5 | 65 | 69.37 | 0 |
| 6 | 50 | 45.67 | 0 |
| 6 | 70 | 77.56 | 1 |

- A. 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.

Table 2: Covariance Matrix

| | TimeReading | TimeTV | Happiness | Gender |
|---|---|---|---|---|
| TimeReading | 3.0545455 | -20.3636364 | -10.350091 | -0.0818182 |
| TimeTV | -20.3636364 | 174.0909091 | 114.377273 | 0.0454545 |
| Happiness | -10.3500909 | 114.3772727 | 185.451422 | 1.1166364 |
| Gender | -0.0818182 | 0.0454545 | 1.116636 | 0.2727273 |

**Covariance is a calculation which indicates how one variables varies compared to another. Negative covariation indicates that as one variable increases the other will decrease and vice versa. A positive covariation means that as one variable increases or decreases the other will increase or decrease accordingly.**

- B. 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.

**It appears that the "Time (spent) Reading" variable is measured in hours whereas "Time (spent watching) TV" is given in minutes. The happiness score appears to be on a 1-100 scale. Lastly, the gender is a binary measurement with a value of 1 corresponding to one gender and a value of 0 corresponding to the other. The difference between the measurement of hours or minutes between TimeReading and TimeTV would make interpreting the covariance slightly more challenging. A solution would be scaling Reading to minutes, multiply values by 60, or TV to hours, divide by 60, would give us two variables measured on the scale. This would allow us to compare the covariances directly instead of trying to compare apples to oranges.**

- C. 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?

**Pearson's Correlation Test is my chosen method. I chose this because it is the default method when performing correlation tests in R. I predict that the parity(whether it is positive or negative) will directly correspond with the parity of the values of the covariance coefficient. Negative covariances will have negative correlation and positives will have have positive.**

- D. Perform a correlation analysis of:
  - i. All variables

| | TimeReading | TimeTV | Happiness | Gender |
|---|---|---|---|---|
| **TimeReading** | 1 | -0.8831 | -0.4349 | -0.08964 |
| **TimeTV** | -0.8831 | 1 | 0.6366 | 0.006597 |
| **Happiness** | -0.4349 | 0.6366 | 1 | 0.157 |
| **Gender** | -0.08964 | 0.006597 | 0.157 | 1 |

- ii. A single correlation between two a pair of the variables

```
##
##  Pearson's product-moment correlation
##
## data:  tr and tt
## 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
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  tr and tt
## 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
```

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

**TimeTV and TimeReading have a strong negative relationship. TimeReading has a weak negative relationship with Happiness. Finally, TimeTV and Happiness have a positive relationship. The relationships between Gender and the other variables is very weak and therefore practically non-existent. A positive relationship implies that as one variable increases or decreases the other does as well. A negative relationship implies that when one variable increases or decreases the other does the opposite.**

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

Table 4: Correlation Coefficient

| | TimeReading | TimeTV | Happiness | Gender |
|---|---|---|---|---|
| **TimeReading** | 1 | -0.8831 | -0.4349 | -0.08964 |
| **TimeTV** | -0.8831 | 1 | 0.6366 | 0.006597 |
| **Happiness** | -0.4349 | 0.6366 | 1 | 0.157 |
| **Gender** | -0.08964 | 0.006597 | 0.157 | 1 |

Table 5: Coefficient of Determination **My conclusions for the correlation coefficients are the same as my responses in Question D part iv. With resepct to the coefficient of determination my conclusions would be based on the commonly accepted interpretation of the coefficient of determination. The coefficient of determination is widely understood or viewed as a measure of how well data fits a linear model. The most significant result is the ~78% coefficient of determination between the Reading and TV time variables. This would cause me to have a large amount of confidence in the correlation calculation of those two variables.**

| | TimeReading | TimeTV | Happiness | Gender |
|---|---|---|---|---|
| **TimeReading** | 1 | 0.7798 | 0.1891 | 0.008036 |
| **TimeTV** | 0.7798 | 1 | 0.4052 | 4.352e-05 |
| **Happiness** | 0.1891 | 0.4052 | 1 | 0.02465 |
| **Gender** | 0.008036 | 4.352e-05 | 0.02465 | 1 |

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

**Based on my analysis, watching TV gave students less time to read and vice versa. This is**

**due to the negative correlation. It is also possible that increased time spent reading led to decreased time for watching TV and because the potential source of the effect could be on either end of this two way street we can't definitively say that one caused the other but we can say with a fair amount of certainty that there is some type of causal relationship beyond simple coincidence. The p-value being lower than .05 and the similarity of results across multiple methods and tests along with the over .80 correlation and the ~80% coefficient of determination causes me to pretty sure about the conclusion of my analysis based upon the results of my calculations.**

- G. 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.

Table 6: Partial Correlation Estimate controlled for Gender

| | Time Reading | Time Watching TV | Happiness score |
|---|---|---|---|
| **Time Reading** | 1 | -0.8729 | 0.5977 |
| **Time Watching TV** | -0.8729 | 1 | 0.3516 |
| **Happiness score** | 0.5977 | 0.3516 | 1 |

**I controlled for gender, effectively removing it from the equation. Normally, this creates a hypothetical scenario where you can test for the correlation of two or more variables while keeping one constant. However, in this case it doesn't really affect my results because the gender variable didn't have strong correlations anyways and the before numbers are very similar to the after numbers.**