ASSIGNMENT 5

Ramsey King

2021-04-30

Student Survey

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.

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.

Survey Covariance Calculations

```
##              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
```

I would use these calculations to get a sense of how related the variables are to each other. Mainly what I would look for is the sign of the number.

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.

It appears that the measurement for TimeReading is in hours, TimeTV is in minutes, Happiness is a percentage ranging from 0 to 100, and Gender is Binary. If we converted the time of measure ment for TimeReading to hours, and then recalculated the covaiance, we would get:

```
##              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
```

This could potentially an issue because now that we have changed the standard of measurements, the values for the covariance has changed. The better alternative would be to calculate the correlation between these variables in place of the covariance.

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?

For this test, I will choose the Pearson correlation. The reason why I have decided to use this test is because it allows for us standardize the covariance where we can gain a better effect of the variables between each other. My prediction is that as Reading time increases, so will happiness. I may be overly optimistic about this.

iv. Perform a correlation analysis of:

1. All variables

```
##             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
```

2. A single correlation between a pair of the variables (time spent reading and happiness)

```
## [1] -0.4348663
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  survey$TimeReading and survey$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8801821  0.4176242
## sample estimates:
##        cor
## -0.4348663
```

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

It appears that the following correlations exists: Time spent reading and Time spent watching TV and Happiness are negatively correlated, meaning the more time spent reading, the less time you watch TV and the less happy you are. I am not sure at this point that we can make a determination of how happiness, TV time, and reading time is correlated by gender.

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

```
##             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
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143
## TimeTV      0.779808529 1.0000000000 0.40520352 0.0000435161
## Happiness   0.189108726 0.4052035234 1.00000000 0.0246527174
## Gender      0.008035714 0.0000435161 0.02465272 1.0000000000
```

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

Watching more TV causes students to read less. The coefficient of determination, or R^2, between Time Reading and Time TV is 0.779, or 77.9%. This means that only 22.1% of the variability is accounted for by the other variables. Considering that there is not much room for other variables to affect reading, watching more TV causes students to read less. If we also look at the correlation coefficient between these two variables, we see that value as -0.88. This would suggest that is one variable increases (time watching TV), the other variable decreases (time reading) at almost the same rate.

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.

The first calculation is the correlation between the time spent reading and happiness while controlling for the time spent watching TV.

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

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     survey
```

```
##    estimate  p.value statistic  n gp  Method
## 1 0.3516355 0.319059  1.062425 11  1 pearson
```

Based on these results there is a positive correlation to happiness when reading time increases. This is a different result when time watching TV is not used as a control variable.

The second calculation is the correlation between the time spent watching TV reading and happiness while controlling for the time reading.

```
##    estimate    p.value statistic  n gp  Method
## 1 0.5976513 0.06804372  2.108388 11  1 pearson
```

Based on these results, there is still a positive correlation between the time spent watching TV and happiness.

The third calculation is the correlation between time spent reading and happiness while controlling for the gender.

```
##     estimate   p.value statistic  n gp  Method
## 1 -0.4277985 0.2174682 -1.338679 11  1 pearson
```

This result shows that there is a negative correlation between the time spent reading and happiness when gender is the controlling variable. I am not sure I understand that means because I feel like there needs to be separate statistical analysis computed for the same gender.