4.1 Assignment: Student Survey

Roni Kaakaty

6/27/2020

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.

```
##              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. The covariance function is used to find the variance between variables from their expected values. This is often used to identify whether or not the variables tend to move in tandem or have an inverse relationship from each other. In this example, the covariance function shows us that time reading has an inverse relationship with time on TV and happiness. Time on TV moves in tandem with happiness and has an inverse relationship with time reading. Happiness works in tandem with time spend on TV and has an inverse relationship with time spent reading.

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.

A. Time reading is being measured in hours, time on TV is being measured in minutes, happiness is an index (unit-less), and gender is given as 1 for male and 0 for female. Changing the measurement variables would have an impact on the covariance calculation as covariance doesn't use a standardized unit, it only tells us if variables have an increasing or decreasing relationship with each other. If we are attempting to find the covariance of different units we are unable to do that objectively since the units are not the same. We would not be able to tell if the covariance is large or small, just if they work in tandem or inverse of each other. A better alternative would be to multiply the standard deviations of the variables together then divide the covariance of the variables in question by that multiplied standard deviation amount. This will provide a range of -1:+1. +1 correlation indicates a positive correlation, 0 indicates no linear relationship, and -1 indicates a negative relationship.

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?

I chose to do a point-biserial correlation between happiness and gender. I used this correlation test because one of my variables (gender) is considered a discrete dichotomy.

```
# Happiness vs. Gender
cor.test(student_df$Happiness, student_df$Gender)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_df$Happiness and student_df$Gender
## t = 0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4889126  0.6917342
## sample estimates:
##       cor
## 0.1570118
```

```
gender_variability <- (.16)^2 * 100
```

This test confirms that gender only plays a tiny role in happiness measured in the data set (2.56%) variability. This leaves us with 97.44% variabilty that isn't attributed to the gender of the individual as to how it contributes to their happiness.

4. Perform a correlation analysis of:

   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
```

A single correlation between two a pair of the variables

```
# Time reading vs Time on TV

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

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

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

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

Describe what the calculations in the correlation matrix suggest about the relationship between the vari

There is a negative correlation between the time reading variable and the time on tv and happiness variables. Time watching tv has a positive correlation with happiness and a negative correlation with time reading. The happiness variable has a positive correlation with time watching tv and a negative correlation with time spent reading.

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

#correlation coefficient

```
##             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
```

#coefficient of determination

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

These results tell me that that the biggest factors of time spent reading in this data set, is the time spent watching tv variable, which shares 78% variability. This leaves only 22% variablity unaccounted for in determining time spent reading. The results also tell us that the happiness variable shares 41% variabiliy with the time spent on TV variable. While significant, there is still 59% variability that is unaccounted for that could lead to the happiness rating.

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

A. Yes, based on the correlation tests performed for question 4, there is a negative correlation (-.88) between students who spent more time watching than students who spent time reading. This inverse relationship meant that as one of the variables went up in time spent, the other went down.

7. 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.

A. For the partial correlation, the "controlled" variable that I chose was the happiness variable. The two non-controlled variables that I selected were the time spent reading variable and the time spent watching tv variable. When compared with each other, there is a strong inverse relationship that confirms the earlier correlation tests that showed that the more time spent watching TV means that there is going to be less time spent reading. When one variable goes up, the other goes down and vice versa.

```
pcor.test(x=student_df$TimeReading, y=student_df$TimeTV, z=student_df$Happiness)
```

```
##    estimate      p.value statistic  n gp  Method
## 1 -0.872945 0.0009753126 -5.061434 11  1 pearson
```