Exercise 9 - Student Survey

Shilpa Kolekar

October 4th 2020

Question: 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.

# read a CSV file into studnet\_df dataframe  
students\_df <- read.csv("C:/Users/shilp/Documents/GitHub/dsc520/data/student-survey.csv")  
  
cov(students\_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

By looking at the data, we can indicate that TimeTv and TimeReading are negatively correlated with each other. Which means when TimeTv increases, TimeReading decreases and vise versa. Happiness is negatively correlated with TimeReading but positively correlated with TimeTv. When TimeTv increases Happiness increases, but when TimeReading increases happiness decreases. Gender is the variable which is part of the dataset but does not impact any other variables.

Question: 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.

Each variable in survey data is different than other. TimeReading data is numeric and unit is hours. TimeTV is also numeric and unit is minutes. Happiness is measure in percentage and has floating values and Gender is in 0 and 1 so it could be Male = 0 and Female = 1 or vice versa. If units for variables are changed, covariance will drastically change and give different output.

correlation is the better alternative to covariance becuase correlation is in the range of 1 and -1 and it does not change evenif the unit of variable changes.

Question: 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?

cor(students\_df, method = "pearson")

## 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(students\_df, method = "spearman")

## TimeReading TimeTV Happiness Gender  
## TimeReading 1.00000000 -0.90725363 -0.4065196 -0.08801408  
## TimeTV -0.90725363 1.00000000 0.5662159 -0.02899963  
## Happiness -0.40651964 0.56621595 1.0000000 0.11547005  
## Gender -0.08801408 -0.02899963 0.1154701 1.00000000

cor(students\_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

There are three methods that can be used to perform correlation tests and they are Pearson, Spearman, and Kendall. if you compare numbers in each methods they are same or close to other method numbers. This explains that relationship between all three methods are almost the same.

The Kendall method does not support rcorr() function. Pearson is most appropriate for measurements taken from an interval scale, while the Spearman is more appropriate for measurements taken from ordinal scales. Out of three Pearson method is the best fit for the test.

Question: D (1)

Perform a correlation analysis of: All variables

cor(students\_df)

## 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

Question: D (2)

A single correlation between two a pair of the variables

cor.test(students\_df$TimeReading, students\_df$TimeTV)

##   
## Pearson's product-moment correlation  
##   
## data: students\_df$TimeReading and students\_df$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

Question: D (3)

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

cor.test(students\_df$TimeReading, students\_df$TimeTV, conf.level = 0.99)

##   
## Pearson's product-moment correlation  
##   
## data: students\_df$TimeReading and students\_df$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

Question: D (4)

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

* TimeReading has high negative correlation with TimeTV.
* TimeReading has low negative correlation with Happiness.
* TimeReading has no correlation with Gender.
* TimeTV has low positive correlation with Happiness.
* TimeTV has no correlation with Gender.
* Happiness has no correlation with Gender.

95% confidence level is the default level. At that confidence level, the correlation will be between -0.6021920 and -0.9694145. The p value is 0.0003153, which is less than 0.05. This confirms high negative correlation between TimeReading and TimeTV.

At 99% confidence level, the correlation will be between -0.4453124 and -0.9801052. The p value is 0.0003153, which is less than 0.05. This confirms high negative correlation between TimeReading and TimeTV.

Question: E

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

cor(students\_df)

## 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(students\_df)^2

## 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

Correlation coefficient measures relationship between two variables and tells how strong the relationship is, Coefficient of determination measures the amount of variability in one variable that is shared by another. coefficient of determination is a squared value of correlation coefficient, which is always positive. TimeReading and TimeTV variables have high negative correlation based on correlation coefficient. TimeReading shares 77.98% of variability of TimeTV.

Question: F

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

Both the variables show strong negative relationship and high variability. This means that increase in one can decrease another, and they share 77.98% of variability. This confirms that Watching more TV causes students to read less.

Question: 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.

library(ggm)  
pcor(c("TimeReading", "TimeTV", "Happiness"), var(students\_df))

## [1] -0.872945

TimeReading and TimeTV are the variables for which the partial correlation test is being performed while controlling Happiness variable. The correlation coefficient is -0.872945 which is very close to the correlation coefficient between TimeReading and TimeTV of -0.88306768 without any control variables. The correlation between TimeReading and TimeTV does not change much even after controlling the Happiness variable.