Exercise 9: Student Survey

Kolekar, Ajit

October 4th, 2020

# Assignment description

**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 the StudentSurvey.csv file.**

students\_df <- read.csv("C:/Users/ShilpaandAjit/Documents/GitHub/dsc520/data/student-survey.csv")

## 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.**

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

The covariance of two variables in a data set measures how the two are linearly related. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate a negative linear relationship between the variables.

The results indicate that TimeReading has negative linear relationship with TimeTV, Happiness, and gender. The results also indicate that TimeTV has positive linear relationship with Happiness and gender. Happiness also has positive linear relationship with Gender based on the results.

## 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.**

The survey variables are defined with different data types. TimeReading, TimeTV, and Happiness are numeric data types, and Gender is a categorical data type. The variables also have different units and the units are not consistent across variables such as TimeReading and TimeTV.

There is no maximum range for covarience. The covariance value may change with the measurement being used for the variables or if the unit of measurement is changed. This could create problem with determining the strength of the covariance.

The better alternative to covariance is correlation since the correlation can be in the range of -1 and +1. Closer the value to -1 and +1 would be that the correlation is strong between the variables. Negative correlation shows negative relationship and positive correlation shows positive relationship. Correlation is independent of the unit of measurement so it doesn’t matter if the two variables are measured in different ways.

## 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. When the correlation is calculated using all three methods, the direction of relationship matched for all three methods however the strength varied among the methods.

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. For these reasons, it makes sense to use Pearson method 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

* 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.

## 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

The default confidence interval is 95%. if you were to repeat the data gathering process and analysis identically, over and over, there is a 95% confidence that the correlation will be between -0.6021920 and -0.9694145. This confirms high negative coorelation between TimeReading and TimeTV. The p value is 0.0003153, which much less than 0.05. This indicates that the relation TimeReading and TimeTV is significant.

## 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

The confidence interval is set to 99%. if you were to repeat the data gathering process and analysis identically, over and over, there is a 99% confidence that the correlation will be between -0.4453124 and -0.9801052. This confirms high negative coorelation between TimeReading and TimeTV. The p value is 0.0003153, which much less than 0.05. This indicates that the relation TimeReading and TimeTV is significant.

## Question d.4

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

Based on the correlation analysis of allvariables:

* 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.

Based on the correlation matrix with default confidence level of 95%:

The default confidence interval is 95%. if you were to repeat the data gathering process and analysis identically, over and over, there is a 95% confidence that the correlation will be between -0.6021920 and -0.9694145. This confirms high negative coorelation between TimeReading and TimeTV. The p value is 0.0003153, which much less than 0.05. This indicates that the relation TimeReading and TimeTV is significant.

Based on the correlation matrix with confidence level of 99%:

The confidence interval is set to 99%. if you were to repeat the data gathering process and analysis identically, over and over, there is a 99% confidence that the correlation will be between -0.4453124 and -0.9801052. This confirms high negative coorelation between TimeReading and TimeTV. The p value is 0.0003153, which much less than 0.05. This indicates that the relation TimeReading and TimeTV is significant.

## 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

While correlation coefficient measures the relationship and strength between two variables, coefficient of determination measures the amount of variability in one variable that is shared by other. coefficient of determination is a squared value of correlation coefficient.The value of coefficient of determination is always positive since it’s a squared value. To understand the relationship between TimeReading and TimeTV, TimeReading and TimeTV highly and negatively correlated using 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.**

Based on the analysis, I can confirm that Watching more TV causes students to read less. Both the variables show strong negative relationship and high variability. This means that increase in one can decrease other, and they share 77.98% variability.

## 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

The partial correlation is being tested for TimeReading and TimeTV variables while controlling Happiness variable. The value with the correlation is -0.872945 which is very close to the correlation between TimeReading and TimeTV of -0.88306768. This tells us that the correlation between TimeReading and TimeTV does not change much even after controlling the Happiness variable.