ASSIGNMENT 7.2.2

Kiran Komati

2021-04-30

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_knit$set(root.dir = 'C:/Users/kiran/dsc520')
```

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.

```
studentsurvey_df <- read.csv("assignments/assignment07/student-survey.csv")
```

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.

```
cov(studentsurvey_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
```

Covariance is used to determine how the two variables are related to one another. The above results show that Timereading,TimeTV and TimeReading,Happiness variables are negatively related which means that as one variables increases the other increases or viceversa. Where as TimeTV and Happiness variables are positively related ..

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 seems that the hours is used for time reading and minutes for time tv. Converting time reading to minutes will increase the covariance. The problem is that we won’t be able to compare the co-variances in an objective way unless they are measured in same units.The better alternative would be to use correlation with with we can measure how strongly they are correlated.

```
cov(studentsurvey_df$TimeReading*60,studentsurvey_df$TimeTV)
```

```
## [1] -1221.818
```

```
cov(studentsurvey_df$TimeReading,studentsurvey_df$TimeTV)
```

```
## [1] -20.36364
```

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?

I will choose pearson correlation test as both the variables have interval data.But I'm not really sure why we should not use Kendall's Tau as the data set provided is small and there are few tied ranks as well as we order the variables TimeReading and TimeTv.. The test will yield negative correlation as the covariance has already yielded a negative relation between these two variables. Correlation is nothing but the scaled form of covariance.

iv.Perform a correlation analysis of:

1.All variables

```
cor(studentsurvey_df,use="everything",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
```

2.A single correlation between two a pair of the variables

```
cor(studentsurvey_df$TimeReading,studentsurvey_df$TimeTV,use="everything",method="pearson")
```

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

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

```
cor.test(studentsurvey_df$TimeReading,studentsurvey_df$TimeTV,method="pearson",conf.level = 0.99)
```

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

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

Value of -.8830 means that there is a strong negative correlation between timeTV and TimeReading and 0.636 suggests that there is positive correlation between TimeTV and Happiness. Gender has weak correlation with other variables.

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

Coefficient of determination is the square of the correlation coefficient. The coefficient of determination for the variables TimeReading and TimeTV is -0.8830677*-0.8830677=0.77980856278329. If we multiply this by 100 the value is 77.98 which means that TimeReading shares 77.98% variability with TimeTV.

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

Even though the two variables TimeTV and TimeReading have a strong negative correlation it doesn't necessarily mean that one caused the other as there can be other variables or factors affecting them which is known as third-variable problem. and also the correlation coefficient doesn't give us the direction of causality.

viii.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.

I'm doing the partial correlation between Timereading and TimemTV variables, by controlling the "Happiness". This has little to no effect as the partial correlation is almost same as the correlation when Happiness is not controlled.

```
library("ggm")
```

```
## Warning: package 'ggm' was built under R version 4.0.5
```

```
pcor(c("TimeReading","TimeTV","Happiness"),var(studentsurvey_df))
```

```
## [1] -0.872945
```