Week 7 - Exercise 7.2

Ganesh Kale

May 2nd 2021

ASSIGNMENT 05

Load Libraries

```
library(ggplot2)
library(ggm)
```

Load the `data/r4ds/heights.csv` to

```
heights_df <- read.csv("dsc520/data/r4ds/heights.csv")

head(heights_df,2)
```

```
##    earn   height    sex ed age  race
## 1 50000 74.42444   male 16  45 white
## 2 60000 65.53754 female 16  58 white
```

Using `cor()` compute correlation coefficients for

height vs. earn

```
cor(heights_df$height,heights_df$earn)
```

```
## [1] 0.2418481
```

age vs. earn

```
cor(heights_df$age,heights_df$earn)
```

```
## [1] 0.08100297
```

ed vs. earn

```
cor(heights_df$ed,heights_df$earn)
```

```
## [1] 0.3399765
```

Spurious correlation The following is data on US spending on science, space, and technology in millions of today's dollars and Suicides by hanging strangulation and suffocation for the years 1999 to 2009. Compute the correlation between these variables

```
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)

cor(tech_spending,suicides)
```

```
## [1] 0.9920817
```

Exercise 7.2 - Student Survey

load the student survey data

```
students <- read.csv('dsc520/data/student-survey.csv')
head(students,2)
```

```
##   TimeReading TimeTV Happiness Gender
## 1           1     90      86.2      1
## 2           2     95      88.7      0
```

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.

For all the variables

```
cov(students)
```

```
##               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
```

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

Correlation for Time spent on Reading vs Watching TV

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

The covariance between TimeReading and TimeTV is negative with value -20, which is low number so its weak negative covariance, which tells watching tv reduces the reading time but the scale/units of both of these variables are different minutes and hours, so we can not compare covariance in an objective way, its not measured on standardized scale.

2. What measurement is being used for the variables?

TimeReading - Time of reading is measured in hours.

TimeTV - Time spent to watch TV is measured in minutes

Happiness - It seems happiness score has been measured on scale 1-10 scale and transformed to percentage.

Gander - This is binary variable and measurement has been converted to 0 for Female and 1 for Male.

Explain what effect changing the measurement being used for the variables would have on the covariance calculation.

```
students$TimeReading_mins <- students$TimeReading * 60
cov(students$TimeTV,students$TimeReading_mins)
```

Changing TimeReading to minutes and then checking the covariance for TimeReading vs TimeTV

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

When changed Time spent on reading from hours to minute the covariance value increased from -20 to -1222,which is large number, this explains there is strong negative relationship between these variables.That means more students watch TV will spend less time in reading. Since we can say larger the covariance stronger the relationship but there is not exact number to tell how much large value to say there is strong relationship. when we compare -20 with -1222 definitely we can say that changing the scale of variables gives correct covariance between them.

Would this be a problem? Explain and provide a better alternative if needed.

Yes, If two variables do not have same scales, then we cannot truly assume the covariance value to gauge the relationship between the variables. This problem can be fixed by dividing the covariance by standard deviation, which gives us Correlation Coefficient.

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?

Pearson's Correlation Test will be performed to see the correlation between TimeReading and TimeTV. The Pearson's Correlation test is used to see the correlation of two continuous variables, because TimeReading and TimeTV are continuous variables. Based on the covariance result we can say that the Pearson's correlation will result negative correlation between TimeReading and TimeTV.

4. Perform a correlation analysis

4.1. All Variables

```
cor(students)
```

```
##                 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_mins  1.00000000 -0.883067681 -0.4348663 -0.089642146
##                 TimeReading_mins
## TimeReading           1.00000000
## TimeTV               -0.88306768
## Happiness            -0.43486633
## Gender               -0.08964215
## TimeReading_mins      1.00000000
```

4.2. A single correlation between two a pair of the variable

```
cor(students$TimeReading,students$TimeTV,method = "pearson")
```

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

```
cor(students$TimeReading,students$Happiness,method = "pearson")
```

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

```
cor(students$TimeReading,students$Gender,method = "kendall")
```

```
## [1] -0.07824608
```

```
cor(students$TimeTV,students$Happiness,method = "pearson")
```

```
## [1] 0.636556
```

```
cor(students$TimeTV,students$Gender,method = "kendall")
```

```
## [1] -0.02507849
```

```
cor(students$Happiness,students$Gender,method = "kendall")
```

```
## [1] 0.09847319
```

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

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

```
##
##  Pearson's product-moment correlation
##
## data:  students$TimeReading and students$TimeTV
## t = -5.6457, df = 9, p-value = 0.0001577
## alternative hypothesis: true correlation is less than 0
## 99 percent confidence interval:
##  -1.0000000 -0.5131843
## sample estimates:
##        cor
## -0.8830677
```

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

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

```
cor.test(students$TimeReading,students$Gender,method = "kendall",conf.level = 0.99)
```

```
## Warning in cor.test.default(students$TimeReading, students$Gender, method =
## "kendall", : Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  students$TimeReading and students$Gender
## z = -0.27832, p-value = 0.7808
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.07824608
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  students$TimeTV and students$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

```
cor.test(students$TimeTV,students$Gender,method = "kendall",conf.level = 0.99)
```

```
## Warning in cor.test.default(students$TimeTV, students$Gender, method =
## "kendall", : Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  students$TimeTV and students$Gender
## z = -0.091705, p-value = 0.9269
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.02507849
```

```
cor.test(students$Happiness,students$Gender,method = "kendall",conf.level = 0.99)
```

```
## Warning in cor.test.default(students$Happiness, students$Gender, method =
## "kendall", : Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  students$Happiness and students$Gender
## z = 0.36515, p-value = 0.715
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## 0.09847319
```

4.4. Describe what the calculations in the correlation matrix suggest about the relationship between the variables.

Be specific with your explanation.

The TimeReading is negatively related to the TimeTV with r value = -0.883067681, we can say that students spending more time watching TV spending less time reading. The TimeReading and Happiness negatively related, r = -0.4348663, Students are spending more time reading are less happy. The TimeTV and Happiness are positively related, r = 0.6365560, students watch more TV are more happy.

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

```r
cor(students$TimeReading,students$TimeTV,method = "pearson")
```

correlation coefficient

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

These two variables TimeReading and TimeTV are negatively correlated, Students who spend more time watching TV getting less time for eading.

```r
cor(students$TimeReading,students$TimeTV,method = "pearson")^2
```

coefficient of determination

```
## [1] 0.7798085
```

The Coefficient of Determination is = 0.78 or 78%, this means students time spent on watching TV causing 78% variation in time spent on reading. There is still 22% variability to be accounted for other variables causing less time for reading.

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

No, There is correlation between TimeReading and TimeTV, because correlation does not imply causation. There could be third factor that impacting student to read less, here correlation does not show the direction of causality. This statistics does not tell us that reading less causing watching more TV.

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.

```
pc <- pcor(c('TimeReading','TimeTV','Happiness'),var(students))
pc
```

The 3 variables - TimeReading, TimeTV and Happiness, Happiness is the variable we are going to control to see if high happiness impacting the time to read and watch TV

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

```
pc^2
```

```
## [1] 0.762033
```

```
pcor.test(pc,1,11)
```

```
## $tval
## [1] -5.061434
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.0009753126
```

Based on partial correlation value -0.87, tells that there is negative correlation between these two variables when controling the Happiness, we found Happiness has positive correlation with Watching TV but negative correlation with Reading Time, but here controling Happiness did not see much difference in the correlation values which is almost same as original correlation value -0.88. The R squared of Partial Correlation is 0.76 or 76%, that means students time spent on watching TV causing 76% variation in time spent on reading, which is similar to R-squared value 78%. The p value is less than 0.05 means the correlation is statistically significant. We do not see significant difference in correlation and partial correlation value after controlling the students happiness