ASSIGNMENT 5

Andrew Jordan

2022-05-01

Question 1

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

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

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

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

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

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

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

```
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
```

Question 2

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?

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.

```
students_df<-read.csv("data/student-survey.csv")
cov(students_df[,c("TimeReading","TimeTV","Happiness","Gender")])
```

```
##               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
```

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.

```
head(students_df)
```

```
##   TimeReading TimeTV Happiness Gender
## 1           1     90     86.20      1
## 2           2     95     88.70      0
## 3           2     85     70.17      0
## 4           2     80     61.31      1
## 5           3     75     89.52      1
## 6           4     70     60.50      1
```

Each variable appears to have its own measurement, though this is based on what I would consider a reasonable assumption for each variables measurement. Time from TimeReading(int) would appear to refer to an hour measurement, whereas TimeTv(int) appears to be in minutes. This is based upon the assumption that reporting one minute of time spent reading and 90 hours of time spent watching TV would be unlikely to occur within the same time frame that the survey is assumed to cover. Happiness(float) appears to be either a point value or percentile, but whether that is a point value determined by survey response of how happy the participant is or how happy a given participant self-reports compared to the other participants, is unclear. Gender(int) appears to be a binary "yes/no" measurement, but has no indication what gender is the baseline "yes".

At this point, the only measurement I believe could be changed and maintain any sort of data integrity without seeing the original survey would be to change the TimeReading from (the assumed measurement of)hours to minutes to match TimeTv, or vice versa. As seen in the results below, changing this measurement to more accurately reflect the assumed responses shows a much greater relationship between TimeReading and the other variables.

```
students_df_altered<-read.csv("data/student-survey.csv")
students_df_altered$TimeReading<-students_df_altered$TimeReading*60
cov(students_df_altered)
```

```
##              TimeReading        TimeTV   Happiness      Gender
## TimeReading 10996.363636 -1.221818e+03 -621.005455 -4.90909091
## TimeTV      -1221.818182  1.740909e+02  114.377273  0.04545455
## Happiness    -621.005455  1.143773e+02  185.451422  1.11663636
## Gender         -4.909091  4.545455e-02    1.116636  0.27272727
```

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 have chosen to use Pearson's correlation test as the data appears to be on an interval scale. I would expect the TimeTV and Happiness variables to have a negative correlation, however when examining the actual data the variables appear as if they will have a positive correlation.

```
x<-students_df[,"TimeTV"]
y<-students_df[,"Happiness"]
cor.test(x,y,method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05934031 0.89476238
## sample estimates:
##      cor
## 0.636556
```

4. Perform a correlation analysis of:

All Variables

```
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
```

A single correlation between a pair of the variables

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

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

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

```
cor.test(students_df$TimeReading,students_df$TimeTV,method = "pearson",conf.level = .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
```

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

The results show that the relationship between the TimeReading and TimeTV variables have a negative correlation, in that as more time is spent reading, less time is spent watching TV.

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

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

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

```
cor(students_df$TimeReading,students_df$TimeTV)^2
```

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

Looking at the above results, one can conclude that there appears to be a relatively high correlation between TimeReading and TimeTV, with TimeReading accounting for up to approximately 78% of variation in TimeTV.

6.Based on your analysis can you say that watching more TV caused students to read less? Explain. Watching more TV does indeed appear to cause students to read less. The negative correlation of .88 shows that as one of these variables rises in value the other decreases, showing that as more time was spent watching TV, less time was spent reading (and vice versa).

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.

```
library(ppcor)
```

```
## Loading required package: MASS
```

```
students_df_abridged<-students_df[, c("TimeReading","TimeTV","Happiness")]
#Partial correlation of TimeReading and TimeTV while controlling Happiness
pcor(students_df_abridged)
```

```
## $estimate
##             TimeReading     TimeTV Happiness
## TimeReading   1.0000000 -0.8729450 0.3516355
## TimeTV       -0.8729450  1.0000000 0.5976513
## Happiness     0.3516355  0.5976513 1.0000000
##
## $p.value
##              TimeReading       TimeTV  Happiness
## TimeReading 0.0000000000 0.0009753126 0.31905895
## TimeTV      0.0009753126 0.0000000000 0.06804372
## Happiness   0.3190589526 0.0680437248 0.00000000
##
## $statistic
##             TimeReading    TimeTV Happiness
## TimeReading    0.000000 -5.061434  1.062425
## TimeTV        -5.061434  0.000000  2.108388
## Happiness      1.062425  2.108388  0.000000
##
## $n
## [1] 11
##
## $gp
## [1] 1
##
## $method
## [1] "pearson"
```

After wrestling with this concept for a while, I believe I have correctly calculated the partial correlation while controlling Happiness, however I don't fully understand the concept well enough to understand and/or explain how changes my interpretation of the results.