Assignment Week 7

Karlie Schwartzwald

2022-07-22

Assignment 05

Set the working directory to the root of your DSC 520 directory

```
getwd()
```

```
## [1] "C:/Users/karli/OneDrive/Documents/Data_Science/DSC520_Stats_for_DS/DSC520/dsc520"
```

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

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

```
##    earn   height    sex ed age  race
## 1 50000 74.42444   male 16  45 white
## 2 60000 65.53754 female 16  58 white
## 3 30000 63.62920 female 16  29 white
## 4 50000 63.10856 female 16  91 other
## 5 51000 63.40248 female 17  39 white
## 6  9000 64.39951 female 15  26 white
```

Using `cor()` compute correlation coefficients for height vs. earn

```
cor(heights_df[, c("height", "earn")])
```

```
##           height      earn
## height 1.0000000 0.2418481
## earn   0.2418481 1.0000000
```

age vs. earn

```
cor(heights_df[, c("age", "earn")])
```

```
##             age       earn
## age  1.00000000 0.08100297
## earn 0.08100297 1.00000000
```

ed vs. earn

```
cor(heights_df[, c("ed", "earn")])
```

```
##             ed      earn
## ed   1.0000000 0.3399765
## earn 0.3399765 1.0000000
```

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)
print('The correlation between tech spending and suicides is:')
```

```
## [1] "The correlation between tech spending and suicides is:"
```

```
print(cor(tech_spending,suicides))
```

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

Student Survey

```
survey_df <- read.csv("student-survey.csv")
head(survey_df)
```

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.

```
##   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
```

```
cov(survey_df)
```

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.

```
##             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 indicates the relationship of two variables whenever one of those variables changes. In these results we can see that TimeReading has a negative covariance with TimeTV, Happiness, and Gender. This means that as TimeReading goes up, those variables move down. However, TimeTV has a positive covariance with Happiness, and thus, when TimeTV goes up so does happiness.

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. TimeReading appears to be measured in hours while TimeTV appears to be measured in minutes. Happiness is on a scale of 1-100 and Gender takes on a binary value. Changing the measurement being used for the variables will change the magnitude of the covariance value. This is a problem, because without a normalized measure of covariance, we can't really compare covariances between variables because magnitudes are meaningless. An improvement would be to use correlation which is normalized.

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? Perform a correlation analysis of:

All variables I will use a point-biserial correlation coefficient test correlation between Gender and all other variables because Gender has binary values to compare against continuous interval variables. I believe that the positive and negative signs on each correlation will match that of the covariance between each variable indicated above from the previous problem. For the rest of the analysis, I will be using the Pearson Correlation Coefficient because the rest of the variables are interval.

```
# Pearson's product-moment correlations
print(cor.test(survey_df$Happiness, survey_df$Gender))
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  survey_df$Happiness and survey_df$Gender
## t = 0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4889126  0.6917342
## sample estimates:
##       cor
## 0.1570118
```

```r
print(cor.test(survey_df$TimeReading, survey_df$Gender))
```

```
##
##  Pearson's product-moment correlation
##
## data:  survey_df$TimeReading and survey_df$Gender
## t = -0.27001, df = 9, p-value = 0.7932
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6543311  0.5392294
## sample estimates:
##         cor
## -0.08964215
```

```r
print(cor.test(survey_df$TimeTV, survey_df$Gender))
```

```
##
##  Pearson's product-moment correlation
##
## data:  survey_df$TimeTV and survey_df$Gender
## t = 0.01979, df = 9, p-value = 0.9846
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5956354  0.6040812
## sample estimates:
##         cor
## 0.006596673
```

```r
# Pearson Correlation Coefficient
cor(survey_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
```

```r
cor.test(survey_df$TimeReading, survey_df$TimeTV, method =
"pearson", conf.level = 0.95)
```

A single correlation between two a pair of the variables

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

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

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

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

```
cor(survey_df)
```

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

```
##             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
```

The correlation matrix tells us that the more time a person spends reading, the less time they will be watching TV. It will also reduce their happiness, but not affecting happiness nearly as much as it affects TV time.It also tells us that the more time a person spends watching TV, the less time they spend reading and the happier they are. Although, the affect on happiness is less significant than the effect on reading time. Gender is not closely correlated with any of the other variables, so we can conclude that Gender does not coincide much with more reading time or tv time.

```
cor(survey_df)^2
```

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

```
##             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
```

We can say that TimeReading shares 78.0% of the variability in TimeTV, 18.9% of the variability in Happiness. Gender shares 0.01% of the variability in TimeReading and none of the variability in TimeTV, and only 2.5% of the variability in happiness. TimeTV accounts for 40.5% of the variability in Happiness.

Based on your analysis can you say that watching more TV caused students to read less? Explain. No, we cannot conclude that watching more TV caused students to read less. This is because all that we have measured is correlation, not causation. In other words, we don't know whether watching more TV caused students to read less, or whether reading less caused students to watch less TV. Furthermore, there could be a third variable that is the cause of both TimeTV and TimeReading.

```
library(ggm)
```

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.

```
## Warning: package 'ggm' was built under R version 4.2.1
```

```
pcor<-pcor(c("TimeTV", "Happiness", "TimeReading"), var(survey_df))
print(pcor)
```

```
## [1] 0.5976513
```

In the code above we compared the partial correlation between TimeTV and Happiness, controlling for TimeReading. I appears that about half of the correlation between TimeTV and Happiness can be explained by variation in TimeReading.Therefore, the true correlation between TimeTV and Happiness is diminished.