Exercise 7.2

Nicole Buros

April 28th, 2022

Exercise 7.2.1, Assignment 5:

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/njack/OneDrive/Documents/DSC 520/dsc520")

## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("data/r4ds/heights.csv")

## 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 dollar
## 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 7.2.2, 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.

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.

```
setwd("C:/Users/njack/OneDrive/Documents/DSC 520/dsc520")
stusurv_df <- read.csv("data/student-survey.csv")

time_read_conv <- c(stusurv_df$TimeReading * 60)
cov(time_read_conv, stusurv_df$Happiness)
```

```
## [1] -621.0055
```

```
cov(stusurv_df$TimeTV, stusurv_df$Happiness)
```

```
## [1] 114.3773
```

a) We would use covariance of the survey variables reading time and TV time with happiness to see if they vary in the same direction.Based on the results, it would appear that time spent reading and percentage of happiness vary in opposite directions, and time spent watching TV and happiness vary in the same direction.

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.

a) Time Reading is in hours, Time TV is in minutes, Happiness is a percentage, and Gender is a 0/1 to correlate with Female/Male, presumably. Changing the measurement to minutes for the Time Reading variable would allow for it to be the same as Time TV and would allow for covariance/correlation to be correctly calculated.

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?

```
setwd("C:/Users/njack/OneDrive/Documents/DSC 520/dsc520")
stusurv_df <- read.csv("data/student-survey.csv")

time_read_conv <- c(stusurv_df$TimeReading * 60)
cor.test(time_read_conv, stusurv_df$Happiness, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  time_read_conv and stusurv_df$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates:
##        cor
## -0.4348663
```

```
cor.test(stusurv_df$TimeTV, stusurv_df$Happiness, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  stusurv_df$TimeTV and stusurv_df$Happiness
## 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
```

a) I chose to use the Pearson's correlation test since it is typically the most popular type used. I would predict a negative correlation with Reading & Happiness, and positive with TV & Happiness based on my covariance results.

iv. Perform a correlation analysis of: -All variables -A single correlation between two a pair of the variables -Repeat your correlation test in step 2 but set the confidence interval at 99% -Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```
setwd("C:/Users/njack/OneDrive/Documents/DSC 520/dsc520")
stusurv_df <- read.csv("data/student-survey.csv")

time_read_conv <- c(stusurv_df$TimeReading * 60)
cor(time_read_conv, stusurv_df$TimeTV)
```

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

```
cor(stusurv_df$TimeReading, stusurv_df$Happiness)
```

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

```
cor(stusurv_df$TimeReading, stusurv_df$Gender)
```

```
## [1] -0.08964215
```

```
cor(stusurv_df$TimeTV, stusurv_df$Happiness)
```

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

```
cor(stusurv_df$TimeTV, stusurv_df$Gender)
```

```
## [1] 0.006596673
```

```
cor(stusurv_df$Happiness, stusurv_df$Gender)
```

```
## [1] 0.1570118
```

```
cor.test(time_read_conv, stusurv_df$Happiness, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  time_read_conv and stusurv_df$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates:
##        cor
## -0.4348663
```

```
cor.test(time_read_conv, stusurv_df$Happiness, method = "pearson", conf.level = .99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  time_read_conv and stusurv_df$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(time_read_conv, stusurv_df$Happiness, method = "pearson")^2
```

```
## [1] 0.1891087
```

a) Changing the confidence interval did not have an effect on the resulting correlation.

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

a) The correlation coefficient is -.43, and the coefficient of determination is .19. Based on these results, I would say this means that although time spent reading highly correlated with happiness, it only could possibly account for 19% of it. That leave a lot of room for Happiness to be impacted by other variables.

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

a) I don't think it's possible to say from this data, they are highly negatively correlated (-.88) which suggests the more time a student spends reading, the less time they spend watching TV, but I can't say that it is the cause.

vii. 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.

```r
setwd("C:/Users/njack/OneDrive/Documents/DSC 520/dsc520")
stusurv_df <- read.csv("data/student-survey.csv")
library(ggm)
## Controlling for TimeTV
pcor(c("Happiness", "Gender", "TimeTV"), var(stusurv_df))
```

```
## [1] 0.1981457
```

a) I'm not sure the partial correlation result changes my interpretation much, other than that Gender may have a larger impact on the correlation than I originally though since the number was much lower than the correlation for just Happiness & TimeTv.

References

- R for Everyone (Lander 2014)
- Discovering Statistics Using R (Field, Miles, and Field 2012)

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using r*. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.

Lander, J. P. 2014. *R for Everyone: Advanced Analytics and Graphics*. Addison-Wesley Data and Analytics Series. Addison-Wesley. https://books.google.com/books?id=3eBVAgAAQBAJ.