assignment_05_DonahueAlan

Alan Donahue

24 July 2021

Assignment 05

```
# the working directory to the root of your DSC 520 directory
setwd("C:/Users/Alan Donahue/Documents/data science masters/DSC 520 Stats/GIT/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, use = "everything", method = "pearson")
```

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

```
#age vs. earn
cor(heights_df$age, heights_df$earn, use = "everything", method = "pearson")
```

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

```
#ed vs. earn
cor(heights_df$ed, heights_df$earn, use = "everything", method = "pearson")
```

```
## [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, use = "everything", method = "pearson")
```

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

Student Survey

Question 1

```
# the working directory to the root of your DSC 520 directory
setwd("C:/Users/Alan Donahue/Documents/data science masters/DSC 520 Stats/GIT/dsc520")

#loading the student-survey data set
studsurvey_df <- read.csv("data/student-survey.csv")

#Covariance of variables
cov(studsurvey_df$TimeReading, studsurvey_df$TimeTV, use = "everything", method = "pearson")
```

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

```
cov(studsurvey_df$TimeReading, studsurvey_df$Happiness, use = "everything", method = "pearson")
```

```
## [1] -10.35009
```

```
cov(studsurvey_df$TimeReading, studsurvey_df$Gender, use = "everything", method = "pearson")
```

```
## [1] -0.08181818
```

```
cov(studsurvey_df$TimeTV, studsurvey_df$Happiness, use = "everything", method = "pearson")
```

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

```
cov(studsurvey_df$TimeTV, studsurvey_df$Gender, use = "everything", method = "pearson")
```

```
## [1] 0.04545455
```

```
cov(studsurvey_df$Happiness, studsurvey_df$Gender, use = "everything", method = "pearson")
```

```
## [1] 1.116636
```

I would use the covariance calculation because it will inform me the direction of the relationship that the two variables have. For instance, the covariance between TimeReading and TimeTV is negative, which means that the two variables have an inverse relationship. On the other hand, TimeTV and Happiness have a positive covariance, which means that the two variables move in tandem.

Question 2

Looking at the data, it seems that TimeReading is based on hours, while TimeTV is based on minutes. Additionally, Happiness looks to be on a sliding scale from what I assume is 1 to 100. Finally, Gender is represented by 1 and 0 for the two different genders.

Changing the measurement being used for the data would alter the number for covariance. However, it still wouldn't alter whether a covariance is positive or negative. The purpose of covariance is to determine the relationship the relationship between two different variables. So in this case, the positive and negative aspect is what is important. Overall, having the number change would not truly affect the answer of covariance between two variables.

Question 3

I plan to use Pearson's r because the information is not ordinal which is needed for Spearnman's r or Kendall's tau. I would predict that there is a negative relationship between TimeReading and TimeTV because I am assuming that most people would focus on one or the other instead of both.

Question 4

```
# the working directory to the root of your DSC 520 directory
setwd("C:/Users/Alan Donahue/Documents/data science masters/DSC 520 Stats/GIT/dsc520")

#loading the student-survey data set
studsurvey_df <- read.csv("data/student-survey.csv")

#finding correlation of all variables
cor(studsurvey_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
```

```
#single correlation between two variables
cor.test(studsurvey_df$TimeReading, studsurvey_df$TimeTV, use = "everything", method = "pearson", conf.l
```

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

```
#99% confidence interval
cor.test(studsurvey_df$TimeReading, studsurvey_df$TimeTV, use = "everything", method = "pearson", conf.l
```

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

The correlation matrix shows that there is an inverse relationship between TimeReading and TimeTV. Additionally, there is an inverse relationship between TimeReading and Happiness. However, there is a positive relationship between TimeTV and Happiness.

Question 5

```
cor(studsurvey_df)^2
```

```
##             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
```

Based off the coefficient of determination, TimeReading and TimeTV share 77.98% of the variance. However, we cannot determine if TimeReading is the cause for TimeTV and vice versa.

Question 6

No I cannot say that watching more TV caused students to read less. There is potential for a third variable that is unaccounted for. Additionally, correlation coefficients say nothing about which variable causes the other to change.

Question 7

```
library(ggm)
pcor(c("TimeReading", "TimeTV", "Happiness"), var(studsurvey_df))
```

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

I performed a partial correlation on TimeReading and TimeTV, while Happiness was the variable I was "controlling." There was a sleight decrease in the correlation coefficient, but it doesn't change my interpretation of the results because it was such a small change.