7.2 Exercise

Harlan Wittlieff

10/17/2021

Assignment 04

1. Using `cor()` compute correclation coefficients for:

a. height vs. earn

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

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

b. age vs. earn

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

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

b. ed vs. earn

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

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

c. 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

```r
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
```

Student survey

a. 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.

```r
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
```

Calculating the covariance gives insight on to whether or not two variables are related to each other. A positive covariance means that as one variable deviates from the mean, the other deviates in the same direction. A negative covariance means that as one variable deviates from the mean, the other deviates in the opposite 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.

- TimeReading
  - This field appears to contain ranked values with ties.
- TimeTV
  - This field contains a numeral value from 50-95 in increments of 5. No units are noted.
- Happiness
  - This field provides a numerical value for happiness. My assumption would be happiness is provided on a scale of 0-100.
- Gender

- This field contains a binary value for gender. Which gender relates to which value is unspecified. Since these variables contain different scales of measurement, comparing the covariance between different variables provides little value. A better method is looking at the correlations between the variables.

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?

I'm going to use Kendall's tau to calculate correlation between TimeReading and TimeTv. I chose this test since it's the suggested method when dealing with small data sets containing tied ranked data. My prediction is that TimeReading and TimeTV will be negatively correlated as time spent doing one activity is time not available for the other.

iv. Perform a correlation analysis of:

```
cor(students_df, method = "kendall")
```

1. All Variables

```
##             TimeReading      TimeTV   Happiness      Gender
## TimeReading  1.00000000 -0.80454045 -0.28894280 -0.07824608
## TimeTV      -0.80454045  1.00000000  0.46304237 -0.02507849
## Happiness   -0.28894280  0.46304237  1.00000000  0.09847319
## Gender      -0.07824608 -0.02507849  0.09847319  1.00000000
```

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

2. A single correlation between a pair of the variables.

```
## [1] -0.8045404
```

```
bootTau <-function(students_df,i)cor(students_df$TimeReading[i],students_df$TimeTV[i], use = "complete.
library(boot)
boot_kendall <- boot(students_df, bootTau, 2000)
boot.ci(boot_kendall, conf = 0.99)
```

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

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_kendall, conf = 0.99)
```

```
##
## Intervals :
## Level      Normal              Basic
## 99%   (-1.0973, -0.5209 )   (-1.2218, -0.6091 )
##
## Level     Percentile            BCa
## 99%   (-1.0000, -0.3873 )   (-0.9775, -0.2122 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

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

- TimeReading & TimeTV
  - -.80
  - These two variables are strongly negatively correlated. However, as TimeReading is a ranked value this means that as the rank decreases for TimeReading, TimeTV increases.
- TimeReading & Happiness
  - These two variables are slightly negatively correlated. As TimeReading rank decreases, happiness increases.
- TimeReading & Gender
  - -.08 correlation. From the results TimeReading & gender have little correlation.
- TimeTv & Happiness
  - There is some correlation seen between these two variables. As more time is spent watching tv, happiness increases.
- TimeTv & Gender
  - Little correlation is seen.
- Happiness & Gender
  - Little correlation is seen.

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

Correlation coefficient:

```
cor(students_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
```

Coefficient of determination:

```
cor(students_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
```

- 78% of TimeReading's variability is shared with TimeTV, 19% with Happiness and 1% with Gender
- 41% of TimeTv's variability is shared with Happiness, and 0% with Gender
- 2% of Happiness's variability is shared with Gender

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

Assuming ranked values for TimeReading where 1 is the most time spent reading out of all samples, from our data more time spent reading correlates to more time spent watching TV. The two had a correlation of -.8. Which equates to a strong negative correlation. As TimeReading's rank decreases, TimeTV increases.

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.

```
library(ggm)
pcor(c("TimeReading","TimeTV","Gender"), var(students_df))
```

```
## [1] -0.8860628
```

Comparing the correlation between TimeReading & TimeTV while controlling for Gender resulted in a correlation of .87. This is a minor change from our correlation that was not controlled for Gender meaning that gender has very little impact on TimeReading & TimeTV.