DSC 520 Exercise 7.2

Taylor Imhof

10/17/2021

7.2.1 Complete assignment05

```r
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/taylo/Documents/dsc520-assignments")

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

## Using `cor()` compute correlation coefficients for
## height vs. earn
height_earn_cov <- cor(heights_df$height, heights_df$earn) # 0.242
sprintf('Correlation between height and earnings: %s', round(height_earn_cov,3))
```

```
## [1] "Correlation between height and earnings: 0.242"
```

```r
### age vs. earn
age_earn_cov <- cor(heights_df$age, heights_df$earn) # 0.081
sprintf('Correlation between age and earnings: %s', round(age_earn_cov,3))
```

```
## [1] "Correlation between age and earnings: 0.081"
```

```r
### ed vs. earn
ed_earn_cov <- cor(heights_df$ed, heights_df$earn) # 0.340
sprintf('Correlation between education and earnings: %s', round(ed_earn_cov,3))
```

```
## [1] "Correlation between education and earnings: 0.34"
```

```r
## 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)

# 0.9992 appear very correlated (almost perfect)
tech_suicide_cov <- cor(tech_spending, suicides)
sprintf('Correlation between tech spending and suicides: %s', round(tech_suicide_cov, 3))
```

```
## [1] "Correlation between tech spending and suicides: 0.992"
```

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

7.2.2.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

```
# set wd to have access to student survey data
setwd("C:/Users/taylo/Documents/dsc520-assignments")

# read student survey csv into data frame
stuData <- read.csv('data/student-survey.csv')

# run cov on newly created student survey data and explain why calculation would
# be used and what the results indicate
cov(stuData)
```

```
##              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
```

Covariance values indicate how much two variables vary together. If the values are high, then this would indicate a strong relationship between the two variables. Looking over the results of running cov(), it would appear that time spent watching TV strongly varies positively (114.38) with the happiness metric. This would lead me to believe that people who watch more TV have a higher sense of happiness. Another thing I noticed was that time spent reading appears to vary negatively (-20.36) with time spent watching TV. While this makes sense intuitively (if you spending more time doing something, you will spend less time doing something else), it was interesting to see this result numerically. It would also be important to note that even though there appears to be a relationship between these values after calculating covariance, further work is needed to establish a true connection. These values just help point my research efforts in the right direction.

7.2.2.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

```
# examine student survey data variables
stuData
```

```
##    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
## 7            4     75     81.46      0
## 8            5     60     75.92      1
## 9            5     65     69.37      0
## 10           6     50     45.67      0
## 11           6     70     77.56      1
```

```r
hours_to_min <- stuData$TimeTV / 60
stuData[2] <- cbind(hours_to_min)
cov(stuData)
```

```
##             TimeReading        TimeTV  Happiness        Gender
## TimeReading   3.05454545 -0.3393939394 -10.350091 -0.0818181818
## TimeTV       -0.33939394  0.0483585859   1.906288  0.0007575758
## Happiness   -10.35009091  1.9062878788 185.451422  1.1166363636
## Gender       -0.08181818  0.0007575758   1.116636  0.2727272727
```

It is difficult to say without knowing the exact details of what metrics were used when gathering the data for this survey, but it looks like the two time variables (TimeReading and TimeTV) are measured using different units of time. I would assume that the TimeReading variable is measured in hours and the TimeTV variable is measured in minutes (if they are measuing time spent per day). It could be that the TimeTV is hours watched per week, but without knowing the specific deatils of the study, I cannot conclude this. Also, the gender variable appears to have been codified (male = 1, female = 0 or vice versa). Also, not knowing the details of the study, it isn't clear which number is associated with each gender. I went ahead a converted all of the values for TimeTV to minutes (simply dividing by 60) and found this drastically changed the results. The covariance values between TV time and happiness dropped significantly (from 114.38 to 1.91), which while still positively related is much weaker than the original values produced. Also, while the read time and TV time were still negatively related, the strength of the relationship was much weaker (-20 to -0.34). So, changing the measurements used can have a drastic results on the values, which can in turn lead researchers to coming to false conclusions about the relationships between the data. Keeping the metrics for similar data types the same (e.g., hours for all time measurements) will help ensure the values are more consistent with how the variables are related.

7.2.2.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?

```r
cor.test(stuData$TimeReading, stuData$TimeTV, method='pearson', alternative = 'less')
```

```
##
##  Pearson's product-moment correlation
##
## data:  stuData$TimeReading and stuData$TimeTV
## t = -5.6457, df = 9, p-value = 0.0001577
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000 -0.6684786
## sample estimates:
##        cor
## -0.8830677
```

```
cor.test(stuData$TimeReading, stuData$TimeTV, method='spearman', alternative = 'less')
```

```
## Warning in cor.test.default(stuData$TimeReading, stuData$TimeTV, method =
## "spearman", : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  stuData$TimeReading and stuData$TimeTV
## S = 419.6, p-value = 5.761e-05
## alternative hypothesis: true rho is less than 0
## sample estimates:
##        rho
## -0.9072536
```

```
cor.test(stuData$TimeReading, stuData$TimeTV, method='kendall', alternative = 'less')
```

```
## Warning in cor.test.default(stuData$TimeReading, stuData$TimeTV, method =
## "kendall", : Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  stuData$TimeReading and stuData$TimeTV
## z = -3.2768, p-value = 0.0005249
## alternative hypothesis: true tau is less than 0
## sample estimates:
##        tau
## -0.8045404
```

I chose to perform each of the three tests that were covered in the reading this week, mostly to get a better feel for using them practically. Before running them, however, I felt that Kendall's tau correlation would be best suited for this data set, as I believe it can be easily ranked and the data set is quite small with only eleven observations. I also passed 'less' into the alternative argument, as I had a hunch that the two variables variables would be negatively correlated for reasons I stated earlier. All of the correlation values returned were negative and had p-values well below the level of significance. Spearman's was the highest/strongest at -0.907 which would indicate the two variables are inversely correlated at high strength. Kendall's was the lowest, though was still quite strong at -0.80. I am not quite sure which value would be the best to use in this situation, but I am sticking to my initial feeling that Kendall's was the best choice as this data set was so small.

7.2.2.iv

Perform a correlation analysis of:

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

```r
# revert student data with original values
setwd("C:/Users/taylo/Documents/dsc520-assignments")
stuData <- read.csv('data/student-survey.csv')

# all variables
cor(stuData)
```

```
##             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
# single correlation between a pair of two variables
cor(stuData$TimeTV, stuData$Happiness)
```

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

```r
# same thing but confidence interval at 99%
cor.test(stuData$TimeTV, stuData$Happiness, conf.level = 0.99, method='pearson')
```

```
##
##  Pearson’s product-moment correlation
##
## data:  stuData$TimeTV and stuData$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

Looking at the values produced in the correlation matrix, there are two numbers that jumped right out at me. The first was between TimeTV and TimeReading (-0.883). Since the sign of the value is negative, when one variable does up, the other goes down. This makes sense again as if you are spending time doing something, you have less time to do something else. The other number that was interesting was the value between Happiness and TimeTV (0.637). While this isn’t as high as the one discussed previously, it is still quite strong ($r > 0.5$). This would lead me to believe that people who spend more time watching TV are generally happier than those who don’t. I was also surprised to see that happiness and TimeReading were negatively correlated, which would mean that people who read more books report being less happy. Now, since the data set is so small it would be unwise to assume these results would be the same if data was sampled from a larger population. Still, it was interesting going through the motions of determining how variables related.

7.2.2.v.

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

```
# calculate correlation coefficient (r)
cor(stuData)
```

```
##             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
```

```
# calculate coefficient of determination (R^2)
cor(stuData) ^ 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
```

Using the coefficient of determination (COD), we can take Perason's R a step further for making the case that one variables causes another. The COD between reading and TV is 0.78, which would indicate that ~80% of the variability in the reading data is shared in the TV data. So, in combination with the high correlation value (r) of -0.883, TimeReading and TimeTV are very strongly related and would be a great choice of features for further research.

7.2.2.vi.

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

In summary, I would conclude that students who reported spending more time watching television did indeed read less than those who did not watch as much television. This was informed by the negative sign of Pearson's r (-0.883) as well as a high value for the coefficient of determination R^2 (0.780). I would have to bring up a personal bias as I would intuitively think these two variables would be inversely related, as I have stated for reasons above (e.g., if you do one, you cannot do the other). It would also be important to point out that there were only eleven observations in the dataset, therefore it would be advised to conduct research on a much larger sample.

7.2.2.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.

```
# import ggm library to use pcor() func for partial correlation
library(ggm)

# use pcor() on data set using gender as the control variable
pcor(c("TimeReading", "TimeTV", "Gender"), var(stuData))
```

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

```
# happiness as control variable
pcor(c("TimeReading", "TimeTV", "Happiness"), var(stuData))
```

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

I chose to examine the two variables that have been the focus for this assignment (TimeReading and TimeTV) and ran a partial correlation while controlling for the gender variable. I was curious as to whether gender had a significant impact on the relationship between watching TV and reading. However, after running the pcor() function, the value for r was very close to the r when simply calculating for Pearson's r (-0.886 vs -0.883). This would lead me to believe that a person's gender does not have a significant impact on the relationship between the time they spend watching TV versus reading books. I was also curious how much impact reported levels of happiness had on this relationship (plus it was as easy as swapping out gender for happiness in the pcor function :D). Again, the value produced was very close (-0.873 vs -0.883) so I would not concluse that happiness levels have any significant impact on the relationship between TimeTV and TimeReading.