DSC520 Week7 Exercise 7.2

Anjani Bonda

January 30th 2022

Part 1 - Assignment05

```
## Set the working directory to the root of your DSC 520 directory
setwd("/Users/anjanibonda/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$height, heights_df$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
```

```
## Correlation using other methods
cor(tech_spending, suicides, method = "kendall")
```

```
## [1] 0.9272727
```

```
cor(tech_spending, suicides, method = "spearman")
```

```
## [1] 0.9727273
```

Part 2 - Student Survey

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("/Users/anjanibonda/DSC520/dsc520")
studentsurvey_df <- read.csv("data/student-survey.csv")
head(studentsurvey_df)
```

```
##   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(studentsurvey_df)
```

```
##              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 is generally used to determine relationship between variables. A positive or negative covariance indicates whether the variables have strong or weak relationship with each other respectively.

Conclusion:

1. "TimeTV" and "Happiness" have a strong relation with positive covariance and more close to each other (114.377).
2. "TimeReading" and "TimeTV" have a weak relation with negative covariance and opposite to each other (-20.363).

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.

```
setwd("/Users/anjanibonda/DSC520/dsc520")
studentsurvey_df <- read.csv("data/student-survey.csv")
str(studentsurvey_df)
```

```
## 'data.frame':    11 obs. of  4 variables:
##  $ TimeReading: int  1 2 2 2 3 4 4 5 5 6 ...
##  $ TimeTV     : int  90 95 85 80 75 70 75 60 65 50 ...
##  $ Happiness  : num  86.2 88.7 70.2 61.3 89.5 ...
##  $ Gender     : int  1 0 0 1 1 1 0 1 0 0 ...
```

Gender is ideally considered a categorical variable and the numerical values here doesn't really indicate what value represents which gender. Also, it looks like TimeTV is in minutes, while TimeReading is in hours which indicates that the units are not consistent leading to incorrect results.

A better alternative is to make the units consistent as either hours/minutes for TimeTV and TimeReading, Gender to be a categorical variable with values of either Male/Female corresponding to 1/0 and Happiness possibly represented in percentage.

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 will choose simple correlation tests between two variables - TimeTV and Happiness and TimeReading and TimeTV, Since these variables seem to have some sort of relation with each other.

```
setwd("/Users/anjanibonda/DSC520/dsc520")
studentsurvey_df <- read.csv("data/student-survey.csv")
cor.test(studentsurvey_df$TimeTV, studentsurvey_df$Happiness)
```

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

```
cor.test(studentsurvey_df$TimeReading, studentsurvey_df$TimeTV)
```

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

Conclusion:

Above test results indicate that TimeTV and Happiness have a strong correlation (0.63) compared to TimeReading and TimeTV which has a weak/opposite correlation (-0.88) with each other.

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.

```
setwd("/Users/anjanibonda/DSC520/dsc520")
studentsurvey_df <- read.csv("data/student-survey.csv")
# correlation of all variables
cor(studentsurvey_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
```

```
# choosing TimeReading vs Happiness
cor(studentsurvey_df$TimeReading, studentsurvey_df$Happiness)
```

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

```
# with confidence interval of 99%
cor.test(studentsurvey_df$TimeReading, studentsurvey_df$Happiness, conf.level = 0.99)
```

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

Conclusion:

TimeReading has a negative correlation with Happiness. This means as TimeReading goes up, Happiness goes down and vice versa. TimeTV and Happiness are positively correlated. This suggests that more TimeTV leads to more happiness and vice-versa. TimeTV and TimeReading are negatively correlated as well. Gender doesn't seem to have much impact as all of those values are pretty low.

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

```
setwd("/Users/anjanibonda/DSC520/dsc520")
studentsurvey_df <- read.csv("data/student-survey.csv")
corcoff <- cor(studentsurvey_df)
coffd <- corcoff ^ 2
coffd
```

```
##             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
```

Conclusion:

The coefficent of determination values in this case is between 0 and 1 which shows its a good fit, As the values between 0 and 1 indicates the strength of linear regression model.

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

Based on the findings, we can conclude that watching more tv leads to less reading. Both TimeTV and TimeReading are negatively correlated variables, and with an coefficient of determination value nearing 1, we can say that there is a solid goodness of fit. Based on above points, it is safe to say that when tv time goes up,reading time goes down and vice-versa.

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.

```
setwd("/Users/anjanibonda/DSC520/dsc520")
studentsurvey_df <- read.csv("data/student-survey.csv")
library('ppcor')
```

```
## Loading required package: MASS
```

```
pcor.test(studentsurvey_df$TimeReading, studentsurvey_df$TimeTV,studentsurvey_df$Happiness)
```

```
##    estimate      p.value statistic  n gp  Method
## 1 -0.872945 0.0009753126 -5.061434 11  1 pearson
```

Conclusion:

A partial correlation between TimeReading, TimeTV with Happiness (controlling variable) is being performed above. The TimeReading and TimeTV appear to be negatively correlated and that low p-value suggests the same. Therefore, we can conclude that TimeTV and TimeReading are negatively correlated.

References

1. 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
2. R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
3. https://www.rdocumentation.org/packages/ppcor/versions/1.1/topics/pcor.test