Week 7: Assignments 7.2.1 and 7.2.2

Kimberly Cable

April 30 2022

Assignment 7.2.1

```
## 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 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)
```

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

Assignment 7.2.2

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?

Student Survey: first 5 rows:

```
##   TimeReading   TimeTV Happiness Gender
## 1           1 1.500000     86.20      1
## 2           2 1.583333     88.70      0
## 3           2 1.416667     70.17      0
## 4           2 1.333333     61.31      1
## 5           3 1.250000     89.52      1
## 6           4 1.166667     60.50      1
```

Question: *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.

```
##              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
```

The covariance is the indication if two variables are linearly related. You would use this information to determine the best correlation method to use.

- Time Reading and Time TV: Negative linear relationship
- Time Reading and Happiness: Negative linear relationship
- Time Reading and Gender: Negative linear relationship
- Time TV and Happiness: Positive linear relationship
- Time TV and Gender: Positive linear relationship
- Happiness and Gender: Positive linear relationship

Question: *ii*

Examine the Survey data variables.

What measurement is being used for the variables?

- TimeReading (integer) - number of minutes converted from number of hours
- TimeTV (integer) - number of minutes
- Happiness (double) - percentage
- Gender (integer) - categorical conversion to 1 and 0

Explain what effect changing the measurement being used for the variables would have on the covariance calculation.

Since the columns are all of a different measure trying to change them to the same measurement may be difficult. Gender for instance cannot be changed to number of minutes or hours.

Would this be a problem? Explain and provide a better alternative if needed.

Since you are trying basically to compare apples and oranges a better alternative is to standardize the comparison using the standard deviation of each of the variables to make them on the same playing field easier to compare.

Question: ***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?

Because Time Reading and Time TV are both intervals, I will be using the Pearson correlation method. I think the test will be negative as the more time you either read or watch tv the less time you do the other.

Question: ***iv***

Perform a correlation analysis of:

1. All variables

Correlation between the Student Survey variables

```
##             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
```

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

Pearson's Correlation Coefficient between Time Reading and Happiness with a 95% confidence level

```
##
##  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
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates:
##       cor
## -0.4348663
```

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

Pearson's Correlation Coefficient between Time Reading and Happiness with a 99% confidence level

```
##
##  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
```

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

Since the correlation coefficient is -0.43 and the p-value is 0.18, this suggests that they are negatively correlated. When one variable increases the other variable decreases. Since the p-value is greater than 0.05, we reject the null hypothesis that the two variables are not related.

Question: *v*

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

Correlation Coefficient between Time Reading and Time TV

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

Coefficient of Determination between Time Reading and Time TV

```
## [1] 0.7798085
```

The Correlation Coefficient is -0.88 => The Coefficient of Determination is 78%

We can say that Time Watching TV shares 78% of the variability of Time Reading.

Question: *vi*

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

Watching TV is only about 78% responsible for students to read less but there is also another 22% of something else that cause students not to read.

Question: *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.

I am seeing if Gender is controlling the relationship between Time Watching TV and Time Reading.

Partial Correlation between Time Reading and Time TV with Gender as the controlling factor

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

Coefficient of Determination between Time Reading and Time TV with Gender as the controlling factor

```
## [1] 0.7851073
```

Significance of the partial Correlation between Time Reading and Time TV with Gender as the controlling factor.

```
## $tval
## [1] -5.406281
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.0006411949
```

The correlation between Watching TV and Reading is the same even with the controlling variable of Gender so it is the same no matter if you are male or female. Even the Coefficient of determination is the same.

References