Exercise 7

Alissa Trujillo

2022-04-27

Assignment_05

Github Link:
https://github.com/alisssaa/dsc520/blob/master/completed/assignment05/assignment_05_TrujilloAlissa.R

Student Survey

i. Covariance

```
cov(survey_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
```

This covariance table gives us a look at correlations between each of our variables. Positive values mean that a higher value of one variable is typically associated with a higher value of the second variable. Covariance is however, difficult to use to make particularly useful predictions and assumptions due to the magnitude of the variable. By reducing it to a linear scale, we will have a better ability to actually analyze our data.

ii. Measurement

Time reading is measured in hours, while time watching TV is measured in minutes. Happiness is measured on a continuous scale according to survey data provided. Gender is measured in a binary fashion. By converting the time spent reading variable to minutes as well, we should have a more concise measure of covariance and correlation.

```
survey_df$TimeReading <- survey_df$TimeReading * 60
cov(survey_df)
```

```
##             TimeReading        TimeTV   Happiness      Gender
## TimeReading 10996.363636 -1.221818e+03 -621.005455 -4.90909091
## TimeTV      -1221.818182  1.740909e+02  114.377273  0.04545455
## Happiness    -621.005455  1.143773e+02  185.451422  1.11663636
## Gender         -4.909091  4.545455e-02    1.116636  0.27272727
```

By changing the measurement of the variables, the covariance of time reading and all other variables was highly elevated. This is due to minutes having a higher numerical value than hours. This does not change the way we read the table, as the correlations are still positive and negative, just by a different magnitude. Covariance does not have a specific scale, so it still must be rectified in order to make meaningful insights of the data. By calculating the correlation, we will be able to see how the variables interact on a more intuitive scale.

A correlation matrix will be a much more useful way to analyze our data. This will consolidate the values to a scale of -1 to 1. Negative numbers indicate a negative correlation, where a higher value in one variable is typically paired with a lower value of another variable. Positive numbers indicate that one variable is typically high when the other is also high. Correlation values closer to 0 mean that there is little co-dependence, but values closer to 1 or -1 mean that the variables have a higher co-dependence.

```
cor(survey_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
```

This correlation matrix allows us to see the same information on a scale from -1 to 1. Looking at it this way, we are able to determine that there is a very weak correlation (0.157) between gender and TV time, and a very strong negative correlation (-0.883) between time spend reading and time spent watching television. The other correlations have more moderate effects.

iii. Correlation Test

I will be conducting a Pearson's correlation test to test whether the correlation between gender and happiness is significant or not. I have chosen this particular test because I am looking to investigate a bivariate correlation, which is a correlation between two variables. Our data is continuous and seem to have a normal distribution, so it fits the criteria for a Pearson's correlation test. Since gender is a binary variable, it is suitable for a biserial correlation which can be achieved using this test.

```
cor.test(survey_df$Happiness, survey_df$Gender, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  survey_df$Happiness and survey_df$Gender
## t = 0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4889126  0.6917342
## sample estimates:
##       cor
## 0.1570118
```

This correlation test provides a p-value of 0.6448. This value is not significant so we cannot reject the null hypothesis that the true correlation of the variables is equal to 0. Additionally, the confidence interval (-0.489, 0.692) is quite wide and includes 0. Therefore we cannot support any significant correlation between the two variables.

iv. Correlation Analysis

1. All Variables Looking back at the correlation table we made earlier, we can see the correlations between all of our 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
```

Our correlation table shows negative correlations between time reading and all other variables (time spent watching television, happiness, and gender). Time spent watching television has a negative correlation with time spent reading, and a positive correlation with happiness and gender. Happiness shows a negative correlation with time spent reading, but a positive correlation with time spent watching television and gender. Gender shows a negative correlation with time spent reading, but a positive correlation with time spent watching television and happiness.

2. Pair of Variables We can focus on two specific variables and simply look at the correlation between the two. For this example we will be taking a closer look at the relationship between time spent reading and time spent watching television.

```
cor(survey_df$TimeReading, survey_df$TimeTV)
```

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

This output will simply give us one number which represents the correlation between two variables. This output shows that there is a high negative correlation between time spent reading and time spent watching television. To see if it is significant however, we will have to conduct additional tests. We will conduct a Pearson's correlation test at a 95% confidence interval to check the significance.

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

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

This correlation test returns a p-value of 0.003153. This means that the statistic is significant and we can reject the null hypothesis that there is no correlation between time spent reading and time spent watching television.

3. 99% Confidence Interval We can recreate the correlation test above using a 99% confidence interval in order to check if the p-value is still significant with 99% confidence.

```r
cor.test(survey_df$TimeReading, survey_df$TimeTV, conf.level = 0.99)
```

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

After conducting the test again with a 99% confidence interval, the results are the same. The p-value remains at 0.0003153, which is lower than 0.01, meaning that it is still a significant statistic at a higher confidence level. We can once again reject the null hypothesis that the true correlation is 0.

4. Correlation Matrix Looking again at our correlation matrix:

```
##             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
```

We can see that every variable has a correlation of 1 with itself, meaning it is perfectly correlated. This makes sense as the variable will always equal itself. Values closer to 0 mean that the effect size is very small and the relationship is not significant. Values closer to -1 or 1 mean that the effect size is large and the relationship is significant.

Negative numbers denote that the relationships are inverse, meaning that survey respondents that report spending a lot of time reading are significantly less likely to report that they spent a lot of time watching television (as it has a correlation of -0.883). Positive numbers denote that the variables are positively correlated, so survey respondents who indicate that they spend a significant amount of time watching TV are also likely to report higher levels of happiness (with a correlation of 0.637). This does not mean that these tendencies are true in all cases, just that it is more likely.

v. Correlation Coefficient & Coefficient of Determination

The correlation coefficient is shown in our correlation matrix:

```
##             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
```

These are the values we have been looking at thus far when evaluating the correlation between variables. We can calculate the coefficient of determination by squaring the correlations and multiplying them by 100 to convert them into percentages:

```
cor(survey_df)^2 * 100
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

This shows us how much variability for each measure is caused by one another. This means that the amount of variability shared by time watching television and time reading is 77.98%. This means that 22.02% of variability is accounted for by other variables. We can also see that the amount of variability shared between gender and time watching television is very low, only 0.004%, so 99.996% is accounted for by other variables. While we can see that there are correlations between many of our variables, each factor is influenced by many differing variables and no one variable accounts for 100% of variation in another.

vi. Analysis

We cannot say that watching more television has caused students to read less. While there is a high negative correlation between the two variables, we can only confidently say that there is a significant tendency for students who read more to watch less TV and vice versa. These statistics do not provide any evidence of causation in one direction or another, so we can not make any assumptions about causation.

vii. Partial Correlation

Using partial correlation, we can find the correlation of two variables while controlling for other variables. I am going to use this calculation to see what the correlation between happiness and watching television while controlling for time spent reading.

```
pc <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(survey_df))
pc
```

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

This number shows that the correlation between time spend reading and time spent watching television is -0.873 when we control for happiness. This is slightly lower than the initial correlation of -0.883 which took into account all of our data. This shows that happiness only has a small effect on the relationship between time spent reading and time spent watching TV.

```
pcor.test(pc, 1, 10)
```

```
## $tval
## [1] -4.734538
##
## $df
## [1] 7
##
## $pvalue
## [1] 0.002121315
```

Looking at the results of our partial correlation test, we can see that the difference is still significant with a p-value of 0.002. There is still a significant negative correlation between the two variables even after holding happiness constant. In this case, it did not make a significant difference to control for the happiness variable. However, in studies where there are multiple variables interacting in a more complex way, we will be able to use this tool to control for confounding variables and isolate our variables of interest.