Student survey analysis

Jackson Aquino

2022-04-30

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.

- 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. 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.

What stands out to me here is that it seems that TimeReading is in hours while TimeTV is in hours. If we multiply those times by 60 to obtain the number of minutes, it will not affect the correlation between the two variables.

- 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 chose the spearman correlation test as it’s a small sample size and also because we have the categorical variable for gender. As I have already ran a scatterplot I know that the correlation will be negative between time reading and time watching

- Perform a correlation analysis of:

1. All variables

```
##             TimeReading TimeTV Happiness Gender
## TimeReading        1.00  -0.91     -0.41  -0.09
## TimeTV            -0.91   1.00      0.57  -0.03
## Happiness         -0.41   0.57      1.00   0.12
## Gender            -0.09  -0.03      0.12   1.00
##
## n= 11
##
##
## P
##             TimeReading TimeTV Happiness Gender
## TimeReading             0.0001 0.2147    0.7969
## TimeTV      0.0001             0.0694    0.9325
```

```
## Happiness   0.2147      0.0694           0.7353
## Gender      0.7969      0.9325 0.7353
```

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

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

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

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

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

They suggest that students who spend more time watching TV are happier than people who spend more time reading (one has a negative correlation while the other one has a positive correlation). It also suggests that as people spend more time reading, they end up spending less time watching TV (negative correlation).

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

```
## [1] -0.8830677

## [1] 0.7798085
```

With the coefficient of determination equals 0.77, it means that 77% of the time reading can be explained by the time watching TV.

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

There's a negative correlation of 88% between these two variables, with an R square of 77%. The P-value is less than 5%, so it's very unlikely that this is happening by chance. However, the sample size is too small and there could be other factors that cause the smaller reading time, so I'd be hesitant in affirming this is caused by too much TV time.

- 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.

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

I'm controlling for Gender while analyzing the effect of those other two variables. P value is still very low below 1%.