Student Survey Analysis

Luke Syverson

April 30, 2023

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.

```
library(readr)
ss <- read_csv("C:/Users/syversonl/OneDrive - Tyson Online/Documents/GitHub/dsc520/completed/assignment
```

```
## Rows: 11 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): TimeReading, TimeTV, Happiness, Gender
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tr <- ss$TimeReading
tv <- ss$TimeTV
hap <- ss$Happiness
gen <- ss$Gender
```

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.

```
cov(ss, ss)
```

This code calculates the covariance between all the columns in the table. Time Reading has a strong negative covariance to the other variables, meaning that more time reading equates to less of anything else in the sample. Time TV is has a strong positive covariance to Happiness and a positive covariance to gender, which implies that one gender may prefer more TV Time than the other, and Happiness is purported to be higher given greater TV time within this sample. Happiness has a negligible positive covariance with gender, implying neither gender maintains higher happiness than the other. It's difficult to interpret any meaning within the magnitude of covariance, especially considering the scale and range of the variables.

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.

```
View(ss)
```

Happiness seems to be a percentage, TimeTV could be a range from 0:100 or a measure of a unit of time spent within a time frame, Gender is binary and represented with [0, 1], and TimeReading is a whole number that represents some amount of reading, presumably hours within a time frame (day, week). Using a different scale for the 0:100 variables would change the magnitude of comparison between the smaller TimeReading and Gender variables, but would not change the direction of covariance. An alternative scale where the ranges were of similar magnitude would bring slightly more meaning to the magnitude of covariance (perhaps on an exponential basis in the case of Happiness). Happiness has to be measured ordinarily (or binarily), but TimeTV does not have to be measured as such.

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?

Using a point-biserial correlation would account for the occasions where one of the variables is dichotomous, but it is not included in the native r function for correlation. The Spearman rank correlation carries no assumptions of the distributions of data, which makes it suitable for ordinal data such as is provided. Additionally, TimeReading & TimeTV seems to be monotonic in nature, possibly along with Happiness. I suspect positive correlations between TimeTV & Happiness, and negative for TimeReading between both TimeTV and Happiness. I'm not sure what to expect for gender, given the binary nature of the variable.

Perform a correlation analysis of:

All variables

```
cor(ss,ss, method = "spearman")
```

```
##             TimeReading      TimeTV  Happiness      Gender
## TimeReading  1.00000000 -0.90725363 -0.4065196 -0.08801408
## TimeTV      -0.90725363  1.00000000  0.5662159 -0.02899963
## Happiness   -0.40651964  0.56621595  1.0000000  0.11547005
## Gender      -0.08801408 -0.02899963  0.1154701  1.00000000
```

A single correlation between two a pair of the variables

```
cor(hap, tr, method="spearman")
```

```
## [1] -0.4065196
```

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

```
cor.test(hap,tr, method = "spearman", conf.level = 0.99, exact=FALSE) # Exact=FALSE given ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  hap and tr
## S = 309.43, p-value = 0.2147
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.4065196
```

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

It appears there is a moderately negative association of rank (rho = -0.41) between TimeReading and Happiness. The p-value < alpha (0.05) indicates the presence of a definite relationship between the variables, rejecting the null hypothesis. A Pearson's test may understate the magnitude of this relationship.

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

```
model <- lm(hap~tr)
summary(model)
```

```
##
## Call:
## lm(formula = hap ~ tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.628 -10.129   3.960   9.617  14.056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   85.629      9.355   9.153 7.44e-06 ***
## tr            -3.388      2.339  -1.449    0.181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.93 on 9 degrees of freedom
## Multiple R-squared:  0.1891, Adjusted R-squared:  0.09901
## F-statistic: 2.099 on 1 and 9 DF,  p-value: 0.1813
```

It appears that nearly 20% of Happiness variation may be explained by TimeReading.

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

While my analysis focused on reported Happiness according to TimeReading, I noted that TimeTV had a strong negative correlation to TimeReading. Let's test TimeReading and TimeTV to see if there's a definitive relationship.

```
cor.test(tv,tr, method = "spearman", conf.level = 0.99, exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  tv and tr
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9072536
```

It appears there is absolutely a negative relationship according to the tiny p-value.

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.

```
library(ppcor)
```

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

```
pcor.test(tv, tr, gen, method = "spearman")
```

```
##     estimate      p.value statistic  n gp   Method
## 1 -0.9137348 0.0002180915 -6.360723 11  1 spearman
```

The partial correlations (controlling for gender) don't change the correlation result or the p-value, meaning that the analysis interpretation remains unchanged. This could be due to gender's dichotomous nature or the small sample size.