Assignment 5.1 Student Survey

Larry Heckel

April 12, 2019

Assignment

As a data science intern with newly learned knowledge and 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.

a. 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.

b. 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.

c. 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?

d. Perform a correlation analysis of:

   1. All variables
   2. A single correlation between two 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.

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

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

g. Use TV Time and Happiness while controlling for Gender and perform a partial correlation. Explain how this changes your interpretation and explanation of the results.

```
# libraries to be used
library(ggplot2)
library(readr)
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(ggm)
```

```
## Loading required package: igraph
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
##
## Attaching package: 'ggm'
```

```
## The following object is masked from 'package:igraph':
##
##     pa
```

```
## The following object is masked from 'package:Hmisc':
##
##     rcorr
```

```r
library(polycor)

# read the file
studentSurvey <- read_csv("student-survey.csv")
```

```
## Parsed with column specification:
## cols(
##   TimeReading = col_double(),
##   TimeTV = col_double(),
##   Happiness = col_double(),
##   Gender = col_double()
## )
```

```r
View(studentSurvey)
str(studentSurvey)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 11 obs. of  4 variables:
##  $ TimeReading: num  1 2 2 2 3 4 4 5 5 6 ...
##  $ TimeTV     : num  90 95 85 80 75 70 75 60 65 50 ...
##  $ Happiness  : num  86.2 88.7 70.2 61.3 89.5 ...
##  $ Gender     : num  1 0 0 1 1 1 0 1 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   TimeReading = col_double(),
##   ..   TimeTV = col_double(),
##   ..   Happiness = col_double(),
##   ..   Gender = col_double()
##   .. )
```

Section A

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(studentSurvey$TimeReading, studentSurvey$TimeTV)
```

```
## [1] -20
```

```
cov(studentSurvey$TimeReading, studentSurvey$Happiness)
```

```
## [1] -10
```

```
cov(studentSurvey$TimeReading, studentSurvey$Gender)
```

```
## [1] -0.082
```

```
cov(studentSurvey$TimeTV, studentSurvey$Happiness)
```

```
## [1] 114
```

```
cov(studentSurvey$TimeTV, studentSurvey$Gender)
```

```
## [1] 0.045
```

```
cov(studentSurvey$Happiness, studentSurvey$Gender)
```

```
## [1] 1.1
```

Section A Answer

Covariance is a measure of how two variables are related to each other. The measure indicates how each of the variables deviates from its mean value, in relation to the other variable. A positive covariance indicates that as one variable deviates from its mean, the other one deviates in the same direction (positive-positive or negative-negative). A negative covariance indicates that as one deviates from its mean, the other changes in the opposite direction (positive-negative).

In general, the larger the covariance, the stronger the relationship, however the problem with the covariance is that it is not a standard measure, and the magnitude of the value depends on the scale of the two variables. As a result, we are only able to compare the covariance of populations that utilize the same measurement scales, and we cannot compare the covariance of those populations that use different scales.

The results of the covariance calculations indicate a positive relationship between the time spent watching TV and happiness, and a negative relationship between time spent reading and happiness. Additionally, the time spent TV watching and reading are negatively related, indicating that reading time and TV watching time potentially conflict with each other, and the students are doing one at the expense of the other.

Because Gender is represented by a 0/1 value, it is possible to calculate covariances between Gender and the other variables, however we don't know from the data provided which value represents Female, and which represents Male. The covariance calculations involving Gender and the other variables don't appear to show a significant direction, as they yield positive and negative values close to 0. I don't think that, based on the results of these calculations, that I can make a definitive statement on the relationships between Gender and the other variables.

Section B (Question and Answer)

Examine the Survey data variables.

1. What measurement is being used for the variables?

   We actually don't know what measurement is being used for the variables. While all of the data points are integers, we don't know what the scales are, or if the numbers represent word, or which numbers would be considered "better". We could make assumptions based on the column header names, however they would simply be assumptions, and not definitive.

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

   If the measurements being used for the variables were to change, then the covariance values would change. Assuming that the measurement direction would stay the same, then the covariance would stay either positive or negative, but the value of it would change.

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

   Changing the measurement would not be a problem in and of itself. The problem that exists with the current data, and would still exist with newly measured data, is that the values are just numbers, with no definition or meaning to derive insights from.

   The best alternative would be for the two Time Variables to be defined with the same scale, whatever that might be. It could be percentage of free time, actual time in minutes or hours, or some other value, so long as both variables contain the same scale. Additionally, it would be helpful for us to understand how the Happiness scale is defined, and if there were key words provided to help the students to assign a number to their feelings.

Section C

c. 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 would perform a Pearson correlation coefficient test, because it allows us to have a standard set of units for the strength of the relationships. The result of this test ranges from -1 to 1, with the extreme value indicating the strongest positive or negative relationship, and values closer to 0 indicating weak, or no, relationships.

   Without yet running the tests, I believe that the correlations will stay with the positive or negative sign of their covariances.

Section D

d. Perform a correlation analysis of:
   1. All variables (see results below)
      What we see from running the correlation on all variables is as follows:
      a. Strong negative correlation between time watching TV and time reading. It leads to the potential conclusion that the survey defines time in a finite quantity during which the student either reads or watches TV.
      b. Happiness had a strong positive correlation to TV time, and a weaker, but still somewhat significant, negative correlation to reading time.
      c. As all of the correlation values with gender are close to 0, there does not seem to be a gender effect on the relationships.

```
cor(studentSurvey, use="complete.obs", method="pearson")
```

```
##             TimeReading  TimeTV Happiness  Gender
## TimeReading        1.00 -0.8831     -0.43 -0.0896
## TimeTV            -0.88  1.0000      0.64  0.0066
```

```
## Happiness         -0.43  0.6366      1.00  0.1570
## Gender            -0.09  0.0066      0.16  1.0000
```

2. A single correlation between two of the variables (result below)

   The Pearson Correlation between Time Reading and Happiness is -0.43, indicating that as one increases, the other decreases, or a negative relationship.

```
cor(studentSurvey$TimeReading, studentSurvey$Happiness, use="complete.obs", method="pearson")
```

```
## [1] -0.43
```

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

   The p-value result is 0.2, meaning that we accept the null hypothesis that there is a negative relationship between the Time Reading and Happiness.

```
cor.test(studentSurvey$TimeReading, studentSurvey$Happiness, alternative="two.sided", method="pearson",
```

```
##
##  Pearson's product-moment correlation
##
## data:  studentSurvey$TimeReading and studentSurvey$Happiness
## t = -1, df = 9, p-value = 0.2
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.88  0.42
## sample estimates:
##   cor
## -0.43
```

4. Describe what the calculations in the correlation matrix suggest about the relationship

   a. Strong negative correlation between time watching TV and time reading. It leads to the potential conclusion that the survey defines time in a finite quantity during
   b. Happiness had a strong positive correlation to TV time, and a weaker, but still somewhat significant, negative correlation to reading time.
   c. As all of the correlation values with gender are close to 0, there does not seem to be a gender effect on the relationships.

Section E

e. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results. (result below)

   The coefficient of determination is the amount of variability in one variable that is shared by the other. It is a measure of the substantive importance of the relationship between the two variables, and the value is expressed as a percentage.

   In the results below, we see that Time Reading and TV Time have the strongest coefficient of determination, at 78%. This means that 78% of the variation of each of these is shared between them. This is the strongest relationship amongst any of the variables. The second strongest shared variation is between TV Time and Happiness, at just over 40%. The shared variation between Happiness and Reading Time is less, at 19%, and the remaining relationships are relatively insignificant.

Correlation Coefficient

```r
cor(studentSurvey, use="complete.obs", method="pearson")
```

```
##             TimeReading  TimeTV Happiness  Gender
## TimeReading        1.00 -0.8831     -0.43 -0.0896
## TimeTV            -0.88  1.0000      0.64  0.0066
## Happiness         -0.43  0.6366      1.00  0.1570
## Gender            -0.09  0.0066      0.16  1.0000
```

Coefficient of Determination

```r
cor(studentSurvey, use="complete.obs", method="pearson")^2
```

```
##             TimeReading   TimeTV Happiness   Gender
## TimeReading       1.000 0.779809     0.189 0.008036
## TimeTV            0.780 1.000000     0.405 0.000044
## Happiness         0.189 0.405204     1.000 0.024653
## Gender            0.008 0.000044     0.025 1.000000
```

Section F

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

   We cannot determine causality in the negative relationship between TV and Reading times. While they are clearly negatively related, these tests do not determine or imply that one caused the other, only that they are related in that manner.

Section G

Use TV Time and Happiness while controlling for Gender and perform a partial correlation. Explain how this changes your interpretation and explanation of the results. (results below)

```
The partial correlation between TV Time and Happiness, controlling for Gender, is 0.64, which
  is the same as when Gender was not controlled for. What this would tell us is that Gender is
  not an important factor on the relationship between the two variables. The relationship is
  the same for both men and women (although, as I noted above, we don't know which of the 0/1
  pair values is which sex).
```

```r
studentSurvey2 <- studentSurvey[, c("TimeTV","Happiness","Gender")]
partCor <- pcor(c("TimeTV","Happiness","Gender"),var(studentSurvey2))
partCor
```

```
## [1] 0.64
```