Assignment 05 Part 2

David Pahmer

2022-04-27

Student Survey 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.

```
setwd("C:/users/pahme/onedrive/documents/github/dsc520")
studsurv <- read.csv("./data/student-survey.csv")
knitr::opts_chunk$set(echo = TRUE)
```

```

library(ggplot2)
library(Hmisc)
```

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.

```
## 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
```

```
cov(studsurv$TimeReading,studsurv$TimeTV)
```

```
## [1] -20.36364
```

The task is to try to understand the relationship between the time spent watching TV and the time spent reading. Perhaps there is a relationship, and those students who spend more time watching also spend more time reading, or perhaps those who spend more time watching spend less time reading. On the other hand, perhaps there is no relationship between the times, and that some of the people who responded spend more time reading and also more time watching, and some spend more time reading and less time watching, and vice versa.
The covariance calculation can determine whether the two variables are related and in what direction- if the covariance results in a value near 0, then there is little to no relationship. If the value is high positive, then more time reading would go together with more time watching; and if the result is large and negative, then the more time reading will go with less time watching.
Since the result is -20.4 we can say that there seems to be some relationship in that more time reading goes

with less time watching, but we cannot say how strongly that relationship holds, and we also cannot say whether either causes the other.

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. The variables we are looking at are both supposed to measure time. However, one of them (time watching) is showing values from 50 to 90, which seems to be in minutes, although the other one (time reading) shows values from 1 to 6, which could indicate hours or half-hours, or some other type of scale that gives us ordinal values. It would be better to know precisely what the values represent, for more accurate analysis, but since we don't know, let's assume that it represents units of half-hours. Thus, we can convert the times into minutes reading.

```
studsurv$readingmins <- studsurv$TimeReading*30
cov(studsurv$TimeTV, studsurv$readingmins)
```

```
## [1] -610.9091
```

The covariance is still negative, as we expected, but it is much greater, indicating a larger relationship, but actually only indicating larger numbers for the existing relationship so no better understanding is gained. So it isn't a problem, but it isn't a solution either. To learn the strength of the relationship in terms that we can interpret, we need to determine the correlation rather than the covariance.

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? Since we are trying to measure the correlation between two variables, whose distributions are not normal and whose values may be ordinal rather than continuous, let's use the Spearman's rho. We predict that the test will show a negative correlation (inverse relationship) just as the covariance shows. First let's rerun the analysis of the two time variables using Spearman's rho:

```
rcorr(studsurv$TimeReading, studsurv$TimeTV, type="spearman" )
```

```
##       x     y
## x  1.00 -0.91
## y -0.91  1.00
##
## n= 11
##
##
## P
##   x      y
## x        1e-04
## y 1e-04
```

Using this statistic, we see that there seems to be a strong inverse relationship between time spent reading and time spent watching. The p-value is quite low enough to have confidence in this result.

```
cor(studsurv[,c(1:4)], method="spearman")
```

Perform a correlation analysis of: All variables- A single correlation

```
##             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
```

This indicates that in addition to the previous strong correlation, between reading time and watching time, there is an additional correlation between happiness and TV time. Let's explore that further.

between two (a pair) of the variables Let's test for a relationship between happiness and TV watching time. These both seem to be continuous variables, but we don't really know the scales so perhaps the pearson is preferred, or perhaps the Spearman's is preferred, so let's run both:

```
cor.test(studsurv$Happiness, studsurv$TimeTV, method="spearman")
```

```
## Warning in cor.test.default(studsurv$Happiness, studsurv$TimeTV, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  studsurv$Happiness and studsurv$TimeTV
## S = 95.432, p-value = 0.06939
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5662159
```

```
cor.test(studsurv$Happiness, studsurv$TimeTV, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  studsurv$Happiness and studsurv$TimeTV
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05934031 0.89476238
## sample estimates:
##      cor
## 0.636556
```

```

cor.test(studsurv$Happiness, studsurv$TimeTV, method="pearson", conf.level = .99)
```

Repeat your correlation test in step 2 but set the confidence interval at 99% Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```
##
##  Pearson's product-moment correlation
##
## data:  studsurv$Happiness and studsurv$TimeTV
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

```
print("Coefficient of Determination")
```

```
## [1] "Coefficient of Determination"
```

```
cor(studsurv)^2
```

```
##             TimeReading       TimeTV  Happiness       Gender readingmins
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143 1.000000000
## TimeTV      0.779808529 1.0000000000 0.40520352 0.0000435161 0.779808529
## Happiness   0.189108726 0.4052035234 1.00000000 0.0246527174 0.189108726
## Gender      0.008035714 0.0000435161 0.02465272 1.0000000000 0.008035714
## readingmins 1.000000000 0.7798085292 0.18910873 0.0080357143 1.000000000
```

Although the Pearson gave a 95% confidence interval of positive values, the 99% does cross 0, which weakens our confidence in the correlation.

As for the Coefficient of Determination, it shows 40% of the variability of Happiness is shared with the variability of TV watching time, which seems pretty significant (but I don't know enough about how this measure indicates anything). However the variability of reading time is shared with the variability of TV time to the extent of 79%, which looks particularly important (although we already know about the correlation, so there's something subtle going on here).

Based on your analysis can you say that watching more TV caused students to read less? Explain. We can say that there seems to be a correlation between watching more TV and spending less time reading, but we cannot say whether the one causes the other. It's possible that the causality is reversed, and that the students choose their reading time but their watching time is dependent on how much time they have left over; or some other factor is responsible for both of those variables. So we cannot say, based on this analysis, that watching more TV caused students to read less.

```

library(ppcor)
```

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.

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

```
pcor.test(studsurv$Happiness, studsurv$TimeTV, studsurv$Gender, method="spearman")
```

```
##    estimate    p.value statistic  n gp   Method
## 1 0.5736413 0.08294342   1.98082 11  1 spearman
```

Interestingly, when we do a partial correlation using Spearman's rho, controlling for Gender, we now have the relationship between TV time and happiness down from .65 to .57, and more significantly- the finding is no longer statistically significant! $p>0.05$
This suggests that the relationship between TV time and happiness is mediated by Gender (I think).