assignment\_05\_KalyanKalai

Kalaivani Kalyanasundaram

2021-07-25

# 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.

## Github link

<https://github.com/kalaikalyan/hello-world>

## 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.

setwd("/Users/kalya/Documents/Kalai/dsc520")  
students\_df <- read.csv("data/student-survey.csv")  
cov(students\_df$TimeTV,students\_df$TimeReading)

## [1] -20.36364

cov(students\_df$Happiness,students\_df$Gender)

## [1] 1.116636

I would use the covariance between TimeTV and TimeReading to see if there is a correlation/relationship between watching tv and reading time as the question prompts. Also, looking at gender and happiness, I can check if there are any other factors in play here in this question

## 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.

* TimeTV(minutes)
* TimeReading(hours)
* Happiness(Percentage)
* Gender (Binary) Changing the measurment will change the covariance value but it will still not change the strength of the relationship so as a whole, the relationship will stay the same but it wil change the covariance value. I do think a better approach is using similar units for both like minutes or hours.

## 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 will be doing a Pearson’s correlation test. I believe this test wil show that there is some sort of relationhip between TimeReading and TimeTV. I predict that there is a negative correlation because people would rather watch TV than read.

## Perform a correlation analysis of:

### All variables

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

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

setwd("/Users/kalya/Documents/Kalai/dsc520")  
students\_df <- read.csv("data/student-survey.csv")  
rcorr(as.matrix(students\_df))

## TimeReading TimeTV Happiness Gender  
## TimeReading 1.00 -0.88 -0.43 -0.09  
## TimeTV -0.88 1.00 0.64 0.01  
## Happiness -0.43 0.64 1.00 0.16  
## Gender -0.09 0.01 0.16 1.00  
##   
## n= 11   
##   
##   
## P  
## TimeReading TimeTV Happiness Gender  
## TimeReading 0.0003 0.1813 0.7932  
## TimeTV 0.0003 0.0352 0.9846  
## Happiness 0.1813 0.0352 0.6448  
## Gender 0.7932 0.9846 0.6448

### A single correlation between two a pair of the variables

library(Hmisc)  
setwd("/Users/kalya/Documents/Kalai/dsc520")  
students\_df <- read.csv("data/student-survey.csv")  
cor.test(students\_df$TimeReading,students\_df$TimeTV)

##   
## Pearson's product-moment correlation  
##   
## data: students\_df$TimeReading and students\_df$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

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

library(Hmisc)  
setwd("/Users/kalya/Documents/Kalai/dsc520")  
students\_df <- read.csv("data/student-survey.csv")  
cor.test(students\_df$TimeReading,students\_df$TimeTV, conf.level = .99)

##   
## Pearson's product-moment correlation  
##   
## data: students\_df$TimeReading and students\_df$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

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

Based on the different confidence levels, I am 99% confident that there is a negative correlation between time spent watching Tv and time spent reading because the cor value is negative and the confidence intervals are all negative (below 0). There seems to be a negative relationship between TimeTV and TimeReading.

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

setwd("/Users/kalya/Documents/Kalai/dsc520")  
students\_df <- read.csv("data/student-survey.csv")  
cor(students\_df$TimeReading,students\_df$TimeTV)

## [1] -0.8830677

lm <- lm(students\_df)  
summary(lm)$r.squared

## [1] 0.8211648

Based on the negative correlation coefficient and the coefficient of determinination, I conclude there is a negative correlation between TV use and reading. The coefficient of determination vale proves that around 80% of the data fit the regression model which is a strong fit.

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

Athough there is a negative correlation, you cannot say that is the cause because correlation doesn’t mean causation.

library(ppcor)

## Loading required package: MASS

setwd("/Users/kalya/Documents/Kalai/dsc520")  
students\_df <- read.csv("data/student-survey.csv")  
pcor(students\_df)

## $estimate  
## TimeReading TimeTV Happiness Gender  
## TimeReading 1.0000000 -0.8827973 0.4013124 -0.2706036  
## TimeTV -0.8827973 1.0000000 0.6311611 -0.2943135  
## Happiness 0.4013124 0.6311611 1.0000000 0.2833152  
## Gender -0.2706036 -0.2943135 0.2833152 1.0000000  
##   
## $p.value  
## TimeReading TimeTV Happiness Gender  
## TimeReading 0.000000000 0.001615344 0.28437887 0.4812716  
## TimeTV 0.001615344 0.000000000 0.06832112 0.4420392  
## Happiness 0.284378868 0.068321119 0.00000000 0.4600603  
## Gender 0.481271572 0.442039185 0.46006033 0.0000000  
##   
## $statistic  
## TimeReading TimeTV Happiness Gender  
## TimeReading 0.0000000 -4.9720962 1.1592148 -0.7436966  
## TimeTV -4.9720962 0.0000000 2.1528933 -0.8147673  
## Happiness 1.1592148 2.1528933 0.0000000 0.7816064  
## Gender -0.7436966 -0.8147673 0.7816064 0.0000000  
##   
## $n  
## [1] 11  
##   
## $gp  
## [1] 2  
##   
## $method  
## [1] "pearson"

This shows another side of the data, where the correlation doesn’t necessarily have to be negative. There are certainly other factors in play so we can reiterate that correlation does not result in causation.