Week 7-Student Survey

Aarti Ramani

January 27, 2023

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Masters/GitHub/Winter2022/Ramani-DSC520")

## Load the `data/r4ds/heights.csv` to
student_survey <- read.csv("data/student-survey.csv")
nrow(student_survey)
```

```
## [1] 11
```

```
## i. 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.
## TimeReading vs. TimeTV
cor(student_survey$TimeReading,student_survey$TimeTV, method="pearson")
```

```
## [1] -0.8830677
```

```
# -0.8830677 -> There is a negative correlation between the time spent reading and watching TV

## TimeReading vs. Happiness
cor(student_survey$TimeReading,student_survey$Happiness, method="pearson")
```

```
## [1] -0.4348663
```

```
# -0.4348663 -> Time spent reading and Happiness are also negatively correlated

## TimeReading vs. Gender
cor(student_survey$TimeReading,student_survey$Gender, method="pearson")
```

```
## [1] -0.08964215
```

```
# -0.08964215 -> Time spent reading and Gender are also negatively correlated.
# The value is close to 0.

## TimeTV vs. Happiness
cor(student_survey$TimeTV,student_survey$Happiness, method="pearson")
```

```
## [1] 0.636556
```

```r
# 0.636556 -> There is a positive correlation between the time spent watching TV and happiness

## TimeTV vs. Gender
cor(student_survey$TimeTV,student_survey$Gender, method="pearson")
```

```
## [1] 0.006596673
```

```r
# 0.006596673 -> Time spent watching TV and Gender are also positively correlated

## Happiness vs. Gender
cor(student_survey$Happiness,student_survey$Gender, method="pearson")
```

```
## [1] 0.1570118
```

```r
# 0.1570118 -> Happiness and Gender are positively correlated
```

```r
## ii. Examine the Survey data variables.
##What measurement is being used for the variables?
#ANSWER:
#TimeReading - In hours
#TimeTV - In minutes
#Happiness - Numeric 0-100 (0 being lowest and 100 being highest)
#Gender - Binary 0 and 1

##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.
#ANSWER:
#All variables except Gender are numeric. I would like to change the gender to a string.
#In the current times gender is represented in different forms such as queer, male, female, fluid etc.
#I would avoid computing Covariance between Gender and the other variables
#since its not relevant to the research question.
student_survey <- student_survey[,c("TimeReading", "TimeTV", "Happiness")]
```

```r
## iii. 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?
cor.test(student_survey$TimeReading,student_survey$TimeTV, method="pearson")
```

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

```
#ANSWER:
#I choose to calculate correlation between Time spent Reading and Time spent watching TV
#with One-Sample T-Test since they are both in measures of time.
#The variables have a negative correlation and the value is closer to -1 implying both are inversely re
#The more time a person spends in watching TV, the lesser time spent in reading and vice-versa.
```

```
## iv. Perform a correlation analysis of:
##1. All variables
cor(student_survey, use="complete.obs", method = "pearson")
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

```
#cor(student_survey, use="everything")

##2. A single correlation between two a pair of the variables
cor(student_survey$TimeReading, student_survey$Happiness, method = "pearson")
```

```
## [1] -0.4348663
```

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

## TimeReading vs. TimeTV
cor.test(student_survey$TimeReading,student_survey$TimeTV, method="pearson", conf.level = .99)
```

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

```
#ANSWER:
# cor is -0.883 which is closer to a -1 implying a strong negative correlation.
#  - If 0, implies there is no relation between the variables.
# Negative correlation -> variables are inversely related,
#                         if one variable goes up the other goes down.
# As the time spent reading goes up less time is spent watching TV.

## TimeReading vs. Happiness
cor.test(student_survey$TimeReading,student_survey$Happiness, method="pearson", conf.level = .99)
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  student_survey$TimeReading and student_survey$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8801821  0.4176242
## sample estimates:
##        cor
## -0.4348663
```

```
#ANSWER:
# cor is -0.4348663 which is negative but also closer to 0 implying a negative correlation.
#  - If 0, implies there is no relation between the variables.
# Negative correlation -> variables are inversely related,
#                         if one variable goes up the other goes down.
# As the time spent reading goes up lesser happy the student is.

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

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey$TimeTV and student_survey$Happiness
## 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
```

```
#ANSWER:
# cor is 0.636556 which is positive and closer to 1 => positive correlation.
#  - If 0, implies there is no relation between the variables.
# Positive correlation -> variables are directly related,
#                         if one variable goes up the other goes up too.
# As the time spent watching TV goes up more happier the student is.


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

```
#cor(student_survey, use="complete.obs", method = "pearson")
student_survey_matrix <-as.matrix(student_survey)
rcorr(student_survey_matrix)
```

```
##             TimeReading TimeTV Happiness
## TimeReading        1.00  -0.88     -0.43
## TimeTV            -0.88   1.00      0.64
## Happiness         -0.43   0.64      1.00
##
## n= 11
##
##
## P
##             TimeReading TimeTV Happiness
## TimeReading             0.0003 0.1813
## TimeTV      0.0003             0.0352
## Happiness   0.1813      0.0352
```

```
#ANSWER:
# The correlation matrix explains the relationship between variables and a significance value
# TimeReading and TimeTV
#We can gain confidence that there is a strong relationship between TimeReading and TimeTV.
#The variables are negatively correlated.

# TimeReading and Happiness
#We can gain confidence that there is some relationship between TimeReading and Happiness.
#The variables are negatively correlated.

#TimeTV and Happiness
#We can gain confidence that there is a strong relationship between TimeTV and Happiness.
#The variables are positively correlated.
```

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

#student_survey <- student_survey[, c("TimeReading", "TimeTV", "Happiness")]
cor(student_survey)
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

```r
cor(student_survey)^2 * 100
```

```
##             TimeReading    TimeTV Happiness
## TimeReading   100.00000  77.98085  18.91087
## TimeTV         77.98085 100.00000  40.52035
## Happiness      18.91087  40.52035 100.00000
```

```r
# ANSWER:
#variability in time spent reading on time spent watching tv is 77.98%.
#The variables have a inverse correlation

#variability in time spent reading on happiness is 18.91%
#The variables have a moderate correlation

#variability in time spent watching TV on happiness is 40.52%
#The variables have a strong correlation
```

```r
## vi. Based on your analysis can you say that watching more TV caused students
## to read less? Explain.
#ANSWER: Yes, the more time a student spent watching TV lesser time they spent reading
```

```r
## vii. 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(ggm)
```

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

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

```r
pcor(c("TimeReading", "TimeTV", "Happiness"), var(student_survey))
```

```
## [1] -0.872945
```

```r
(pcor(c("TimeReading", "TimeTV", "Happiness"), var(student_survey)))^2*100
```

```
## [1] 76.2033
```

```r
#ANSWER:
#The partial correlation has one control variable - Happiness.
#Time spent reading or watching TV controls the level of happiness
#making happiness the controlled variable.
#The percentage variability of readings vs TV still remains almost the same,
#with or without the controlled variable.
```