# Assignment: ASSIGNMENT 5

# Name: Parker, Breanna

# Date: 1-29-2023

## Set the working directory to the root of your DSC 520 directory

setwd("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1")

## Load the data/r4ds/heights.csv to

heights\_df <- read.csv("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1/data/r4ds/heights.csv")

## Using cor() compute correclation coefficients for

## height vs. earn

library(ggplot2)  
cor(heights\_df$height, heights\_df$earn)

## [1] 0.2418481

### age vs. earn

cor(heights\_df$age, heights\_df$earn)

## [1] 0.08100297

### ed vs. earn

cor(heights\_df$ed, heights\_df$earn)

## [1] 0.3399765

## Spurious correlation

## The following is data on US spending on science, space, and technology in millions of today’s dollars

## and Suicides by hanging strangulation and suffocation for the years 1999 to 2009

## Compute the correlation between these variables

tech\_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)  
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)  
  
cor(tech\_spending, suicides)

## [1] 0.9920817

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.

survey\_data <- read.csv("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1/data/student-survey.csv")

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.

library(corpcor)   
library(readr)  
library(ggpubr)  
library(MASS)  
library(ppcor)  
  
cov(survey\_data$TimeReading,survey\_data$TimeTV)

## [1] -20.36364

cov(survey\_data$TimeReading,survey\_data$Happiness)

## [1] -10.35009

cov(survey\_data$TimeReading,survey\_data$Gender)

## [1] -0.08181818

cov(survey\_data$Happiness,survey\_data$TimeTV)

## [1] 114.3773

cov(survey\_data$Gender,survey\_data$TimeTV)

## [1] 0.04545455

cov(survey\_data$Gender,survey\_data$Happiness)

## [1] 1.116636

#covariance is when you're trying to see if when one variable deviates from a mean, that the other variable follows. The positive values mean that they deviate from the mean in the same direction, while negative is opposing. So, what the data is showing is that the more time someone reads, the less time that they watch TV and vice versa. It also is showing that the more time you spend reading, the less happy that you are. Gender isn't really a numerical value and the survey doesn't explain if 0 or 1 is supposed to be male or female. If one is female than I believe the data is showing that females are happier than males and that females watch more tv than males.

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.

#It looks like Time reading is in hours per week while tv is in minutes possibly per day? or maybe per week. This would definitely effect the numbers. Time watching tv and reading should both be in minutes per week to make sure that the data is being assessed correctly. Happiness is based on a percentage and gender is a binary code. Becasue of this you can’t really calculate covariance too well since there are only two options and they’re not supposed to be numerical. Also, If we should include nonbinary as an option which would complicate using the covariance even more so I don’t think it should be used for gender. What you could do is simply split the groups up and find out the average mean of each one for tv, reading, and happiness so that you can compare the groups that way. Happiness being based as a percentage is fine but I think the numbers should be rounded. I’m not sure how you can really say that someone’s happiness is 15.22 rather than just 15? Seems irrelevant.

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(survey\_data$TimeReading,survey\_data$Happiness, method=c("kendall"))

## [1] -0.2889428

cor(survey\_data$TimeTV,survey\_data$Happiness, method=c("kendall"))

## [1] 0.4630424

#this matches the covariance correlation in that the happiness and timeTV have a positive correlation while the happiness and time reading has a negative one. I used kendall because it does not require the variables to be within a bell curve and it measures the likelihood that two variabels will move in the same direction (even if the rates of the direction are different)

Perform a correlation analysis of: All variables

cor(survey\_data$TimeReading,survey\_data$TimeTV)

## [1] -0.8830677

cor(survey\_data$TimeReading,survey\_data$Happiness)

## [1] -0.4348663

cor(survey\_data$TimeReading,survey\_data$Gender)

## [1] -0.08964215

cor(survey\_data$Happiness,survey\_data$TimeTV)

## [1] 0.636556

cor(survey\_data$Gender,survey\_data$TimeTV)

## [1] 0.006596673

cor(survey\_data$Gender,survey\_data$Happiness)

## [1] 0.1570118

A single correlation between two a pair of the variables 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.

cor.test(survey\_data$TimeReading,survey\_data$Happiness, method=c("kendall"), conf.level=0.99)

## Warning in cor.test.default(survey\_data$TimeReading, survey\_data$Happiness, :  
## Cannot compute exact p-value with ties

##   
## Kendall's rank correlation tau  
##   
## data: survey\_data$TimeReading and survey\_data$Happiness  
## z = -1.1921, p-value = 0.2332  
## alternative hypothesis: true tau is not equal to 0  
## sample estimates:  
## tau   
## -0.2889428

cor.test(survey\_data$TimeTV,survey\_data$Happiness, method=c("kendall"), conf.level = 0.99)

## Warning in cor.test.default(survey\_data$TimeTV, survey\_data$Happiness, method =  
## c("kendall"), : Cannot compute exact p-value with ties

##   
## Kendall's rank correlation tau  
##   
## data: survey\_data$TimeTV and survey\_data$Happiness  
## z = 1.9582, p-value = 0.05021  
## alternative hypothesis: true tau is not equal to 0  
## sample estimates:  
## tau   
## 0.4630424

#the z for time reading vs happiness is negative and it represents the deviations from the mean. since the p value is greater tha 0.05 in both cases (though just barely in the second case), means that there is a statistical correlation between the two variables, but that doesn’t mean that they are significant.

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

cor(survey\_data$TimeReading,survey\_data$Happiness)

## [1] -0.4348663

cor(survey\_data$TimeTV,survey\_data$Happiness)

## [1] 0.636556

cor(survey\_data$TimeReading,survey\_data$Happiness)^2

## [1] 0.1891087

cor(survey\_data$TimeTV,survey\_data$Happiness)^2

## [1] 0.4052035

#this determines how strong the two variables are related to each other. Since it is closer to 0 than 1, it means that the variables are not closely related to each other. Though maybe tv has a bigger impact on happiness than reading does.

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

#It appears that way in the data. However, this could be caused because you only have so many hours in the week and those who read, it’s a hobby that you can’t do while also watching tv, so it cuts into tv time. However, if someone’s hobby is to crochet. They can easily do that while watching tv so it doesn’t cut into that time.

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.

cor(survey\_data$TimeTV, survey\_data$TimeReading)

## [1] -0.8830677

pcor.test(survey\_data$TimeTV, survey\_data$TimeReading, survey\_data$Happiness,)

## estimate p.value statistic n gp Method  
## 1 -0.872945 0.0009753126 -5.061434 11 1 pearson

#these results ended up being the same, whether we control for happiness or not, meaning that they are not connected.