ASSIGNMENT 5

Naveen Bagam

2022-01-30

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

## setwd(“/Users/nbaga/dsc520”)

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

heights\_df <- read.csv("/Users/nbaga/dsc520/data/r4ds/heights.csv")  
  
## Using `cor()` compute correclation coefficients for  
## height vs. earn  
cor(heights\_df$height, heights\_df$earn, method = "pearson")

## [1] 0.2418481

### age vs. earn  
cor(heights\_df$age, heights\_df$earn, method = "pearson")

## [1] 0.08100297

### ed vs. earn  
cor(heights\_df$ed, heights\_df$earn, method = "pearson")

## [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   
  
##The survey data is located in this StudentSurvey.csv file.  
  
## 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.   
  
student\_survey <- read.csv("/Users/nbaga/dsc520/data/student-survey.csv")  
  
 ## 1) Time Reading vs Happiness:-   
   
cov(as.numeric(student\_survey$TimeReading),as.numeric(student\_survey$Happiness), method = "pearson")

## [1] -10.35009

## Expalantion of result:-  
   
 ## The Timereading and Happiness negatively related, in other words , when the Time of reading increases the Happiness is decreased .  
  
## 2) Time TV vs Happiness   
 cov(as.numeric(student\_survey$TimeTV),as.numeric(student\_survey$Happiness), method = "pearson")

## [1] 114.3773

##Explanation :-  
   
 ##The TimeTV and Happiness positively ralted, when TV time increased the   
 ## happiness is increased.  
   
## 3) Time TV vs TimeReading   
 cov(as.numeric(student\_survey$TimeTV),as.numeric(student\_survey$TimeReading), method = "pearson")

## [1] -20.36364

## EXplanation:-   
   
 ## TimeTV and TimeReading are negatively related, when ever people watch   
 ## more tv theor reading habbit is reduce.   
   
   
 ## II) 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.  
 ## Explanation  
   
 ## The measurement is used for the 3 variables Time TV in minutes and TimeReading in hours and the Happiness is in the Percentage measure(%) , if we can keep all our variables in same measure the results would be correct.  
   
## 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?  
   
 ##EXPLANATION  
   
 ##relationship between two variables is to use the Pearson correlation coefficient, which is a measure of the linear association between two variables. It always takes on a value between -1 and 1 where:  
  
## -1 indicates a perfectly negative linear correlation between two variables  
## 0 indicates no linear correlation between two variables  
## 1 indicates a perfectly positive linear correlation between two variables  
   
##Perform a correlation analysis of:  
 ## 1)All variables  
 cor(student\_survey)

## TimeReading TimeTV Happiness Gender  
## TimeReading 1.00000000 -0.883067681 -0.4348663 -0.089642146  
## TimeTV -0.88306768 1.000000000 0.6365560 0.006596673  
## Happiness -0.43486633 0.636555986 1.0000000 0.157011838  
## Gender -0.08964215 0.006596673 0.1570118 1.000000000

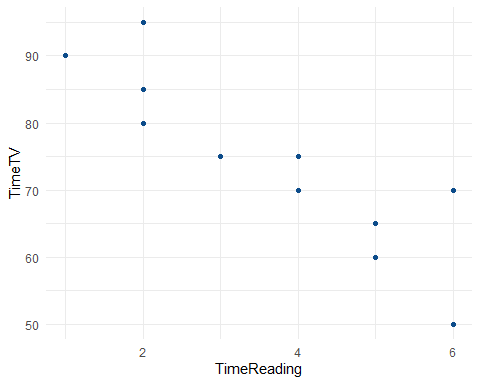
## 2) A single correlation between two a pair of the variables  
 cor(student\_survey$TimeReading,student\_survey$TimeTV)

## [1] -0.8830677

## 3) Repeat your correlation test in step 2 but set the confidence interval at 99%   
 cor.test(student\_survey$TimeReading,student\_survey$TimeTV,conf.level=0.99)

##   
## 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

## 4) Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.  
 ## Explanation:-   
 ## Based on the calculation found that the 99 percent confidence   
 ## interval for the variables taken are : ## -0.9801052 -0.4453124  
 ## and the T-Value is t = -5.6457 and the p-value = 0.0003153 and the  
 ## sample correlation is -0.8830677 whcih means the variables are  
 ## negitively related  
 ## variable increases the other one decreases and if one decreases the other one increase as well.  
   
 ## 5) Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.  
library(ggplot2)  
 ggplot(student\_survey) + aes(x = TimeReading, y = TimeTV) + geom\_point(colour = "#0c4c8a") +  
 theme\_minimal()



## 6) Based on your analysis can you say that watching more TV caused students to read less? Explain.  
 ##Explanation  
   
 ##Yes, Based on the analysis we can conclude that watching TV more is causing less in reading.  
   
 ## 7) 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(ppcor)

## Loading required package: MASS

pcor(student\_survey)

## $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"

## From the above analysis , we can conclude that the if one variable is going up another variabvke is going dwon , when the TC time incerses the reading time is reduced.