ASSIGNMENT 7.2

Sharath Chandra Tummanapally

2021-10-12

# 1.

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

## Using cor() compute correclation coefficients for

### height vs. earn

## [1] 0.2418481

### age vs. earn

## [1] 0.08100297

### ed vs. 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

## [1] 0.9920817

# 2.

## 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.

## 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 TimeTV Happiness Gender  
## TimeReading 3.05454545 -20.36363636 -10.350091 -0.08181818  
## TimeTV -20.36363636 174.09090909 114.377273 0.04545455  
## Happiness -10.35009091 114.37727273 185.451422 1.11663636  
## Gender -0.08181818 0.04545455 1.116636 0.27272727

### Covariance is a statistical term used to measures the direction of the linear relationship between the data vectors.We are using cov() function to test the relationship with each survey variable. Higher positive numbers meaning a closer relationship between the two variables, lower negative numbers meaning an inverse relationship and numbers near zero meaning no relationship.

### Results observations:

### According to the results, TimeReading vs TimeTV, TimeReading vs Happiness, TimeReading vs Gender have negative values which means they have inverse relationship. TimeTv vs Happiness, Happiness vs Gender, Gender vs TimeTv, Gender vs Happiness have positive values which means they have closer relationship. TimeTv vs Gender is near to zero which indicates no relation.

## 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.

### Measurements that are used: TimeReading - Ratio, TimeTv - Ratio, Happiness - Ordinal, Gender - Nominal. I believe that Gender since its measurement level Nominal, the impact on the covriance calcuation would be significant if its changed to ratio. Other than that, we can see the covariance calculation will not significantly change if we change the measurements of remaining variables.

## 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?

### We are going to use the Spearman correlation as it can be used to evaluate a monotonic relationship between two variables — Continous or Ordinal and it is based on the ranked values for each variable rather than the raw data. In this dataset, to perform this test we could choose TimeTv and Happiness data variables, it yields a positive correlation.

## Perform a correlation analysis of:

## All variables

## 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

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

## [1] -0.8830677

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

##   
## Pearson's product-moment correlation  
##   
## data: student\_survey\_df$TimeReading and student\_survey\_df$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 observation of the above correlation matrix, TimeReading (Time spent on reading by the students) is inversely propotional to TimeTv (Time spend on watching Tv by the students) since the Correlation Coefficient is negative and close to -1. Further more TimeTv and Happiness data variables, it yields a near positive correlation based on the matrix in which the coefficeint is almost near to 1.

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

## [1] -0.8830677

## [1] 0.7798085

### The negative sign of r tells us that the relationship is negative — as TimeTv increases, TimeReading decreases — as we expected. Because r is fairly close to -1, it tells us that the linear relationship is fairly strong, but not perfect. The r squared value tells us that 77.9 % of the variation in the time of reading is reduced by taking into account the time of watching Tv.

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

### Based on the analysis we found the TimeReading is decreased when TimeTV is increased. We calculated Correlation coefficient. It shows the negative value and its closed to -1. It means that those variables are in negative correlation. So, we can say that watching more TV caused to read less.

## 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

## $estimate  
## TimeReading TimeTV Happiness  
## TimeReading 1.0000000 -0.8729450 0.3516355  
## TimeTV -0.8729450 1.0000000 0.5976513  
## Happiness 0.3516355 0.5976513 1.0000000  
##   
## $p.value  
## TimeReading TimeTV Happiness  
## TimeReading 0.0000000000 0.0009753126 0.31905895  
## TimeTV 0.0009753126 0.0000000000 0.06804372  
## Happiness 0.3190589526 0.0680437248 0.00000000  
##   
## $statistic  
## TimeReading TimeTV Happiness  
## TimeReading 0.000000 -5.061434 1.062425  
## TimeTV -5.061434 0.000000 2.108388  
## Happiness 1.062425 2.108388 0.000000  
##   
## $n  
## [1] 11  
##   
## $gp  
## [1] 1  
##   
## $method  
## [1] "pearson"

### Partial correlation is run when we want to find the relation between two variables while controlling the third variable.

### Here the two variables are TimeReading and TimeTv, with the third variable - Happiness

### The partial correlation between TimeReading and TimeTv is - 0.8729450, which is almost a negative correlation. As TimeReading increases, TimeTv tends to decrease while the Happiness variable is controlled.

### The p-value for this partial correlation is 0.0009753126, which is not statistically significant at α = 0.05.