assignment_07_KanaparthiVenkata

Venkata Kanaparthi

5/1/2021

We will be utilizing GitHub for some of the exercises in this course. This is to give you more experience with interacting with this tool that will be used in most courses during the program.

The GitHub repository that will be used is: https://github.com/bellevue-university/dsc520

1.) Complete assignment05

```
library(ggm)
```

```
## Warning: package 'ggm' was built under R version 4.0.5
```

```
library(ggplot2)
```

Using `cor()` compute correlation coefficients for

1.) height vs. earn

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("data/r4ds/heights.csv")
cor(heights_df$height, heights_df$earn)
```

```
## [1] 0.2418481
```

2.) age vs. earn

```
cor(heights_df$age, heights_df$earn)
```

```
## [1] 0.08100297
```

3.) 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.

Q1:

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. Variables Considered: TimeReading, TimeTV

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")
cov(studSurvey_df$TimeReading, studSurvey_df$TimeTV)
```

```
## [1] -20.36364
```

Ans:

Variables Considered: TimeReading, TimeTV

Covariance measures the directional relationship between the returns on two variables. A positive covariance means that asset returns move together while a negative covariance means they move inversely.

The covariance of the variables TimeReading and TimeTV from Student Survey is negative which indicates an inverse relationship. i.e; one variable deviates from mean(increases) and the other deviates from the mean in the opposite direction(decreases)

Q2:

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.

Variables Considered: TimeReading, TimeTV

Ans:

Looks like Minutes are used to TimeTV and Hours for TimeReading After changing the measurement units to minutes for TimeReading, there is a significant increase in covariance. It would have been a problem if there is a change to the covariance in the other direction.

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey_new.csv")
cov(studSurvey_df$TimeReading, studSurvey_df$TimeTV)
```

```
## [1] -1221.818
```

Q3:

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?

Variables Considered: TimeReading, TimeTV

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")
cor(studSurvey_df$TimeReading, studSurvey_df$TimeTV, method="spearman")
```

```
## [1] -0.9072536
```

```
cor.test(studSurvey_df$TimeReading, studSurvey_df$TimeTV, method="spearman")
```

```
## Warning in cor.test.default(studSurvey_df$TimeReading, studSurvey_df$TimeTV, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  studSurvey_df$TimeReading and studSurvey_df$TimeTV
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9072536
```

Ans:

Will choose a spearmans correlation because of the non parametric nature. The correlation between two variables is very large. as the significance value for this correlation coefficient is less than 0.05, it can be concluded that there is a significant relationship between these variables. Time spent in reading increases when there is a decrease in time spent on TV

Q4:

Perform a correlation analysis of:

1.) All variables

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")
cor(studSurvey_df)
```

```
##                 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
## TimeReadingInMin  1.00000000 -0.883067681 -0.4348663 -0.089642146
##                 TimeReadingInMin
## TimeReading           1.00000000
## TimeTV               -0.88306768
## Happiness            -0.43486633
## Gender               -0.08964215
## TimeReadingInMin      1.00000000
```

2.) A single correlation between two a pair of the variables

Variables Considered: TimeReading, TimeTV

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")
cor(studSurvey_df$TimeReading, studSurvey_df$TimeTV, use="complete.obs", method = "spearman")
```

```
## [1] -0.9072536
```

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

Variables Considered: TimeReading, TimeTV

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")
cor.test(studSurvey_df$TimeReading, studSurvey_df$TimeTV, use="complete.obs", method = "spearman", conf
```

```
## Warning in cor.test.default(studSurvey_df$TimeReading, studSurvey_df$TimeTV, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  studSurvey_df$TimeReading and studSurvey_df$TimeTV
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9072536
```

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

Variables Considered: TimeReading, TimeTV

Ans:

Correlation is negative which indicates that both the variables are inversely related

As the confidence interval does not cross 0, it indicates us the value of correlation is negative, so we can confidently say that time spent on tv and reading are negatively related

Q5:

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

Variables Considered: TimeReading, TimeTV

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")

# correlation coefficient
correCoeff <- cor(studSurvey_df$TimeReading, studSurvey_df$TimeTV, use="complete.obs", method = "spearma

# coefficient of determination
coffDet <- cor(studSurvey_df$TimeReading, studSurvey_df$TimeTV, use="complete.obs", method = "spearman"

correCoeff
```

```
## [1] -0.9072536
```

```
coffDet
```

```
## [1] 0.8231091
```

```
coffDetinPercnt <- coffDet*100
coffDetinPercnt
```

```
## [1] 82.31091
```

Ans:

After converting the Coeff of Determination into percent, it indicates that 77% of time reading is dependent on time spent on TV

Q6:

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

Ans:

Yes, watching more TV definitely reduces the time to read. Because of the negative correlation that we had on these two variables

Q7:

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.

Variables Considered: TimeReading, Happiness, TimeTV (Controlling Variable)

```
setwd('E:/MSDS-SEM2/DSC520/CodingAssignments/DSC520KANAPARTHI')
studSurvey_df <- read.csv("data/student-survey.csv")
# correlation coefficient
correCoeff <- cor(studSurvey_df$TimeReading, studSurvey_df$Happiness, use="complete.obs", method = "spea
correCoeff
```

```
## [1] -0.4065196
```

```
partCorr <- pcor(c("TimeReading","Happiness","TimeTV"),var(studSurvey_df))
partCorr
```

```
## [1] 0.3516355
```

Ans:

Correlation Coefficient is coming as negative when comparing TimeReading and Happiness which indicates the inverse relation between these two variables.

When a controlling variable of TimeTV is brought into picture it completely changed the perception to a positive correlation which says that the relation between these two variables is direct