assignment_07_PonisserilRadhakrishnan

Radhakrishnan Ponisseril

7/24/2021

```
library(ggm)
library(ggplot2)
```

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.

1:

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.

```
setwd("/Users/RajeevP/dsc520")
studSurvey_df <- read.csv("data/student-survey.csv")
cov(studSurvey_df$TimeReading, studSurvey_df$TimeTV)
```

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

Covariance measures the directional relationship between the returns on two variables. The covariance of the variables TimeReading and TimeTV from Student Survey is negative which indicates an inverse relationship.

2:

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: TimeReading, TimeTV Minutes are used for TimeTV and Hours for TimeReading. After changing the measurement units to minutes for TimeReading, there is a significant increase in covariance.

```
setwd("/Users/RajeevP/dsc520")
studSurvey_df <- read.csv("data/student-survey-updated.csv")
cov(studSurvey_df$TimeReading, studSurvey_df$TimeTV)
```

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

3:

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?

```
setwd("/Users/RajeevP/dsc520")
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
```

I will choose spearmans correlation because of the non parametric nature.

4:

Perform a correlation analysis of:

All variables

```
setwd("/Users/RajeevP/dsc520")
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
```

A single correlation between two a pair of the variables

```
setwd("/Users/RajeevP/dsc520")
studSurvey_df <- read.csv("data/student-survey.csv")
cor(studSurvey_df$TimeReading, studSurvey_df$TimeTV, use="complete.obs", method = "spearman")
```

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

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

```
setwd("/Users/RajeevP/dsc520")
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
```

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

Correlation is negative which indicates that both the variables are inversely related. As the confidence interval did not cross zero, it indicates that the value of correlation is negative. Time spent on tv and reading are negatively related

5:

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

```
setwd("/Users/RajeevP/dsc520")
studSurvey_df <- read.csv("data/student-survey.csv")

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

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

correCoef
```

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

```
coffDtn
```

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

```
coffDtninPercnt <- coffDtn*100
coffDtninPercnt
```

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

77% of time reading is dependent on time spent on TV

6:

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

Yes, watching more TV reduces the time to read because of the negative correlation between these variables

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.

Variables: TimeReading, Happiness Controlling Variable: TimeTV

```
setwd("/Users/RajeevP/dsc520")
studSurvey_df <- read.csv("data/student-survey.csv")
# correlation coefficient
correCoef <- cor(studSurvey_df$TimeReading, studSurvey_df$Happiness, use="complete.obs", method = "spea
correCoef
```

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

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

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

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

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