RMarkdown Assignment Week 7

Dipika Sharma

May 2nd 2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Add Citations

- R for Everyone (Lander 2014)
- Discovering Statistics Using R (Field, Miles, and Field 2012)

Assignment 05

Using `cor()` compute correlation coefficients for

1. height vs. earn

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

2. age vs. earn

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

3. ed vs. earn

```
## [1] 0.3399765
```

4. 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
```

Student Survey

Use R to calculate the covariance of the Survey variables

```
## [1] 0.636556
```

```
## [1] -0.4348663
```

Provide an explanation of why you would use this calculation and what the results indicate.

Answer: I use cor function to see the relationship between the Time reading with Happiness and Time TV with Happiness. Using Cor function I found that Time TV and Happiness is giving us positive correlation where as Time reading and Happiness is negative correlation which mean more time students spent in watching TV their happiness increases but if student spent more time in reading their happiness decreases.

Examine the Survey data variables. What measurement is being used for the variables?

Answer: We have four variables in student survey data.

```
## 'data.frame':    11 obs. of  4 variables:
##  $ TimeReading: int  1 2 2 2 3 4 4 5 5 6 ...
##  $ TimeTV     : int  90 95 85 80 75 70 75 60 65 50 ...
##  $ Happiness  : num  86.2 88.7 70.2 61.3 89.5 ...
##  $ Gender     : int  1 0 0 1 1 1 0 1 0 0 ...
```

```
## TimeReading      TimeTV   Happiness      Gender
##   "integer"   "integer"   "numeric"   "integer"
```

Using str and sapply functions we can clearly see TimeReading, TimeTV and Gender variables are integer where as Happiness is numeric variable.

TimeReading, TimeTV are interval variables. Gender is nominal variable with value 0 or 1 and Happiness is Ratio variable.

Explain what effect changing the measurement being used for the variables would have on the covariance calculation.

Answer: Depending on the changes it can have significant effect on the covariance or it might have slight changes. Covariance is used to find out the relationship of of 2 variables or we can say finding out the dependency of one variable on other. If we will make any change to any varibale it will have some effect on covariance.

Would this be a problem? Explain and provide a better alternative if needed.

Answer: I would use nominal variables for Gender - Male and Female. currently we using the numeric value 0 and 1 which is not very clear in this way we can see different relationship between other variables for specific Gender It would be interesting to see if that will change the overall relationship between the variables.

Choose the type of correlation test to perform.

```
## [1] 0.636556
```

explain why you chose this test?

Answer: I would like to see relationship between time spent on TV and happiness.

Make a prediction if the test yields a positive or negative correlation?

Answer: It is positive correlation which indicate that if student spent more time watching TV their happiness also increase.

Perform a correlation analysis of:

1. 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
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  ssurvey_df$TimeTV and ssurvey_df$TimeReading
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9694145 -0.6021920
## sample estimates:
##        cor
## -0.8830677
```

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

```
##
##  Pearson's product-moment correlation
##
## data:  ssurvey_df$TimeTV and ssurvey_df$TimeReading
## 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.

Answer: After looking at the correlation matrix we can clearly see that TimeReading and TimeTV show negative correlation which mean with increase of one variable the second variable will decrease so if student spent more time in watching TV they will spent less time in reading. Hence variables are opposite. If we look at TimeTV and Happiness we see positive correlation, we will see student are more happy, if they spent more time watching TV so if one variable increase the other variable will also increase. Time reading and Happiness is negative correlation which mean if student spent more time in reading their happiness decrease.and lastly all the gender are showing negative correlation with Time reading where as gender are showing positive correlation wit TimeTV and Happiness.

Calculate the correlation coefficient and the coefficient of determination,

Correlation coefficient

```
## [1] -0.8830677
```

Coefficient of determination

```
## [1] 0.7798085
```

Describe what you conclude about the results.

Answer: Correlation define the strength of the relationship between an independent and dependent variable and coefficient of determination tell us to what extent the variance of one variable explains the variance of the second variable. In our case coefficient of determination is .77 then approximately 70% of the observed variation can be explained by the inputs.

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

Answer:

```
## [1] -0.8830677
```

Using the Cor function on TimeTV and Time Reading we can cleary see bot variable have negative correlation which stat that if student spent more time in TV then they will spent less time reading.

Pick three variables and perform a partial correlation

```
## Loading required package: MASS
```

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

Documenting which variable you are "controlling."

Answer: Happiness is the variable we are controlling.

Explain how this changes your interpretation and explanation of the results.

Answer: We already know that TimeTV and TimeReading are negative correlation that is if one variable increases second variable will decrease. After calculating partial correlation between the two where happiness is the controlling variable. p value is 0.0009 which indicate if we can control happiness the relationship between two can show significant changes and it might improve.

References

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using r*. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.

Lander, J. P. 2014. *R for Everyone: Advanced Analytics and Graphics*. Addison-Wesley Data and Analytics Series. Addison-Wesley. https://books.google.com/books?id=3eBVAgAAQBAJ.