ASSIGNMENT 5 - Exercise 9: Student Survey

Ragunath Gunasekaran

2020-10-06

a. 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.

```
# Covariance of the Survey variables
cov(student_sur_df)
```

```
##              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 between the Survey variables - TimeReading and TimeTV
cov(student_sur_df$TimeReading,student_sur_df$TimeTV)
```

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

```
# covariance between the Survey variables - TimeReading and Happiness
cov(student_sur_df$TimeReading,student_sur_df$Happiness)
```

```
## [1] -10.35009
```

```
# covariance between the Survey variables - TimeReading and Gender
cov(student_sur_df$TimeReading,student_sur_df$Gender)
```

```
## [1] -0.08181818
```

```
# covariance between the Survey variables - TimeTV and Happiness
cov(student_sur_df$TimeTV,student_sur_df$Happiness)
```

```
## [1] 114.3773
```

```
# covariance between the Survey variables - TimeTV and Gender
cov(student_sur_df$TimeTV,student_sur_df$Gender)
```

```
## [1] 0.04545455
```

```r
# covariance between the Survey variables - Happiness and Gender
cov(student_sur_df$Happiness,student_sur_df$Gender)
```

```
## [1] 1.116636
```

```r
# Covariance of the Survey variables in table format
pander(cov(student_sur_df), caption ="Covariance of the Survey variables")
```

Table 1: Covariance of the Survey variables

	TimeReading	TimeTV	Happiness	Gender
TimeReading	3.055	-20.36	-10.35	-0.08182
TimeTV	-20.36	174.1	114.4	0.04545
Happiness	-10.35	114.4	185.5	1.117
Gender	-0.08182	0.04545	1.117	0.2727

a. Conclusion:

As per Covariance between Student Survey variables,

1. Positive Values indicate more related with variables each other. Happiness and Time TV are more related (+114.4)
2. Negative Values indicate opposite related with variables each other. Reading and Time TV are more opposite related (-20.36)

b. 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.

```r
# Examine the Survey data variables
str(student_sur_df)
```

```
## '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 ...
```

```r
head(student_sur_df)
```

```
##   TimeReading TimeTV Happiness Gender
## 1           1     90     86.20      1
## 2           2     95     88.70      0
## 3           2     85     70.17      0
## 4           2     80     61.31      1
## 5           3     75     89.52      1
## 6           4     70     60.50      1
```

```
# Calculate Covariance, Variance, COrrelation  for variables
cov(student_sur_df)
```

```
##               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
```

```
cor(student_sur_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
```

```
var(student_sur_df)
```

```
##               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
```

b. Conclusion

1. What measurement is being used for the variables?

As per str(student_sur_df), we have seen TimerReading, TimeTV,Gender are Integer and Happiness in Decimal. (double) Hence Survey data has 2 formats.

2. 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

As per str(student_sur_df), The survey data is coming in 2 formats.As per data, 1. Time Reading –> Time Reading in Hours 2. Time TV –> Time TV in Minutes 3. Happiness –> Happiness in Percentage 4. Gender –> Male or Female

I used both covariance and Variance and see that same results.

c. 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?

I want to perform both positive and negative correlation test. Also I want to choose Happiness, Since it has both positive and negative value with other variables.

Perform a correlation analysis of:

All variables A single correlation between two a pair of the variables Repeat your correlation test in step 2 but set the confidence interval at 99% Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```
# All variables
pander(cor(student_sur_df), caption ="Correlation of the all Survey variables")
```

Table 2: Correlation of the all Survey variables

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.8831	-0.4349	-0.08964
TimeTV	-0.8831	1	0.6366	0.006597
Happiness	-0.4349	0.6366	1	0.157
Gender	-0.08964	0.006597	0.157	1

```
#A single correlation between two a pair of the variables
cor(student_sur_df$TimeReading,student_sur_df$Happiness, method=c("spearman"))
```

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

```
cor(student_sur_df$TimeReading,student_sur_df$Happiness, method=c("pearson"))
```

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

```
cor(student_sur_df$TimeReading,student_sur_df$Happiness, method=c("kendall"))
```

```
## [1] -0.2889428
```

```
cor(student_sur_df$TimeTV,student_sur_df$Happiness, method=c("kendall"))
```

```
## [1] 0.4630424
```

```
cor(student_sur_df$TimeTV,student_sur_df$Happiness, method=c("pearson"))
```

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

```
cor(student_sur_df$TimeTV,student_sur_df$Happiness, method=c("kendall"))
```

```
## [1] 0.4630424
```

```
#Repeat your correlation test in step 2 but set the confidence interval at 99%
cor.test(student_sur_df$TimeReading,student_sur_df$Happiness, conf.level = 0.99)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  student_sur_df$TimeReading and student_sur_df$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8801821  0.4176242
## sample estimates:
##        cor
## -0.4348663
```

```
cor.test(student_sur_df$TimeTV,student_sur_df$Happiness, conf.level = 0.99)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  student_sur_df$TimeTV and student_sur_df$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

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

confidence interval/correlation coefficient at 99% Reading -0.8801821 0.4176242 TV -0.1570212 0.9306275

sample estimates/correlation coefficient Reading -0.4348663 TV 0.636556

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

```
corvalue <- cor(student_sur_df$TimeReading,student_sur_df$Happiness)
corvalue
```

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

```
corvalue2 <- corvalue ^ 2
corvalue2
```

```
## [1] 0.1891087
```

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

```r
cor.test(student_sur_df$TimeReading,student_sur_df$TimeTV)
```

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

g. 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.

As we know, correlation between happiness and watching tv is 0.4052035, but after gender addition this changed.

References

1. 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.

2. R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

3. Xie, Yihui. 2016. Bookdown: Authoring Books and Technical Documents with R Markdown. Boca Raton, Florida: Chapman; Hall/CRC. https://github.com/rstudio/bookdown.

4. https://bookdown.org/yihui/rmarkdown-cookbook