ASSIGNMENT 5

Abhijit Mandal

2020-10-03

Question 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.

```
library(ggm)
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

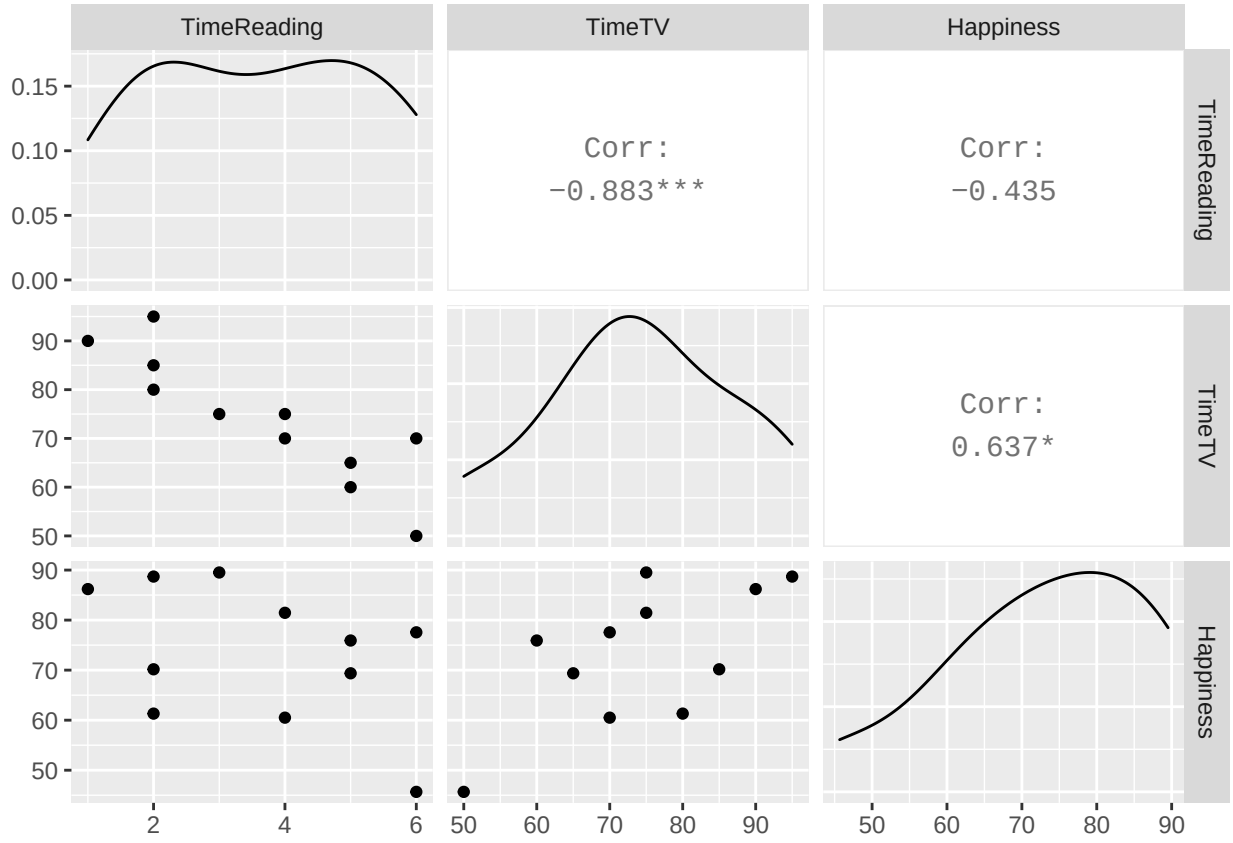
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
setwd("~/Documents/GitHub/dsc520.git ")
## Load the `data/r4ds/heights.csv` to
students_df <- read.csv("data/student-survey.csv")

#create the matrix of the student data for variables TimeReading, TimeTV and Happiness
cor(students_df[, c("TimeReading", "TimeTV", "Happiness")])
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

```
#lets draw a graph for the correlation
GGally::ggpairs(students_df[, c("TimeReading", "TimeTV", "Happiness")])
```



Graph

```
#Heatmap
library(ggm)
library(GGally)
library(dplyr)
library(ggplot2)
setwd("~/Documents/GitHub/dsc520.git ")
# load the reshape package for melting the data
library(reshape2)

# load the scales package for some extra plotting features
library(scales)

## Load the `data/r4ds/heights.csv` to
students_df <- read.csv("data/student-survey.csv")

# build the correlation matrix
studentCor <- cor(students_df[, c("TimeReading", "TimeTV", "Happiness")])

# melt it into the long format
studentMelt <- melt(studentCor, varnames=c("x", "y"), value.name="Correlation")
```

```r
# order it according to the correlation
studentMelt <- studentMelt[order(studentMelt$Correlation), ]

# display the melted data
studentMelt
```

```
##             x           y Correlation
## 2      TimeTV TimeReading  -0.8830677
## 4 TimeReading      TimeTV  -0.8830677
## 3   Happiness TimeReading  -0.4348663
## 7 TimeReading   Happiness  -0.4348663
## 6   Happiness      TimeTV   0.6365560
## 8      TimeTV   Happiness   0.6365560
## 1 TimeReading TimeReading   1.0000000
## 5      TimeTV      TimeTV   1.0000000
## 9   Happiness   Happiness   1.0000000
```

```r
## plot it with ggplot
# initialize the plot with x and y on the x and y axes
ggplot(studentMelt, aes(x=x, y=y)) + geom_tile(aes(fill=Correlation)) +
       scale_fill_gradient2(low=muted("red"), mid="white",
                                          high="steelblue",
                                         guide=guide_colorbar(ticks=FALSE, barheight=10),
                                         limits=c(-1, 1)) +  theme_minimal() + labs(x=NULL, y=NULL)
```

HeatMap

Answer for A

I am using Pearson correlation for this calculation as it ony requires that data are interval for it to be an accurate measure of the linear relationship between two variables, we see that the TimeReading is negatively related to TimeTv with pearson correlation of r = -0.883, this is a reasonably big effect, so we can conclude that as Tv Time increases the Reading time decreases. Also, we see that the TimeReading is negatively related to Happiness with pearson correlation of r =-0.434, again this is a big effect, so we can conclude that as Reading time increases the happiness decreases

Question 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.

Answer for B:

In the Survey data variables, there are following variables with the mentioned measurement as per my assumption: TimeReading - numeric value (hours) TimeTV - numeric value (minutes) Happiness - float value (int percentage) Gender - numeric value (1 represents male and 0 female)

Effect of changing the measurement in covariance

lets convert the TimeReading to Minutes and get the covariance

```
setwd("~/Documents/GitHub/dsc520.git ")
students_df_new <- read.csv("data/student-survey.csv")
students_df_new$TimeReading <- students_df_new$TimeReading * 60
students_df_new
```

```
##    TimeReading TimeTV Happiness Gender
## 1           60     90     86.20      1
## 2          120     95     88.70      0
## 3          120     85     70.17      0
## 4          120     80     61.31      1
## 5          180     75     89.52      1
## 6          240     70     60.50      1
## 7          240     75     81.46      0
## 8          300     60     75.92      1
## 9          300     65     69.37      0
## 10         360     50     45.67      0
## 11         360     70     77.56      1
```

```
cor(students_df_new[, c("TimeReading", "TimeTV", "Happiness")])
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

clearly we can see that there is no effect after changing the measurement

Question 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?

Answer For C

I chose the pearson method with .95 level of confidence with the prediction that the correlation of Reading time vs TV Time will be less than zero

```
cor.test(students_df_new$TimeReading, students_df_new$TimeTV, alternative = "less", method = "pearson",
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeReading and students_df_new$TimeTV
## t = -5.6457, df = 9, p-value = 0.0001577
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000 -0.6684786
## sample estimates:
##        cor
## -0.8830677
```

Pearson method with .95 level of confidence with the prediction that the correlation of Reading time vs Happiness will be less than zero

```
cor.test(students_df_new$TimeReading, students_df_new$Happiness, alternative = "less", method = "pearso
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeReading and students_df_new$Happiness
## t = -1.4488, df = 9, p-value = 0.09067
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000  0.1151482
## sample estimates:
##        cor
## -0.4348663
```

Pearson method with .95 level of confidence with the prediction that the correlation of Tv Time time vs Happiness will be greater than zero

```
cor.test(students_df_new$TimeTV, students_df_new$Happiness, alternative = "greater", method = "pearson"
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeTV and students_df_new$Happiness
```

```
## t = 2.4761, df = 9, p-value = 0.01761
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.1691762 1.0000000
## sample estimates:
##      cor
## 0.636556
```

Question D: Perform a correlation analysis of:

D. 1. All variables

```
cor(students_df_new, use = "complete.obs", method = "pearson")
```

Answer to D.1

```
##             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
```

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

```
cor(students_df_new$TimeReading, students_df_new$TimeTV, use = "complete.obs", method = "pearson")
```

Answer to D.2

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

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

```
cor.test(students_df_new$TimeReading, students_df_new$TimeTV, alternative = "less", method = "pearson",
```

Answer to D.3

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeReading and students_df_new$TimeTV
## t = -5.6457, df = 9, p-value = 0.0001577
## alternative hypothesis: true correlation is less than 0
```

```
## 99 percent confidence interval:
##  -1.0000000 -0.5131843
## sample estimates:
##        cor
## -0.8830677
```

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

Answer to D.4

The calculation suggest that the Reading time is inversely related to TV time at .99 confidence level, i.e. if the TV time increases then the Reading time decreases and vice versa.

Question E

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

Answer to E:

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following object is masked from 'package:ggm':
##
##     rcorr

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
rcorr(as.matrix(students_df_new[, c("TimeReading", "TimeTV", "Happiness")]))
```

```
##             TimeReading TimeTV Happiness
## TimeReading        1.00  -0.88     -0.43
## TimeTV            -0.88   1.00      0.64
## Happiness         -0.43   0.64      1.00
##
## n= 11
##
##
## P
##             TimeReading TimeTV Happiness
## TimeReading             0.0003 0.1813
## TimeTV      0.0003             0.0352
## Happiness   0.1813      0.0352
```

The output of the above correlation shows that

```
- Time TV  is negatively related to Reading Time with a Pearson correlation coefficient of r = -0.88 and
- This significance value tells us that the probability of getting a correlation coefficient this big i
- Hence, we can gain confidence that there is a genuine relationship between TVTime and ReadingTime Our
- So we can say that all of the correlation coefficients are significant.
```

```
coeffDet <- (-0.88) * (-0.88) * 100
coeffDet
```

Coefficient of Determination for TimeTv vs TimeReading

```
## [1] 77.44
```

The coefficient of determination came out to be 77.44%, this means that TVTime is highly correlated with ReadingTime and can account for 77.44% of variation in ReadingTime

```
cor(students_df_new)^2 * 100
```

For all variables

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

```
1. TimeTv account for 40.52% of variation in Happiness
2. Happiness accounts for 18.91% variation in TimeReading
```

Question F

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

Answer to F:

Yes, based on the analysis on the calculation done in previous step we can say that watching more TV caused students to read less.

Question 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.

Answer to G:

1. Partial coorelation between TimeReading and TimeTV controlling Happiness

```
library(ggm)
students_df2 <- students_df_new[, c("TimeReading", "TimeTV", "Happiness")]


pc <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(students_df2))
pc
```

```
## [1] -0.872945
```

```
pc <- pc^2

pc
```

```
## [1] 0.762033
```

```
pcor.test(pc, 1, 11)
```

```
## $tval
## [1] 3.328537
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.01040702
```

So we see that the partial corelation between TimeReading and TimeTV came nearly same 76.20% keeping Happiness in control

2. Partial coorelation between TimeTV and Happiness controlling TimeReading

```
pc2 <- pcor(c("TimeTV", "Happiness", "TimeReading"), var(students_df2))
pc2
```

```
## [1] 0.5976513
```

```r
pc2 <- pc2^2

pc2
```

```
## [1] 0.3571871
```

```r
pcor.test(pc2, 1, 11)
```

```
## $tval
## [1] 1.08163
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.3109403
```

So we see that the partial corelation between TimeTV and Happiness came around 35.71% keeping TimeReading in control