DSC520_7_StudentSurvey_ZimmerAlexis

2022-07-24

Set the working directory to the root of your DSC 520 directory

```
setwd("C:/Users/alexi/OneDrive/Documents/GitHub/dsc520")
Survey_df <- read.csv("data/student-survey.csv")
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

Load libraries

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.1
```

```
##
## 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)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.1
```

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

```
library(formatR)
```

```
## Warning: package 'formatR' was built under R version 4.2.1
```

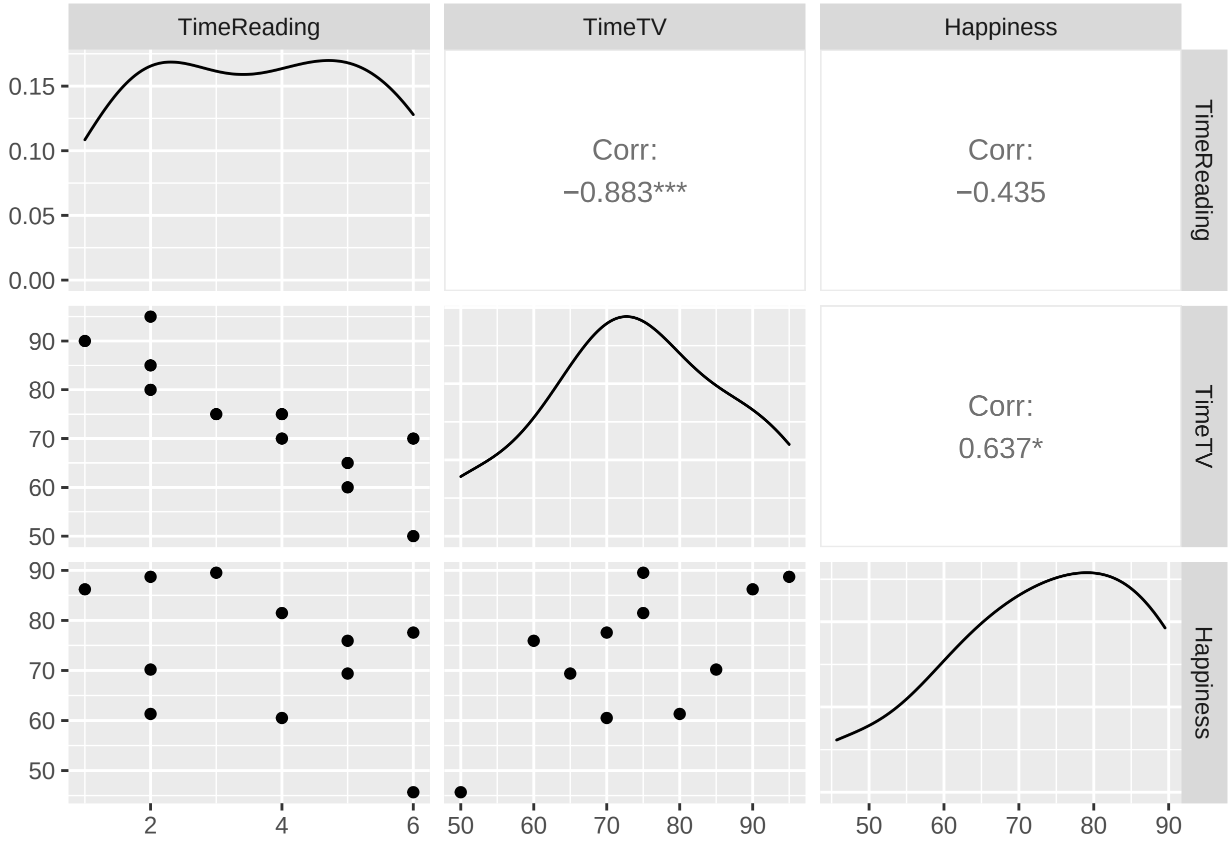
1. Use R to calculate the covariance of the Survey variables and provide an explanation of why you would use this calculation and what the results indicate.

```
## I chose Pearson's correlation, since we can see a linear
## relatinoship within our survey data. TimeReading is
## negatively related to TimeTV. With a Pearon correlation
## of -0.8830677, we can assume that as TimeTV increaes,
## TimeReading decreases. In a similar idea, TimeReading is
## negatively related to happiness with an r-value of
## -0.4348663, so we can assume that as TimeReading
## increases, Happiness decreases.

cor(Survey_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
```

```
GGally::ggpairs(Survey_df[, c("TimeReading", "TimeTV", "Happiness")])
```



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 = hours (numeric) TimeTV =
## minutes (numeric) Happiness = int % (float) Gender =
## binary (numeric)

setwd("C:/Users/alexi/OneDrive/Documents/GitHub/dsc520")
Survey2_df <- read.csv("data/student-survey.csv")
Survey2_df$TimeReading <- Survey2_df$TimeReading * 60
Survey2_df
```

```
##    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(Survey2_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
```

```
## If we changed the measurement being used for the
## variables, the effect on the covariance calculation
## would be null so not a problem. No alternative is
## needed.
```

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?

```
## Pearon's Correlation with a .95 Confidence level,
## assuming that the correlation of TimeTV and Happiness
## will result in > 0
cor.test(Survey_df$TimeTV, Survey_df$Happiness, alternative = "less",
    method = "pearson", conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Survey_df$TimeTV and Survey_df$Happiness
## t = 2.4761, df = 9, p-value = 0.9824
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000  0.8702006
## sample estimates:
##      cor
## 0.636556
```

```r
## Pearson's Correlation with a .95 Confidence level,
## assuming that the correlation of ReadingTime and TVTime
## will result in < 0
cor.test(Survey_df$TimeReading, Survey_df$Happiness, alternative = "less",
    method = "pearson", conf.level = 0.95)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  Survey_df$TimeReading and Survey_df$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
```

4. Perform a correlation analysis of:
5. all variables
6. a single correlation between two a pair of the variables
7. Repeat your correlation test in step 2 but set the confidence interval at 99%
8. Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```r
## 1.
cor(Survey_df, use = "complete.obs", method = "pearson")
```

```
##             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
```

```r
## 2.
cor(Survey_df$TimeReading, Survey_df$TimeTV, use = "complete.obs",
    method = "pearson")
```

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

```r
## 3.
cor.test(Survey_df$TimeReading, Survey_df$TimeTV, alternative = "less",
    method = "pearson", conf.level = 0.99)
```

```
##
##  Pearson’s product-moment correlation
##
```

```
## data:  Survey_df$TimeReading and Survey_df$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
```

```
## 4.  The calculations in the correlation matrix suggest
## ReadingTime is inversely related to TimeTV at a
## confidence interval of 99%, so we can safely assume that
## as ReadingTime increases, TimeTV decreases (and vice
## versa)
```

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

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.2.1

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

```
rcorr(as.matrix(Survey_df[, 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
```

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

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

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

```
##             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
```

```
## Based on the results, I conclude that TimeTV is
## negatively related to TimeReading with an r value of
## -0.88 and a p-value of 0.0003. With a significance value
## close to null, the probability of getting a large
## correlation coefficient in an n-size of 11 if H0 = true
## is incredibly low. Thus, we can assume that there is a
## genuine relationship between TimeTV and TimeReading.
## Additionally, all of our correlation coefficients are
## significant.  The coefficient of Determination is
## 77.44%, which shows that TimeTV is highly correlated
## with TimeReading. TimeTV accounts for 40.52% of the
## variability in Happiness, and Happiness accounts for
## 18.92% of variability in TimeReading
```

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

```
## Based on my analysis, I can say that watching more TV
## caused students to read less, as the calculations showed
## a strong inverse relationship between the two, in which
## watching more TV negatively affected time reading.
```

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.

```
library(ggm)
```

```
## Warning: package 'ggm' was built under R version 4.2.1
```

```
##
## Attaching package: 'ggm'
```

```
## The following object is masked from 'package:Hmisc':
##
##     rcorr
```

```
Survey_df <- Survey_df[, c("TimeReading", "TimeTV", "Happiness")]

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

```
## There is a partial correlation between TimeReading and
## TimeTV if Happiness is controlled, resulting in ~87.30%

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

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

```
pc2 <- pc2^2

pc2
```

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

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

```
## $tval
## [1] 1.08163
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.3109403
## There is a partial correlation between TimeTV and
## Happiness if TimeReading is controlled, resulting in
## ~35.72%
```