Exercise 7.2 (Includes Assignment 5)

John Manzo

May 2nd 2021

Assignment 5

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

```
setwd("C:/Users/manzo/Documents/Bellevue/Classes/DSC 520/Personal GitHub/DSC520")
```

Load the `data/r4ds/heights.csv`

```
heights_df <- read.csv("data/r4ds/heights.csv")
```

Using `cor()` compute correclation coefficients for

height vs. earn

```
cor(heights_df$height, heights_df$earn)
```

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

age vs. earn

```
cor(heights_df$age, heights_df$earn)
```

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

ed vs. earn

```
cor(heights_df$ed, heights_df$earn)
```

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

Spurious correlation - Compute the correlation between these variables

```r
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
cor(tech_spending, suicides)
```

```
## [1] 0.9920817
```

Exercise 7.2.2

i. 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.

```r
cov(survey_df[, c(1, 2:4)])
```

```
##              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 is a statistacal measure for deirming if two variables are related to one another, in that as they each deviate from their means, their movement moves in the same direction (positive or negative). Significate findings:

1. As TimeReading increases, TimeTV and Happiness decrease
2. As TimeTV increases, Happiness increases
3. On average, Gender "1" reads less, watches more TV, and is happier than Gender "2"

ii. 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.

1. TimeReading & TimeTV appear to use the same time-based ratio units, perhaps hours. These two variables can be compared objectively.
2. Happiness appears to use a interval measurement based on an unknown attributes and calculations. Lacking standardization relative to the time variables, covariance statistics comparing happiness to time reading or time watching TV can obscure the magnitude of covariance returned due to lack or standardization between happiness and the time-based variables. This
3. Gender uses a nominal variable (1 or 0) to delineate between two different genders. Because the mean of the variable of the variable depends on the sum of the of the variables divided by overall count, the magnitude of covariance returned when comparing gender with the other variables depends on the ratio of 1s to 0s.
4. Likely the best way to improve upon the interpretability of the output would be to standardize the units of TimeReading, TimeTV, and Happiness based on the variables' standard deviation to derive a correlation coefficient between +1 (perfect possitive correlation) and -1 (perfect negative correlation).

iii. 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 chose the Pearson test for corrlation between the primary variables of interest, TimeReading and TimeTV, because both variables use ratio measures and normal distribution is assumed.

- Prediction: A strong negative correlation between the variables based on the results of the covariance statistic above (-20.36). Because a negative correlation is predicted, the "alternative" command will be employed.

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

```
## 
##  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
## 95 percent confidence interval:
##  -1.0000000 -0.6684786
## sample estimates:
##        cor 
## -0.8830677
```

iv-1. Perform a correlation analysis of all variables.

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

```
##             TimeReading TimeTV Happiness Gender
## TimeReading        1.00  -0.88     -0.43  -0.09
## TimeTV            -0.88   1.00      0.64   0.01
## Happiness         -0.43   0.64      1.00   0.16
## Gender            -0.09   0.01      0.16   1.00
## 
## n= 11 
## 
## 
## P
##             TimeReading TimeTV Happiness Gender
## TimeReading             0.0003 0.1813    0.7932
## TimeTV      0.0003             0.0352    0.9846
## Happiness   0.1813      0.0352           0.6448
## Gender      0.7932      0.9846 0.6448
```

iv-2. Perform a correlation analysis of a single correlation between a pair of the variables.

```
cor.test(survey_df$TimeTV, survey_df$Happiness, method = "pearson")
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  survey_df$TimeTV and survey_df$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05934031 0.89476238
## sample estimates:
##      cor 
## 0.636556
```

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

```
cor.test(survey_df$TimeTV, survey_df$Happiness, method = "pearson", conf.level = 0.99)
```

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

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

> Based on P-values less than 0.05 one can conclude with at least 95% confidence that a negative correlation between TimeReading and TimeTV is expected to be replactable. The same is true of the posative correlation between TimeTV and Happiness, albeit with less confiedence that between TimeReading and TimeTV. Conversely, given their high P-values, correlations between TimeReading and Happiness (negative), TimeReading and Gender (negative), TimeTV and Gender (positive), and Happiness and Gender (positive) cannot be reasonably expected to be replactable.

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

```
cor(survey_df) # correlation coefficient
```

```
##             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
```

```
cor(survey_df)^2 # coefficient of determination
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143
## TimeTV      0.779808529 1.0000000000 0.40520352 0.0000435161
## Happiness   0.189108726 0.4052035234 1.00000000 0.0246527174
## Gender      0.008035714 0.0000435161 0.02465272 1.0000000000
```

The negative correlation between TimeReading and TimeTV has a strong fit with nearly 78% of the variance of each variable being explained by variance of the other. Additionally, despite a low P-value (0.035) for the TimeTV/Happiness correlation noted above, the low R-squared value (0.405) between the two variables indicates a week fit with a majority of variance within each variable not explainable by variance within the other.

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

No. While the negative correlation between TimeReading and TimeTV is strong, this correlation alone does not infer causation. For instance, there might exist other ways students spend their time other than watching TV not captured by the survey that resulted in less time spent reading. Perhaps the the programming available on TV at the time of the survey appealed to some students more than others; this might change week-to-week or month-to-month. Perhaps the time spent reading was higher for those with heavy course loads at the time of the survey. There negative correlation between time spending reading and watching TV is real, but causation has not been proved.

vii. 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.

```
pc <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(survey_df))
pc  # R
```

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

```
pc^2  # R-squared
```

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

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

```
## $tval
## [1] -5.061434
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.0009753126
```

> Controlling for Happiness shows a purer, less noisy correlation relationship between TimeReading and TimeTV. Though previous results suggested the variances between TimeReading/Happiness (negative; R^2 = 0.189) and TimeTV/Happiness (positive, R^2 = 0.405) might be of some signifence, the impact of Happiness on the correlation between TimeReading and TimeTV is largely negligible. With a >99% level of confidence, based on the P and T values, with Happiness controlled the changes in the correlation coefficient and the coefficient of determination between TimeReading and TimeTV are respectively 0.01 and 0.018. That said, one might reason a student's Happiness may be a response variable relative to TimeReading and TimeTV, but Happiness itself does little to impact the relationship between TimeReading and TimeTV.

ASSIGNMENT 5 CODE

```
# Assignment: ASSIGNMENT 5
# Name: Manzo, John
# Date: 2021-05-02

## Set the working directory to the root of your DSC 520 directory
setwd("~/Bellevue/Classes/DSC 520/Personal GitHub/DSC520")

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

## Using `cor()` compute correclation coefficients for
## height vs. earn
cor(heights_df$height, heights_df$earn)
### age vs. earn
cor(heights_df$age, heights_df$earn)
### ed vs. earn
cor(heights_df$ed, heights_df$earn)

## 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
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731,
29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
cor(tech_spending, suicides)
```

ASSIGNMENT 5 CONSOLE OUTPUT

```
> # Assignment: ASSIGNMENT 5
> # Name: Manzo, John
> # Date: 2021-05-02
>
> ## Set the working directory to the root of your DSC 520 directory
> setwd("~/Bellevue/Classes/DSC 520/Personal GitHub/DSC520")
>
> ## Load the `data/r4ds/heights.csv` to
> heights_df <- read.csv("data/r4ds/heights.csv")
>
> ## Using `cor()` compute correclation coefficients for
> ## height vs. earn
> cor(heights_df$height, heights_df$earn)
[1] 0.2418481
> ### age vs. earn
> cor(heights_df$age, heights_df$earn)
[1] 0.08100297
> ### ed vs. earn
> cor(heights_df$ed, heights_df$earn)
[1] 0.3399765
>
> ## 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
> tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731,
29449)
> suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
> cor(tech_spending, suicides)
[1] 0.9920817
```

EXERCISE 7.2 R MARKDOWN CODE

~~~
---
title: "Exercise 7.2 (Includes Assignment 5)"
author: "John Manzo"
date: May 2nd 2021
output:

  pdf_document: default
---

# Assignment 5

## Set the working directory to the root of your DSC 520 directory
```{R}
setwd("C:/Users/manzo/Documents/Bellevue/Classes/DSC 520/Personal GitHub/DSC520")
```
```{R setup, include=FALSE, echo=FALSE}
require("knitr")
opts_knit$set(root.dir = "C:/Users/manzo/Documents/Bellevue/Classes/DSC 520/Personal
GitHub/DSC520")
```

## Load the `data/r4ds/heights.csv`
```{R}
heights_df <- read.csv("data/r4ds/heights.csv")
```

## Using `cor()` compute correclation coefficients for
## height vs. earn
```{R}
cor(heights_df$height, heights_df$earn)
```
## age vs. earn
```{R}
cor(heights_df$age, heights_df$earn)
```
## ed vs. earn
```{R}
cor(heights_df$ed, heights_df$earn)
```

## Spurious correlation - Compute the correlation between these variables
```{R}
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731,
29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
cor(tech_spending, suicides)
```
\newpage
# Exercize 7.2.2
```{R, echo=FALSE}
survey_df <- read.csv("assignment05/student-survey.csv")
```
## i. 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.
~~~

~~~
```{R}
cov(survey_df[, c(1, 2:4)])
```

> Covariance is a statistacal measure for deirming if two variables are related to one
another, in that as they each deviate from their means, their movement moves in the same
direction (positive or negative). Significate findings:

1. As TimeReading increases, TimeTV and Happiness decrease
2. As TimeTV increases, Happiness increases
3. On average, Gender "1" reads less, watches more TV, and is happier than Gender "2"

## ii. 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.
1. TimeReading & TimeTV appear to use the same time-based ratio units, perhaps hours. These
two variables can be compared objectively.
2. Happiness appears to use a interval measurement based on an unknown attributes and
calculations. Lacking standardization relative to the time variables, covariance statistics
comparing happiness to time reading or time watching TV can obscure the magnitude of
covariance returned due to lack or standardization between happiness and the time-based
variables. This
3. Gender uses a nominal variable (1 or 0) to delineate between two different genders. Because
the mean of the variable of the variable depends on the sum of the of the variables divided by
overall count, the magnitude of covariance returned when comparing gender with the other
variables depends on the ratio of 1s to 0s.
4. Likely the best way to improve upon the interpretability of the output would be to
standardize the units of TimeReading, TimeTV, and Happiness based on the variables' standard
deviation to derive a correlation coefficient between +1 (perfect possitive correlation) and -
1 (perfect negative correlation).

\newpage

## iii. 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 chose the Pearson test for corrlation between the primary variables of interest,
TimeReading and TimeTV, because both variables use ratio measures and normal distribution is
assumed.

- Prediction: A strong negative correlation between the variables based on the results of the
covariance statistic above (-20.36). Because a negative correlation is predicted, the
"alternative" command will be employed.

```{R}
cor.test(survey_df$TimeReading, survey_df$TimeTV, method = "pearson", alternative = "less")
```

## iv-1. Perform a correlation analysis of all variables.
```{R, echo=FALSE, include=FALSE}
library(Hmisc)
```
```{R}
rcorr(as.matrix(survey_df[, c("TimeReading", "TimeTV", "Happiness", "Gender")]))
```
\newpage
~~~

~~~
## iv-2. Perform a correlation analysis of a single correlation between a pair of the
variables.
```{R}
cor.test(survey_df$TimeTV, survey_df$Happiness, method = "pearson")
```

## iv-3. Repeat your correlation test in step 2 but set the confidence interval at 99%.
```{R}
cor.test(survey_df$TimeTV, survey_df$Happiness, method = "pearson", conf.level = 0.99)
```
## iv-4. Describe what the calculations in the correlation matrix suggest about the
relationship between the variables. Be specific with your explanation.
> Based on P-values less than 0.05 one can conclude with at least 95% confidence that a
negative correlation between TimeReading and TimeTV is expected to be replactable. The same is
true of the posative correlation between TimeTV and Happiness, albeit with less confiedence
that between TimeReading and TimeTV. Conversely, given their high P-values, correlations
between TimeReading and Happiness (negative), TimeReading and Gender (negative), TimeTV and
Gender (positive), and Happiness and Gender (positive) cannot be reasonably expected to be
replactable.

\newpage
## v. Calculate the correlation coefficient and the coefficient of determination, describe
what you conclude about the results.
```{R}
cor(survey_df) # correlation coefficient
cor(survey_df)^2 # coefficient of determination
```
> The negative correlation between TimeReading and TimeTV has a strong fit with nearly 78% of
the variance of each variable being explained by variance of the other. Additionally, despite
a low P-value (0.035) for the TimeTV/Happiness correlation noted above, the low R-squared
value (0.405) between the two variables indicates a week fit with a majority of variance
within each variable not explainable by variance within the other.

## vi. Based on your analysis can you say that watching more TV caused students to read less?
Explain.
> No. While the negative correlation between TimeReading and TimeTV is strong, this
correlation alone does not infer causation. For instance, there might exist other ways
students spend their time other than watching TV not captured by the survey that resulted in
less time spent reading. Perhaps the the programming available on TV at the time of the survey
appealed to some students more than others; this might change week-to-week or month-to-month.
Perhaps the time spent reading was higher for those with heavy course loads at the time of the
survey. There negative correlation between time spending reading and watching TV is real, but
causation has not been proved.

\newpage
## vii. 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.
```{R, include=FALSE, echo=FALSE}
library(ggm)
```
```{R}
pc <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(survey_df))
pc  # R
pc^2  # R-squared
pcor.test(pc, 1, 11)
```
~~~

```
> Controlling for Happiness shows a purer, less noisy correlation relationship between
TimeReading and TimeTV. Though previous results suggested the variances between
TimeReading/Happiness (negative; R^2 = 0.189) and TimeTV/Happiness (positive, R^2 = 0.405)
might be of some signifence, the impact of Happiness on the correlation between TimeReading
and TimeTV is largely negligible. With a >99% level of confidence, based on the P and T
values, with Happiness controlled the changes in the correlation coefficient and the
coefficient of determination between TimeReading and TimeTV are respectively 0.01 and 0.018.
That said, one might reason a student's Happiness may be a response variable relative to
TimeReading and TimeTV, but Happiness itself does little to impact the relationship between
TimeReading and TimeTV.
```

EXERCISE 7.2 R MARKDOWN OUTPUT

```
processing file: execize_7.2_ManzoJohn.Rmd
  |..                                                                    |   3%
  ordinary text without R code

  |....                                                                  |   6%
label: unnamed-chunk-1
  |.......                                                               |   9%
label: setup (with options)
List of 2
 $ include: logi FALSE
 $ echo   : logi FALSE

  |.........                                                             |  12%
  ordinary text without R code

  |...........                                                           |  16%
label: unnamed-chunk-2
  |.............                                                         |  19%
  ordinary text without R code

  |...............                                                       |  22%
label: unnamed-chunk-3
  |..................                                                    |  25%
  ordinary text without R code

  |....................                                                  |  28%
label: unnamed-chunk-4
  |......................                                                |  31%
  ordinary text without R code

  |........................                                              |  34%
label: unnamed-chunk-5
  |..........................                                            |  38%
  ordinary text without R code

  |............................                                          |  41%
label: unnamed-chunk-6
  |...............................                                       |  44%
  ordinary text without R code

  |.................................                                     |  47%
label: unnamed-chunk-7 (with options)
List of 1
 $ echo: logi FALSE

  |...................................                                   |  50%
  ordinary text without R code

  |.....................................                                 |  53%
label: unnamed-chunk-8
  |.......................................                               |  56%
  ordinary text without R code
  |..........................................                            |  59%
label: unnamed-chunk-9
  |............................................                          |  62%
  ordinary text without R code

  |..............................................                        |  66%
label: unnamed-chunk-10 (with options)
List of 2
 $ echo   : logi FALSE
 $ include: logi FALSE

  |................................................                      |  69%
label: unnamed-chunk-11
  |..................................................                    |  72%
  ordinary text without R code

  |....................................................                  |  75%
label: unnamed-chunk-12
  |.......................................................               |  78%
  ordinary text without R code

  |.........................................................             |  81%
label: unnamed-chunk-13
  |...........................................................           |  84%
  ordinary text without R code

  |.............................................................         |  88%
label: unnamed-chunk-14
  |...............................................................       |  91%
  ordinary text without R code

  |..................................................................    |  94%
label: unnamed-chunk-15 (with options)
List of 2
 $ include: logi FALSE
 $ echo   : logi FALSE

  |....................................................................  |  97%
label: unnamed-chunk-16
  |......................................................................| 100%
  ordinary text without R code


output file: execize_7.2_ManzoJohn.knit.md

"C:/Program Files/RStudio/bin/pandoc/pandoc" +RTS -K512m -RTS execize_7.2_ManzoJohn.utf8.md --
to latex --from markdown+autolink_bare_uris+tex_math_single_backslash --output
execize_7.2_ManzoJohn.tex --lua-filter "C:\Users\manzo\Documents\R\win-
library\4.0\rmarkdown\rmarkdown\lua\pagebreak.lua" --lua-filter
"C:\Users\manzo\Documents\R\win-library\4.0\rmarkdown\rmarkdown\lua\latex-div.lua" --self-
contained --highlight-style tango --pdf-engine pdflatex --variable graphics --variable
"geometry:margin=1in"

Output created: execize_7.2_ManzoJohn.pdf
```