7.2Exercises.R

Rahul Rajeev

2023-01-24

```r
# Assignment: 7.2 Exercises
# Name: Rajeev, Rahul
# Date: 2023-01-23

## Load the ggplot2 package
library(ggplot2)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/rahul/Documents/Bellevue/DSC 520")

## 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
```

```r
### age vs. earn
cor(heights_df$age, heights_df$earn)
```

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

```r
### ed vs. earn
cor(heights_df$ed, heights_df$earn)
```

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

```r
## 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
```

```r
# Assignment: Student Survey Covariance
# Name: Rajeev, Rahul
# Date: 2023-01-23
# libraries
library(ppcor)
```

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

```r
library(dplyr)
```

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

## The following object is masked from 'package:MASS':
##
##     select

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

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

```r
# loading student survey dataset
student_df <- read.csv("data/student-survey.csv")

## 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.

# covariance matrix
cov(student_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
```

```r
# Covariance calculates the direction of the relationship between two variables
# whether one increases with the other, or one increases while the other
# decreases, or if there is no direction at all.
# a positive covariance means that both variables are high or low at the same
# time, a negative covariance means that one variable is high and the other
# is low and a covariance close to 0 implies no direction.

# From our results, timereading and timetv + timereading and happiness have
# negative covariance meaning that as one increases, the other decreases
# timetv and happiness has a positive covariation meaning that as one increases
# or decreases, the other one follows suit.
# timereading and gender, timetv and gender, and happiness and gender all have
# covariance close to 0 meaning that the data between them has no direction

## 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.
```

```
head(student_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
```

```
# based off first glance, time read is probably in terms of just hours,
# time tv is based off of minutes, and happiness is most likely a sort of
# percentage, gender is a binary value that goes between 0 and 1, based off
# of male and female.

# I think changing the units of time read and time tv to both hours or both
# minutes could adjust the value of covariance, but it would make the data look
# better. Using a percentage for happiness is interesting, not sure how they
# calculated it, but keeping it as the value should be relatively fine since
# happiness can't really be measured in terms of minutes and hours. Perhaps
# increasing accuracy of each could benefit the covariance, to 2 decimal places
# changing the gender binary value to male and female as string could look
# better on plots, but not necessary.

## 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?

# I will perform a correlation test on the time reading vs. time tv
# and I predict it will have a negative correlation since spending more time
# reading means less time watching tv

cor.test(student_df$TimeReading, student_df$TimeTV)
```

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

```
# I was correct and the correlation is very close to -1, which means strong
# negative correlation. And the p value is much less than 5% out of the
# 95% confidence interval.

## 4. Perform a correlation analysis of:
## All variables

# correlation matrix
cor(student_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
```

```
## A single correlation between two of the variables
cor(student_df$TimeReading, student_df$TimeTV)
```

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

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

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

# Across the diagonal each of the variables have strong positive correlation of
# 1 for each matching variable.

# Time reading has a strong negative correlation with time tv, a less strong
# negative correlation with happiness, and almost no correlation with gender
# this means that increases/decreases in time reading has a correlation with
# decreases/increases in time tv, and increases/decreases in time reading
# has a correlation with decreases/increases in happiness

# time tv has also a strong negative correlation with time reading, a medium
# strength positive relationship with happiness, and no correlation with gender
# the relationship between timetv and timereading has already been explained,
# but increases/decreases in timetv correlates to increases/decreases in
# happiness

# happiness has a small negative correlation with time reading and a medium
# positive correlation with time tv, and no correlation with gender
# these relationships were explained in the other two answers.

# gender has no correlation with any of the variables across the board,
# and therefore has questionable relationship with the experiment.

## 5. Calculate the correlation coefficient and the coefficient of determination,
## describe what you conclude about the results.
readcor <- cor(student_df$TimeReading, student_df$TimeTV)
```

```
readcor
```

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

```
readdet <- readcor ^ 2
readdet
```

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

```
# the correlation coefficient is very close to -1, which means strong negative
# correlation. And the coefficietn of determination is close to 1, which means
# the model we used can predict an outcome quite well.

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

# No we can't necessarily say that watching more tv caused students to read less
# because correlation doesn't always mean causation. If we knew that tv causes
# people to read less and took that data, then I guess we could. Even though
# our data shows a strong relationship between tv watched and reading, it
# it doesn't prove a causation.

## 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.

three_df = select(student_df, TimeReading:Happiness)
pcor(three_df)
```

```
## $estimate
##             TimeReading     TimeTV Happiness
## TimeReading   1.0000000 -0.8729450 0.3516355
## TimeTV       -0.8729450  1.0000000 0.5976513
## Happiness     0.3516355  0.5976513 1.0000000
## 
## $p.value
##              TimeReading       TimeTV  Happiness
## TimeReading 0.0000000000 0.0009753126 0.31905895
## TimeTV      0.0009753126 0.0000000000 0.06804372
## Happiness   0.3190589526 0.0680437248 0.00000000
## 
## $statistic
##             TimeReading    TimeTV Happiness
## TimeReading    0.000000 -5.061434  1.062425
## TimeTV        -5.061434  0.000000  2.108388
## Happiness      1.062425  2.108388  0.000000
## 
## $n
## [1] 11
## 
## $gp
## [1] 1
## 
## $method
## [1] "pearson"
```

```
# I set Happiness to be the controlled variable. The value of correlation
# between timereading and timetv has slightly changed from -0.88 to -0.87,
# but the correlation between timereading and happiness is now a positive one
# at 0.351 because timereading and happiness are inconsistent with that value.
```