assignment_05_SmitshoekStephen

Stephen Smitshoek

29/04/2022

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

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

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

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

Assignment 05 Code

```
# Assignment: ASSIGNMENT 5
# Name: Smitshoek, Stephen
# Date: 2022-04-27

## Set the working directory to the root of your DSC 520 directory
setwd("C:\\Users\\sksmi\\PeytoAccess\\Personal\\Bellevue\\DSC520\\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
with(heights_df, cor(age, 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)
```

week_07_student_survey_SmitshoekStephen

Stephen Smitshoek

27/04/2022

i.

```
##              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
```

ii.

```
##    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
## 7            4     75     81.46      0
## 8            5     60     75.92      1
## 9            5     65     69.37      0
## 10           6     50     45.67      0
## 11           6     70     77.56      1
```

There are four variables being used,

- TimeReading is measured in hours
- TimeTV is measured in minutes
- Happiness appears to be measured in percentage
- Gender is measured as a binary integer, either 1 or 0

Changing the units of the variables used will change the covariance numbers but it will not nessessarily change the information that is revealed. Covariance tells you the two variables are positively, negativity, or not correlated but the strength of the relationship can only be compared where the units are equal. For example if a third variable was added, TimeSleeping, and it along with TimeTV and TimeReading were all in minutes, you could see the relative strength of the correlation between the three variables but you could not see a relative strength between the relation of TimeSleeping ~ TimeTV and TimeSleeping ~ Gender.

It is best practice to have the two variables share units where possible. For example TimeReading is in hours and TimeTV is in minutes. Because both units are a measure of time it would be best to have them both in the same units.

iii.

```
##
##  Pearson's product-moment correlation
##
## data:  TimeReading and Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates:
##        cor
## -0.4348663
```

A Pearson correlation test was performed assuming that both TimeReading and Happiness are normally distributed and neither are a rank. I predicted that TimeReading and Happiness would be positively correlated. The correlation turned out to be -0.43 which was not what I predicted however the p-value of 0.18 suggests that we cannot assume that the correlation coefficient is not 0, indicating their may be no relationship between Happiness and TimeReading.

iv.

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
```

2.

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

3.

```
##
##  Pearson's product-moment correlation
##
## data:  student.survey$TimeReading and student.survey$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
```

4.

- TimeReading vs TimeTV
  - A value of -0.88 suggests a very strong correlation between time spent reading and time spent watching TV
  - The negative suggests that as time spent reading increases, time spent watching TV decreases
- TimeReading vs Happiness
  - A value of -0.43 suggests a medium correclation between time spent reading and level of happiness
  - The negative suggests that as time spent reading increases, happiness decreases
- TimeReading vs Gender
  - A value of -0.09 suggests there is little to no correlation between gender and time spent reading
- TimeTV vs Happieness
  - A value of 0.63 suggests a strong correlation between time spent watching TV and happiness
  - The positive value suggests that as time spent watching TV increases, happiness also increases.
- TimeTV vs Gender
  - A value of 0.01 suggests there is no correlation between gender and time spent watching TV
- Happiness vs Gender
  - A value of 0.16 suggests that there may be a slight correlation between gender and level of happiness
  - A positive or negative value here has no meaning as gender is not a scale, just a binary (in this data) value

v.

```
##             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
```

```
##             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 correlation coefficients shows how strongly correlated the various variables are, while the coefficient of determination shows how much of the variance in each variable can be attributed to the other variables.

For example time spent reading and time spent watching TV have a very strong correlation (-.88). The coefficient of determination (0.78) suggests that 78% of the variation in TimeTV and TimeReading is could be caused by the other variable. This still does not confirm that these variables are the cause of the variance in the other.

vi.

While it can be said that time spent watching TV and time spent reading are strongly correlated it cannot be said if this is causation. Determining whether watching TV decreases time spent reading or if it is the other way around can not be determined given the data that we have. Additionally it could be a third variable that is causing each of these variables to move and is the actual cause.

vii.

```
## Warning: package 'ggm' was built under R version 4.1.3
```

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

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

Looking at the correlation between TimeReading and TimeTV while holding Happiness constant does not change the correlation coefficient or the coefficient of determination by a significant amount. This suggests that happiness is not having a strong effect on the relationship between TimeReading and TimeTV.

Student Survey Code

```
---
title: "week_07_student_survey_SmitshoekStephen"
author: "Stephen Smitshoek"
date: "27/04/2022"
output: pdf_document
---

```{r include=FALSE}
setwd("C:\\Users\\sksmi\\PeytoAccess\\Personal\\Bellevue\\DSC520\\dsc520")
student.survey <- read.csv("data\\student-survey.csv")
```

i.
```{r echo=FALSE}
cov(student.survey)
```

ii.
```{r echo=FALSE}
student.survey
```
There are four variables being used,

* TimeReading is measured in hours
* TimeTV is measured in minutes
* Happiness appears to be measured in percentage
* Gender is measured as a binary integer, either 1 or 0

Changing the units of the variables used will change the covariance numbers but it will not nessessarily change the information that is revealed. Covariance tells you the two variables are positively, negativity, or not correlated but the strength of the relationship can only be compared where the units are equal. For example if a third variable was added, TimeSleeping, and it along with TimeTV and TimeReading were all in minutes, you could see the relative strength of the correlation between the three variables but you could not see a relative strength between the relation of TimeSleeping ~ TimeTV and TimeSleeping ~ Gender.

It is best practice to have the two variables share units where possible. For example TimeReading is in hours and TimeTV is in minutes. Because both units are a measure of time it would be best to have them both in the same units.

iii.
```{r echo=FALSE}
with(student.survey, cor.test(TimeReading, Happiness, method="pearson"))
```
A Pearson correlation test was performed assuming that both TimeReading and Happiness are normally distributed and neither are a rank. I predicted that TimeReading and Happiness would be positively correlated. The correlation turned out to be -0.43 which was not what I predicted however the p-value of 0.18 suggests that we cannot assume that the correlation coefficient is not 0, indicating their may be no relationship between Happiness and TimeReading.
```

iv.
1.
```{r echo=FALSE}
cor(student.survey)
```
2.
```{r echo=FALSE}
cor(student.survey$TimeReading, student.survey$Happiness)
```

3.
```{r echo=FALSE}
cor.test(student.survey$TimeReading, student.survey$Happiness, conf.level = .
99)
```

4.

* TimeReading vs TimeTV
  + A value of -0.88 suggests a very strong correlation between time spent reading and time spent watching TV
  + The negative suggests that as time spent reading increases, time spent watching TV decreases
* TimeReading vs Happiness
  + A value of -0.43 suggests a medium correclation between time spent reading and level of happiness
  + The negative suggests that as time spent reading increases, happiness decreases
* TimeReading vs Gender
  + A value of -0.09 suggests there is little to no correlation between gender and time spent reading
* TimeTV vs Happieness
  + A value of 0.63 suggests a strong correlation between time spent watching TV and happiness
  + The positive value suggests that as time spent watching TV increases, happiness also increases.
* TimeTV vs Gender
  + A value of 0.01 suggests there is no correlation between gender and time spent watching TV
* Happiness vs Gender
  + A value of 0.16 suggests that there may be a slight correlation between gender and level of happiness
  + A positive or negative value here has no meaning as gender is not a scale, just a binary (in this data) value

v.
```{r echo=FALSE}
cor(student.survey)
cor(student.survey)^2
```

The correlation coefficients shows how strongly correlated the various variables are, while the coefficient of determination shows how much of the variance in each variable can be attributed to the other variables.

For example time spent reading and time spent watching TV have a very strong correlation (-.88). The coefficient of determination (0.78) suggests that 78% of the variation in TimeTV and TimeReading is could be caused by the other variable. This still does not confirm that these variables are the cause of the variance in the other.

vi.
While it can be said that time spent watching TV and time spent reading are strongly correlated it cannot be said if this is causation. Determining whether watching TV decreases time spent reading or if it is the other way around can not be determined given the data that we have. Additionally it could be a third variable that is causing each of these variables to move and is the actual cause.

vii.
```{r echo=FALSE}
library(ggm)
pc <- pcor(c("TimeReading", "TimeTV", "Happiness"), var(student.survey))
pc
pc^2
```
Looking at the correlation between TimeReading and TimeTV while holding Happiness constant does not change the correlation coefficient or the coefficient of determination by a significant amount. This suggests that happiness is not having a strong effect on the relationship between TimeReading and TimeTV.