Week 7 Exercises_WilsonStewart

Stewart Wilson

2022-05-01

Assignment 05

```
# Assignment: ASSIGNMENT 5
# Name: Wilson, Stewart
# Date: 2022-4/-27

## Set the working directory to the root of your DSC 520 directory
setwd("/Users/Stewart/Documents/GitHub/dsc520")

## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("C:/Users/Stewart/Documents/GitHub/dsc520/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 dollar
## 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
```

Student Survey

```
# load student survey files
students <- read.csv("C:/Users/Stewart/Documents/GitHub/dsc520/data/student-survey.csv")
# print head
head(students)
```

```
##   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
```

```
# find covariance of time reading vs time tv
cov(students)
```

```
##              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 Meaning

Covariance is used to see the relationship between the variances of any two variables. A positive covariance indicates that the two variables move in similar directions while a negative covariance indicates they move in opposite directions. For instance, when looking at Time Reading vs Time Watching Tv, there is a negative covariance between the two, meaning that as the latter increases, the former decrease. It does not tell us anything about the strength of that relationship.

The same goes for the rest of the variables. There are negative covariances between TimeReading and Happiness, TimeReading and Gender. There are positive covariances between TimeTV and Happiness, and TimeTV and Gender. These tell us the direction of the relationship but nothing about the strength.

Standard Measurement Issue

It seems that Time Reading is measured in hours, while Time Tv is minutes, Happiness is on a scale from 0 to 100 and Gender is a binary choice, with 0 indicating one gender and 1 another.Covariance is heavily affected by different units of measurements. If they are not the same, then it is hard to tell how big the covariance really is nor can we compare covariances between various variables. In this case, we can partially amend the problem by converting TimeReading into minutes.

However, if we were to attempt to find relationships between variables such as Happiness and TimeTV, we should use correlation measures to have a better idea of the strength of the relationship since correlation is not affected by different standards of measurements.

```
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
```

```
# converting TimeReading into minutes and calculating covariance
students2 <- students %>% mutate(TimeReadingMin= TimeReading * 60)
cov(students2$TimeReadingMin, students2$TimeTV)
```

```
## [1] -1221.818
```

Correlation Analysis

I will be using Spearman's correlation test for TimeReading and TimeTV because they are interval variables and it isn't clear that the data follows a normal distribution. I believe that TimeTV and TimeReading will have a negative correlation.

Correlation Between All Variables

```
cor(students, method="spearman")
```

```
##             TimeReading      TimeTV  Happiness      Gender
## TimeReading  1.00000000 -0.90725363 -0.4065196 -0.08801408
## TimeTV      -0.90725363  1.00000000  0.5662159 -0.02899963
## Happiness   -0.40651964  0.56621595  1.0000000  0.11547005
## Gender      -0.08801408 -0.02899963  0.1154701  1.00000000
```

Correlation Between Time Reading and Time TV

```
cor(students$TimeReading, students$TimeTV, method="spearman")
```

```
## [1] -0.9072536
```

```
# spearman correlation w/ 99% confidence
cor.test(students$TimeReading, students$TimeTV, method="spearman", conf.level=.99, exact=FALSE)
```

Confidence Interval At 99%

```
##
##  Spearman's rank correlation rho
##
## data:  students$TimeReading and students$TimeTV
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9072536
```

What This Means

Looking at the correlation matrix, we can come to a number of preliminary observations. First, Time TV and Time Reading have an incredibly strong negatively correlated relationship of-.9, meaning that as one spends more time reading, one spends less time reading and vice versa. At -.4, Happiness and Time Reading also have a negative correlation, though not as strong as TV and Reading.

Time TV and Happiness have a strong positive correlation; one is happier as they spend more time watching TV (and vice versa).

Since gender in this data set is a discrete binary variable, for Gender and TimeReading to be negatively correlated is to say that one is more likely to be of Gender 0 (let's say woman) the more time is spent reading. Time TV and Gender are also negatively correlated but Gender and Happiness are positively correlated, though the sign matters little since 0 and 1 could be for either gender. The strength of this relationship is pretty small, but we can use this data to find how much gender accounts for the variability between the other variables.

Of course, if we consider that gender is a spectrum than we should use a biserial correlation, which accounts for continuous binaries.

Correlation Coefficients & Coefficients of Determination

The Correlation Coefficients are the correlation calculations conducted in the previous section. To get coefficients of Determination, we square these coefficients

```
coef <- students %>% cor(method="spearman")
# calc coefficient of determination then prints
coef_det <- coef**2
coef_det
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 1.000000000 0.8231091442 0.16525822 0.0077464789
## TimeTV      0.823109144 1.0000000000 0.32060050 0.0008409786
## Happiness   0.165258216 0.3206005004 1.00000000 0.0133333333
## Gender      0.007746479 0.0008409786 0.01333333 1.0000000000
```

The above tells us how much one variable accounts for the variability in another. Time Reading accounts for 82% of the variability in Time TV and 16% of happiness. Time TV accounts for 32% of the variability in Happiness.

The most significant finding here is that TimeTV accounts for 82% of the variability in TimeReading meaning only 18% of its variability is accounted for by other variables.

Causality?

Despite the strong correlation and coefficient of determination between watching TV and reading we cannot say that watching more TV *caused* less reading. This is because it is possible there is a third variable that is responsible for the changes we have observed here. In addition, we cannot determine the direction of causality. It could be that reading less causes watching tv more rather than the other way around.

Partial Correlation

```
library(ggm)
students2 <- students[, c("TimeTV", "Happiness", "TimeReading")]
# partial correlation b/w TimeTV and TimeReading, controlling for Happiness
pc <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(students2))
pc_det <- pc ** 2
# print partial correlation
pc
```

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

```
# print partial correlation of determination
pc_det
```

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

Interpretation

The partial correlation test is a more accurate test for how much Time TV is correlated to Time Reading because it attempts to control for the variance that Happiness accounts for in Time TV and TimeReading. It still does not tell us about causation, but it is a more comprehensive insight into their relationship.