Week 7 Assignment

Eric Vance

January 26 2022

Assignment 5 - Part 1

```
setwd("/Users/eric/Documents/GitHub/dsc520")
heights_df <- read.csv("data/r4ds/heights.csv")
#Measuring correlation between height and earnings
cor(heights_df$height, heights_df$earn)
```

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

```
#Measuring correlation between age and earnings
cor(heights_df$age, heights_df$earn)
```

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

```
#Measuring correlation between education and earnings
cor(heights_df$ed, heights_df$earn)
```

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

```
#Spurious correlation: measuring correlation between tech spending and
suicides
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
```

Assignment 5 - Part 2: Student Survey

Assignment 5 - Part 2: A1

Covariance is used to determine relationships between variables. If variables move in the same direction, they have a positive covariance. Opposites are negative covariance. This is useful to use to see what variables have strong or weak relationships with each other.

```
#Set working dictionary and import data
setwd("/Users/eric/Documents/GitHub/dsc520")
survey_df <- read.csv("data/student-survey.csv")
cov(survey_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
```

Assignment 5 - Part 2: A2 - Examination of Survey Variables

For one, I think Gender should be a categorical variable - there is no point to it being numerical, especially with no key for what number corresponds to what gender. Secondly, it looks like TVTime is measured in minutes, while Reading is measured in hours? If that's the case, units need to be the consistent throughout. Converting minutes to hours would certainly alter the covariance calculation.

Assignment 5 - Part 2: A3

I will be doing a simple correlation test between two variables, time spent watching tv vs happiness. My prediction is that these two will be negatively correlated. As tv time goes up, happiness will go down.

```
setwd("/Users/eric/Documents/GitHub/dsc520")
survey_df <- read.csv("data/student-survey.csv")
cor.test(survey_df$TimeTV, survey_df$Happiness)

## 
##  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
```

Well, I was wrong on this one. Turns out more tv leads to great happiness. Who knew? The correlation coefficient of 0.63 suggests that there is a strong linear relationship between happiness levels and tv time.

Assignment 5 - Part 2: A4 - Correlation Analysis

Here, a correlation analysis will be done between variables in the data.

```
setwd("/Users/eric/Documents/GitHub/dsc520")
survey_df <- read.csv("data/student-survey.csv")

#all variables
cor(survey_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

#two variables - choosing time reading vs happiness
cor(survey_df$TimeReading, survey_df$Happiness)

## [1] -0.4348663

#repeat of above with confidence interval 99%
cor.test(survey_df$TimeTV, survey_df$Happiness, 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
```

Time spent reading has a negative correlation with time spent watching tv. This means as tv time goes up, reading time goes down and vice versa. Gender does not seem to have an impact, as all numbers with gender are close to zero. Reading and happiness are negatively correlated, while tv and happiness are positively correlated. This suggests that more reading leads to lower happiness, while more tv leads to increases in happiness. Not what I would have predicted!

Assignment 5 - Part 2: A5 - Calculate correlation of determination (R2)

I've already done the correlation coefficient above so will not be doing this again.

```
setwd("/Users/eric/Documents/GitHub/dsc520")
survey_df <- read.csv("data/student-survey.csv")

#we know that R-squared is the square of the correlation
rSquared <- function(x)cor(x)^2
rSquared(survey_df)

##             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 R-squared value here is reasonably close to 1, so there is a decent goodness of fit.

Assignment 5 - Part2: A6 - Explanation of Findings

Based on the findings, I can say that watching more tv contributed to less reading. They are heavily negatively correlated, and with an R2 value nearing 1, we can say that there is a solid goodness of fit. Based on those two things, I can safely say as tv time goes up,reading time goes down.

Assignment 5 - Part2: A7 - Partial Correlation and Explanation

```
setwd("/Users/eric/Documents/GitHub/dsc520")
survey_df <- read.csv("data/student-survey.csv")

library('ppcor')

## Loading required package: MASS

pcor.test(survey_df$TimeReading, survey_df$TimeTV,survey_df$Happiness)

##    estimate      p.value statistic  n gp  Method
## 1 -0.872945 0.0009753126 -5.061434 11  1 pearson
```

Here I did a partial correlation between time reading, time watching tv, with happiness controlled. The two appear to still be heavily negatively correlated, and that low of a p-value suggests that this is statistically significant. This really only amplifies my initial assessment. Therefore, I can conclude that tv-time and reading-time are negatively correlated.