Assignment 05 and Student Survey

Rapuru, Supraja

2021-07-25

— Assignment 05 —

Height Correlations

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

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

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

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

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

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

Tech Spending and Suicide Correlation

```
cor(tech_spending, suicides)
```

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

— Student Survey —

Part I

A. Calculate covariance of survey variables

```
cov(ss_df[, c(1:4)])
```

```
##              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
```

B. Why use this calculation?
Covariance in R is calculated by using the cov() function

C. What do the results indicate?
1.Time of reading is negatively related to Time of watching TV, 2.Time of reading is negatively related to Happiness. 3.Time of watching TV is positively related to Happiness. 4.As gender is represented as integer, we can ignore the covariance associated with gender.

Part II

A. What measurement is being used for the variables?
TimeReading: time, in hours (rounded to whole hr) TimeTV: time, in minutes (rounded to nearest 5 min) Happiness: looks like percent. It looks like Happiness index varies from 45.67% to 89.52% Gender - By looking at the values, I assumed that the Gender is measure in boolean.It is not specified that 0 or 1 mean male/female. Need so more info on the variable.

B. Explain what effect changing the measurement being used for the variables
would have on the covariance calculation.
Because time is measured in two different ways, out initial covariance calculation
is not accurate in terms of calculation, but it should still be accurate in terms of showing positive or negative covariance.

C. Would this be a problem? Explain and provide a better alternative if needed.
Differences in how a type of variable is calculated can be problematic. This is true not only for time (min v hours) but also for length (in v ft or imperial v metric) and any other measurement. We should compare like to like to ensure accuracy.

Here we have the corrected covariance after altering the data so that all time
is represented in minutes:

```
cov(ss_edited_df[, c(1:4)])
```

```
##                TimeReadingMin        TimeTV   Happiness      Gender
## TimeReadingMin   10996.363636 -1.221818e+03 -621.005455 -4.90909091
## TimeTV           -1221.818182  1.740909e+02  114.377273  0.04545455
## Happiness         -621.005455  1.143773e+02  185.451422  1.11663636
## Gender              -4.909091  4.545455e-02    1.116636  0.27272727
```

Part III

A. Choose the type of correlation test to perform.
B. Why this test?
C. Make a prediction as to whether or not yield +/- correlation.

checking normality of data

By looking at plots, I can confirm that data is normally distributed. We can used Perason's correlation coefficient to check the correlation between variables.

```
##
##  Pearson's product-moment correlation
##
## data:  ss_edited_df$TimeReadingMin and ss_edited_df$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
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  ss_edited_df$TimeReadingMin and ss_edited_df$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
```

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

Part IV

Correlation analysis of:
A. All variables

```
##                TimeReadingMin       TimeTV  Happiness       Gender
## TimeReadingMin     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
```

B. A single correlation between two of the variables

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

C. Repeat above, but set confidence interval at 99%

```
## 
##  Pearson's product-moment correlation
## 
```

```
## data:  ss_edited_df$TimeReadingMin and ss_edited_df$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
```

D. Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific.
The variables TimeReadingMin and TimeTV are negatively correlated. That is, the more time students spend reading the less time they spend watching TV.

Part V

A. Calc correlation coefficient and coefficient of determination.
Earlier, we looked at correlation between the variables TimeReadingMin and TimeTV and found they were negatively correlated. The coefficient of determination, or Rˆ2, can give us a percentage of variation. That is, we can see how much one variable effects the other.

```
##                TimeReadingMin       TimeTV  Happiness       Gender
## TimeReadingMin     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
```

```
##                TimeReadingMin       TimeTV  Happiness       Gender
## TimeReadingMin    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
```

B. Describe what you conclude about the results.
The coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0.

Looking at coeffiecient of determination between TimeTV and Happiness shared variability is about 40% which would imply that TV time variability effects Happiness upto 40% only, while remaining 60% variability in Happiness must be caused by some other variable.

Part VI

A. Based on analysis, does watching more TV cause students to read less?

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

Looking at coefficient of determination (rˆ2) we can say that variability in TimeReading can cause upto 77% variability in TimeTV. There could be other variables that may cause 23% variability in TimeTV.

Part VII

A. Pick 3 variable and perform a partial correlation.
I will select TimeReadingMin, TimeTV, and Happiness. Run partial correlation between TimeTV and Happiness while controlling TimeReading

B. Be sure to document which variable you are "controlling."
Run partial correlation between TimeTV and Happiness while controlling TimeReading

```
## [1] 0.5976513
```

```
## [1] 0.3571871
```

C. Does this change your interpretation? How, or why not? If we keep TimeReading controlling , the correlation coefficient between TV time and happiness decrease to 0.59 and coefficient of determination has decreased to 35%. This decrease suggests that variation in Happiness was also effected positively by TimeReading by about 5%.