4.1 Assignment: Student Survey

Scott Breitbach

June 22nd 2020

## Covariance

### a. 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.

studentSurvey\_df <- read.csv("completed/Week4/student-survey.csv")  
cov(studentSurvey\_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

Covariance gives us a measure of relatedness of the variables to each other. In this table, where one variable is compared to itself we are given the variance, which is a measure of the average error between the model (or mean) and the actual observations. In the case of variance, the closer to zero, the less overall error meaning the values overall hung out relatively close to the mean. When calculating the covariance between two different variables, you are more or less determining if they are positively or negatively correlated, or if one increases, the other will either increase with it or decrease proportionally. Unless you are using the same units, it can be difficult to compare magnitude of the relatedness.

## Survey Data Variables

### b. 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.

Given that units aren’t provided in the data set and that the survey participants are students, I’m going with the following assumptions: The TimeReading units are in hours and the TimeTV units are in minutes. I went back and forth on the period and ultimately decided these units were probably per day. The Happiness variable is possibly a rating out of 100 and Gender is male or female but it’s not defined which is which.

For the Time units, converting between minutes and hours will have an effect on covariance in that minutes will produce higher numbers than hours. For an easier comparison it would be preferred to have them in the same units. If we knew which gender was which, we could swap the coding and we’d get the negative of the original result. This may provide some information on the magnitude of the covariance, but not the direction.

## Correlation Test

### c. 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 have chosen to compare TimeReading with TimeTV using Pearson’s correlation coefficient because both variables consist of interval data. I believe these will have a negative correlation because the students have a limited supply of time available to be allotted to various tasks.

cor(studentSurvey\_df$TimeReading, studentSurvey\_df$TimeTV, method = "pearson")

## [1] -0.8830677

## Correlation Analysis

### d. Perform a correlation analysis of:

#### d.1. All variables

cor(studentSurvey\_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

#### d.2. A single correlation between a pair of the variables

cor.test(studentSurvey\_df$TimeTV, studentSurvey\_df$Happiness)

##   
## Pearson's product-moment correlation  
##   
## data: studentSurvey\_df$TimeTV and studentSurvey\_df$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

#### d.3. Repeat your correlation test in step 2 but set the confidence interval at 99%

cor.test(studentSurvey\_df$TimeTV, studentSurvey\_df$Happiness, conf.level = 0.99)

##   
## Pearson's product-moment correlation  
##   
## data: studentSurvey\_df$TimeTV and studentSurvey\_df$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

#### d.4. Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

Using the correlation values generated in question d.1., I can say that time spent reading and time spent watching TV are pretty strongly negatively correlated. This makes sense because they both require the use of time, of which there is a limited supply. Time spent reading and happiness are also negatively correlated to a reasonably strong degree. On the flip side, time spent watching TV and happiness have a fairly strong positive correlation, indicating that people who watch more TV are generally happier and people who spend more time reading are generally less happy. The remaining variable, Gender, isn’t really very strongly correlated with any of the variables. The strongest correlation is with Happiness, but even that is relatively small. The correlations between Gender and either time spent reading or time watching TV are both near zero, indicating that there is very little difference between men and women and these other variables.

## Coefficient of Determination

### e. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

# Correlation Coefficient  
cor(studentSurvey\_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

# Coefficient of Determination  
cor(studentSurvey\_df)^2 \* 100

## TimeReading TimeTV Happiness Gender  
## TimeReading 100.0000000 77.98085292 18.910873 0.80357143  
## TimeTV 77.9808529 100.00000000 40.520352 0.00435161  
## Happiness 18.9108726 40.52035234 100.000000 2.46527174  
## Gender 0.8035714 0.00435161 2.465272 100.00000000

As discussed in question d.4., the correlation coefficient tells you on a scale of -1 to 1 how correlated two variables are to each other, with zero being no correlation and +/- 1 being fully correlated, either positively (as one increases the other increases) or negatively (as one increases, the other decreases). The coefficient of determination is the same values squared, then multiplied by 100 to give the percent of variability that the two variables share with each other. The numbers are different, but the results are largely the same in that Gender doesn’t share much variability with any of the other variables. Meanwhile, time watching TV shares 78.0% of its variability with time reading and 40.5% with happiness, leaving 22.0% and 59.5% respectively to be accounted for by other factors. Additionally, happiness shares 18.9% of it’s variability with time spent reading.

## Causation

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

While there is a large (negative) correlation between time spent watching TV and time spent reading, you cannot conclude that one causes the other in either direction, only that if there is an increase in one will likely correspond with a decrease in the other. In this case it is more likely that a third factor, the limited amount of total available time (such as hours in a day) is a contributing factor to the negative correlation between the two variables.

## Partial Correlation

### g. 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.

studentSurvey2\_df <- studentSurvey\_df[, c("TimeReading", "TimeTV", "Happiness")]  
pc <- pcor(c("TimeReading", "Happiness", "TimeTV"), var(studentSurvey2\_df))  
pc

## [1] 0.3516355

pc^2

## [1] 0.1236476

pcor.test(pc, 1, 11)

## $tval  
## [1] 1.062425  
##   
## $df  
## [1] 8  
##   
## $pvalue  
## [1] 0.319059

I was curious about the relationship between time spent reading and happiness when controlling for time spent watching TV, which seemed to have a relatively large correlation with both. When not controlling for time spent watching TV, the correlation between time spent reading and happiness was -0.43, or 18.9% of their variation is shared. When controlling for time spent watching TV, the correlation between the two variables changes from a -0.43 to a +0.35, meaning when we control for the effect of TV watching, the correlation actually changes from a negative to a positive, which makes some sense because there was a very strong negative correlation between tv watching and reading. This leaves 12.4% of shared variance between reading and happiness, the significance of which is shown by the p-value of 0.32, which is not very statistically significant. While time spent watching TV accounts for a large part of the variability in time spent reading and the majority of the variability in happiness, there is still some positive effect between reading and happiness.

## References

* Discovering Statistics Using R (Field, Miles, and Field 2012)

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. <https://books.google.com/books?id=wd2K2zC3swIC>.