DSC 520 Week 5 Correlation Assignment

Dan Wiltse

Jan 12th 2020

Research Question

"Is there a significant relationship between the amount of time spent reading and the time spent watching television?"

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.

```
cov(student_survey2)
```

```
##              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 is a measure of how much two random variable vary together. A large covariance can mean there is a strong relationship between the variables. You can't compare variances with data sets with different scales, as a strong covariance in 1 data set may be weak in another data set, depending on the scales. The larger the numbers, the larger the covariance, so it tells us the variables are related, but doesn't tell us exactly how strong the relationship between the two variables is.

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.

1) TimeReading – Time Reading – Appears to be time in hours
2) TimeTV – Appears to be time in minutes
3) Happiness – Appears to be a rating scale, from 1-100
4) Gender – Appears to be a flag, with 1 being male and 0 being female (or vice versa)

If you are not measuring the same units in the calculation (like hours vs minutes, or inches vs feet), then the units are arbitrary and the covariance may change based on the units compared. This would be a problem if you wanted to compare metrics with different units of measure. Correlation normalizes the differences by including the standard deviations of each metric, so it is dimensionless and gives you a coefficient between -1 and 1, and isn't impacted by the changing in the scale or measurement of the included variables.

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 will use Pearson's R correlation because the data is normally distributed and measured at interval ratio, which are the requirements of using Pearson's R. I predict there will be a negative correlation between time watching TV and time reading, and I also predict there will be a negative correlation between happiness and time spent reading.

D. Perform a correlation analysis of:

1)All variables

```
cor(student_survey2)
```

```
##             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)A single correlation between two of the variables

```
cor.test(student_survey2$TimeReading, student_survey2$TimeTV)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey2$TimeReading and student_survey2$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
```

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

```
cor.test(student_survey2$TimeReading, student_survey2$TimeTV,  conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey2$TimeReading and student_survey2$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
```

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

Time watching TV was significantly correlated with Time Reading, r = -.88, p <.01. Time watching TV was significantly related to Happiness rating. Time Reading was negatively related to Happiness rating.

So while correlation doesn't mean causation, the relationships between the variables tends to show that more time reading is correlated with less Time Watching TV, and more Time watching TV is positively correlated with Happiness Rating, while more time Reading is negatively correlated with Happiness rating.

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

The correlation coefficient between time reading and happiness is -.434 and the correlation of determination is .189, which is the squared value of the correlation coefficient. This shows that time reading is negatively correlated with happiness, and time reading accounts for 18.9% of the variation in happiness scores.

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

As the famous saying goes, correlation does not imply causation, but looking at the correlation data, it does show that time watching TV is highly negatively related to time reading, and the significance value shows that the likelihood of no relationship between the variables is very low.

G. Use TV Time and Happiness while controlling for Gender and perform a partial correlation. Explain how this changes your interpretation and explanation of the results.

```
pc<-pcor(c("TimeTV", "Happiness", "Gender"), var(student_survey2))
```

```
 pc
```

```
## [1] 0.6435158
```

```
 pc^2
```

```
## [1] 0.4141125
```

```
pcor.test(pc,1,11) $pvalue
```

```
## [1] 0.04469059
```

The partial correlation between TV Time and Happiness, while controlling for Gender, was .6435. This has a R squared value of .414, which means that TV time accounts for 41% of the happiness rating, when controlling for Gender. This is very similar to the full correlation between TimeTV and Happiness variables without accounting for Gender.