7.2 Exercise

Felipe Rodriguez

2023-01-29

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

```
setwd("/Users/feliperodriguez/Library/CloudStorage/OneDrive-BellevueUniversity/Github/dsc520/")
student_survey <- read.csv("data/student-survey.csv")
```

Covariance

```
cov(student_survey)
```

```
##              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 the relationship of two variables. When two value trend to be high at the same time, the value is higher. This calculation can be used to see the trends of two values and see how they relate. For example, we can see that happiness and TV Time tend to be low versus happiness and Time Reading. This study can give a quick glance of the trends of the variables.

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.

```
head(student_survey)
```

```
##   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
```

The measurements used are undefined which can create an inconsistency when analyzing covariance. Time reading has a measure of digits less than 5 and Time TV has mesaurements above 70. This gives an indication that Time Reading might be measure in hours as opposed to Time TV might be measured in minutes.

If the values are changed to a consistent measurement, the covariance analysis would be able to yield better results. A better alternative to be to use a consistent form of time amongst the fields Time Reading and Time TV. This will give a clearer output on covariance.

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 am going to be performing a test on happiness and TV Time. I think this test will yield some interesting results since a lot of students spend time watching TV. My prediction is that this test will yield a positive correlation.

Perform a correlation analysis of:

All variables

```
cor(student_survey)
```

```
##             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
```

A single correlation between two a pair of the variables

```
cor.test(student_survey$TimeTV, student_survey$Happiness)
```

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

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

```
cor.test(student_survey$TimeTV,student_survey$Happiness,method="pearson",conf.level = 0.99 )
```

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

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

This calculation is using the Pearson method. The confidence level is set at 99% and the bottom and top levels are -.1570212 and .9306275. With this, it can be assumed that the correlation between TV Time and Happiness will fall between those two values. The correlation found is .636556 which falls within the confidence level. The correlation of the two variables is high which gives us the indication that there is a positive correlation between Happiness and TV Time.

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

Correlation Coefficient

```
cor(student_survey)
```

```
##             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

```
student_survey.lm <- lm(data=student_survey)
summary(student_survey.lm)$r.squared
```

```
## [1] 0.8211648
```

These results show us that there is a strong trend amongst the variables.

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

With the given data, yes that appears to be the trend. However, the measures used make it difficult to give a reliable analysis. They are not consistent measures and these can cause some inconsistencies in the data.

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.

```
library(ppcor)
```

```
## Loading required package: MASS
```

```
pcor.test(student_survey$TimeReading, student_survey$TimeTV, student_survey$Happiness)
```

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

This changes the interpretation because the estimate is low. This value is the partial correlation between the variables. With this value being low, it means that the degree of association between two random variables have a weak relationship.