Student Survey

Kinnick Fox

2023-1-26

I. 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('C://Users/minij/OneDrive/Desktop/Bellevue EDU/DSC520 Stats for DS/Week 7')
library(ppcor)
```

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

```
survey <- read.csv('student-survey.csv')
tv_read <- cor(survey['TimeReading'],survey['TimeTV'])
tv_hap <- cor(survey['Happiness'],survey['TimeTV'])
read_hap <- cor(survey['TimeReading'],survey['Happiness'])
tv_read
```

```
##                 TimeTV
## TimeReading -0.8830677
```

```
tv_hap
```

```
##             TimeTV
## Happiness 0.636556
```

```
read_hap
```

```
##                Happiness
## TimeReading -0.4348663
```

These calculations would be used to find correlation between variables. Time reading and time watching TV have a strong inverse correlation which makes sense because if you are doing one, you are spending less time doing the other. Time watching TV and happiness have a strong correlation while time reading and happiness have a medium inverse correlation. This implies that people who spend more time reading are less happy than people who spend more time watching TV.I did not bother finding the coefficients for gender because it is unclear what 1 and 0 stand for.

II. 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.

It appears that time reading is measured by days where reading took place while time watching TV is measured in minutes. Happiness appears to be measured with a percentage. Gender is measured with a binary integer without indicating what the value actually means. Changing time spent reading to also be measured in minutes (the same as watching TV) should effect the coefficient by making it more accurate. Changing gender to be easily identifiable would also increase readability.

III. 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 would like to use the Kendall correlation test between TimeReading and Happiness. I chose a Kendall test because it uses ranked values as opposed to raw data, similarly to Spearman although Kendall is more robust and therefore preferred. I anticipate that the coefficient will be somewhat inverse.

IV. Perform a correlation analysis of:

1. All variables

```
cor(survey[,],method="kendall")
```

```
##             TimeReading      TimeTV   Happiness      Gender
## TimeReading  1.00000000 -0.80454045 -0.28894280 -0.07824608
## TimeTV      -0.80454045  1.00000000  0.46304237 -0.02507849
## Happiness   -0.28894280  0.46304237  1.00000000  0.09847319
## Gender      -0.07824608 -0.02507849  0.09847319  1.00000000
```

2. A single correlation between two a pair of the variables

```
cor(survey['TimeReading'],survey['TimeTV'])
```

```
##                 TimeTV
## TimeReading -0.8830677
```

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

```
cor.test(survey[,'TimeReading'],survey[,'TimeTV'], conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  survey[, "TimeReading"] and survey[, "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.

Gender appears to have no correlation with any other variable because all coefficients show 0±0.1. Happiness has a moderate correlation to time spent watching TV and a medium inverse correlation to time spent reading. This indicates that individuals who watch TV are more likely to be happy and individuals that read are less likely to be happy. Time spent reading and time spent watching TV have a strong inverse correlation because doing one often means not doing the other.

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

```
survey_lm <- lm(formula = TimeReading ~ TimeTV, data = survey)
summary(survey_lm)
```

```
##
## Call:
## lm(formula = TimeReading ~ TimeTV, data = survey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9452 -0.4922 -0.2846  0.3851  1.8851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30287    1.55704   7.901 2.44e-05 ***
## TimeTV      -0.11697    0.02072  -5.646 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8645 on 9 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7553
## F-statistic: 31.87 on 1 and 9 DF,  p-value: 0.0003153
```

```
cor(survey['TimeReading'],survey['TimeTV'])
```

```
##                 TimeTV
## TimeReading -0.8830677
```

The multiple coefficient of determination is 0.779 meaning that the data stays close to the 1:1 line. The correlation coefficient is -0.883 which means that as one value grows the other will shrink.

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

Yes, the strong inverse correlation coefficient suggests that doing one often means not doing the other which stand to reason because time is the resource used for both measurements.

VII. 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.

```
part_surv <- data.frame(TimeReading = survey["TimeReading"], TimeTV = survey["TimeTV"], Happiness = surv
pcor(part_surv)
```

```
## $estimate
##             TimeReading     TimeTV Happiness
## TimeReading   1.0000000 -0.8729450 0.3516355
## TimeTV       -0.8729450  1.0000000 0.5976513
## Happiness     0.3516355  0.5976513 1.0000000
##
## $p.value
##               TimeReading       TimeTV  Happiness
## TimeReading 0.0000000000 0.0009753126 0.31905895
## TimeTV      0.0009753126 0.0000000000 0.06804372
## Happiness   0.3190589526 0.0680437248 0.00000000
##
## $statistic
##             TimeReading    TimeTV Happiness
## TimeReading    0.000000 -5.061434  1.062425
## TimeTV        -5.061434  0.000000  2.108388
## Happiness      1.062425  2.108388  0.000000
##
## $n
## [1] 11
##
## $gp
## [1] 1
##
## $method
## [1] "pearson"
```

For this partial correlation, the gender variable was controlled. The results and how they can be determined remain unchanged due to the gender variable's correlation being negligible for each other variable.