Student_Survey

Moshe Burnstein

2022-07-22

Calculate covariance of Student Survey

```
student_survey_df <- read.csv('http://content.bellevue.edu/cst/dsc/520/id/resources/student-survey.csv'
str(student_survey_df)
```

```
## 'data.frame':    11 obs. of  4 variables:
##  $ TimeReading: int  1 2 2 2 3 4 4 5 5 6 ...
##  $ TimeTV     : int  90 95 85 80 75 70 75 60 65 50 ...
##  $ Happiness  : num  86.2 88.7 70.2 61.3 89.5 ...
##  $ Gender     : int  1 0 0 1 1 1 0 1 0 0 ...
```

```
student_survey_matrix <- as.matrix(student_survey_df)
cov(student_survey_matrix)
```

```
##              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
```

**The covariance matrix shows the relationship between two variables. A positive covariance indicates that as one variable goes, so does the other. A negative covariance indicates that the variables go in opposite directions. There appears to be a strong positive relationship between Happiness and TimeTv. There appears to be an opposite relationship between TimeReading and TimeTV and between TimeReading and Happiness. The covariance of Gender to all other variables is minimal because Gender is represented by an integer of 1 or 0, so one cannot meaningfully compare.

The TimeReading variable seems to represent hours, while the TimeTv seems to represent minutes. Happiness seems to be on a scale of 1 to 100...with its range between 75 and 90 or so. Gender is arbitrarily assigned a 1 or a 0. One can scale the variables to a standard deviation to get numbers between 0 and 1. If we standardize the covariance, we can assure a value between -1 and 1.**

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.2.1
```

```
## Loading required package: lattice
```

```
## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
rcorr(student_survey_matrix)
```

```
##             TimeReading TimeTV Happiness Gender
## TimeReading        1.00  -0.88     -0.43  -0.09
## TimeTV            -0.88   1.00      0.64   0.01
## Happiness         -0.43   0.64      1.00   0.16
## Gender            -0.09   0.01      0.16   1.00
##
## n= 11
##
##
## P
##             TimeReading TimeTV Happiness Gender
## TimeReading             0.0003 0.1813    0.7932
## TimeTV      0.0003             0.0352    0.9846
## Happiness   0.1813      0.0352           0.6448
## Gender      0.7932      0.9846 0.6448
```

```
cor(student_survey_matrix, method = "spearman")
```

```
##             TimeReading      TimeTV  Happiness      Gender
## TimeReading  1.00000000 -0.90725363 -0.4065196 -0.08801408
## TimeTV      -0.90725363  1.00000000  0.5662159 -0.02899963
## Happiness   -0.40651964  0.56621595  1.0000000  0.11547005
## Gender      -0.08801408 -0.02899963  0.1154701  1.00000000
```

```
rcorr(student_survey_matrix, type = "spearman")
```

```
##             TimeReading TimeTV Happiness Gender
## TimeReading        1.00  -0.91     -0.41  -0.09
## TimeTV            -0.91   1.00      0.57  -0.03
## Happiness         -0.41   0.57      1.00   0.12
## Gender            -0.09  -0.03      0.12   1.00
##
## n= 11
##
##
## P
```

```
##             TimeReading TimeTV Happiness Gender
## TimeReading              0.0001 0.2147    0.7969
## TimeTV      0.0001              0.0694    0.9325
## Happiness   0.2147      0.0694            0.7353
## Gender      0.7969      0.9325 0.7353
```

```
cor(student_survey_matrix, method = "spearman")^2 * 100
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  82.31091442  16.525822   0.77464789
## TimeTV       82.3109144 100.00000000  32.060050   0.08409786
## Happiness    16.5258216  32.06005004 100.000000   1.33333333
## Gender        0.7746479   0.08409786   1.333333 100.00000000
```

Correlation test between 2 variables

```
cor.test(student_survey_df$TimeReading, student_survey_df$Happiness, alternative = "less", method = "ke
```

```
## Warning in cor.test.default(student_survey_df$TimeReading,
## student_survey_df$Happiness, : Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  student_survey_df$TimeReading and student_survey_df$Happiness
## z = -1.1921, p-value = 0.1166
## alternative hypothesis: true tau is less than 0
## sample estimates:
##        tau
## -0.2889428
```

I use Kendall's tau because because this is a small sample and I do not know if it is parametric. There are also tied ranks in TimeReading. I predict a negative correlation because the data shows that as TimeReading increases, so does Happiness decrease. The p-value of 0.12 tells us that we are only 88% probability that the results were not due to something random.

Correlation coefficient and coefficient of determination

```
cor.test(student_survey_df$TimeReading, student_survey_df$Happiness, alternative = "less", method = "ke
```

```
## Warning in cor.test.default(student_survey_df$TimeReading,
## student_survey_df$Happiness, : Cannot compute exact p-value with ties
```

```
##
##  Kendall's rank correlation tau
##
## data:  student_survey_df$TimeReading and student_survey_df$Happiness
## z = -1.1921, p-value = 0.1166
## alternative hypothesis: true tau is less than 0
## sample estimates:
##        tau
## -0.2889428
```

```
(-0.2889428)^2 * 100
```

```
## [1] 8.348794
```

The -0.289 correlation coefficient indicates some negative correlation. Note that Kendall's coefficient is 66-75% less than Pearson's and Spearman, so consider the correlation to be solidly negative.

Watching more tv vs reading

```
tv_vs_read <- cor.test(student_survey_df$TimeTV, student_survey_df$TimeReading, alternative = "less", m
```

```
## Warning in cor.test.default(student_survey_df$TimeTV,
## student_survey_df$TimeReading, : Cannot compute exact p-value with ties
```

```
tv_vs_read
```

```
##
##  Spearman's rank correlation rho
##
## data:  student_survey_df$TimeTV and student_survey_df$TimeReading
## S = 419.6, p-value = 5.761e-05
## alternative hypothesis: true rho is less than 0
## sample estimates:
##        rho
## -0.9072536
```

```
(-0.9072536)^2 * 100
```

```
## [1] 82.31091
```

The p-value of 5.761e-05 indicates that the rho of -0.907 is meaningful and there is a very strong negative correlation. The r^2 of 82% indicates that 82% of the variables' variability is correlated to the other variable. However we cannot discern that watching more tv causes students to read less any more than reading less causes students to watch more tv.

Show correlation between TimeTv and TimeReading, controlling for Happiness

```
library(ggm)
```

```
## Warning: package 'ggm' was built under R version 4.2.1
```

```
##
## Attaching package: 'ggm'
```

```
## The following object is masked from 'package:Hmisc':
##
##     rcorr
```

```
par_cor <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(student_survey_df))
par_cor
```

```
## [1] -0.872945
```

```
par_cor^2
```

```
## [1] 0.762033
```

```
pcor.test(par_cor, 1, 11)
```

```
## $tval
## [1] -5.061434
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.0009753126
```

While the correlation between TimeTV and TimeReading is reduced when controlling for Happiness, there is still a strong negative correlation of 0.873. The R^2 is 76% which shows that 76% of the variance is due solely to the relationship between TimeTV and TimeReading. The p-value is less than 0.001, which indicates significance.