ASSIGNMENT 7.2

Siddhartha Bhaumik

2022-04-30

Add Citations

- R for Everyone
- Discovering Statistics Using R

Exploring Student Survey Data

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

```
## Load the `data/student-survey.csv` to
stud_survey_df <- read.csv("/Users/siddharthabhaumik/Documents/GitHub/dsc520/data/student-survey.csv")

print(stud_survey_df)
```

```
##    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
## 7            4     75     81.46      0
## 8            5     60     75.92      1
## 9            5     65     69.37      0
## 10           6     50     45.67      0
## 11           6     70     77.56      1
```

```
# Get the structure of the data frame
str(stud_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 ...
```

Q1: Calculate the covariance of the Survey variables and provide an explanation of why you would use this calculation and what the results indicate.

```
## Using `cov()` compute covariance for
## Time Reading vs. Time TV
cov(stud_survey_df$TimeReading,stud_survey_df$TimeTV,method = "pearson")
```

```
## [1] -20.36364
```

```
cov(stud_survey_df$TimeReading,stud_survey_df$TimeTV,method = "spearman")
```

```
## [1] -9.775
```

```
## Time Reading vs. Happiness
cov(stud_survey_df$TimeReading,stud_survey_df$Happiness, method = "pearson")
```

```
## [1] -10.35009
```

```
## Time TV vs. Happiness
cov(stud_survey_df$TimeTV,stud_survey_df$Happiness, method = "pearson")
```

```
## [1] 114.3773
```

```
## Time Reading vs. Gender
cov(stud_survey_df$TimeReading,stud_survey_df$Gender, method = "pearson")
```

```
## [1] -0.08181818
```

```
## Time TV vs. Gender
cov(stud_survey_df$TimeTV,stud_survey_df$Gender, method = "pearson")
```

```
## [1] 0.04545455
```

```
## compute covariance for entire dataset
cov(stud_survey_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
```

Answer: From the results, its clear that their is negative relationship between time spent reading when compared with happiness and gender.For time spend watching TV, it shows a high positive relationship with happiness.

Q2: 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.

```
print(stud_survey_df)
```

```
##    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
## 7            4     75     81.46      0
## 8            5     60     75.92      1
## 9            5     65     69.37      0
## 10           6     50     45.67      0
## 11           6     70     77.56      1
```

```
# Get the structure of the data frame
str(stud_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 ...
```

```
# Get the dataframe summary
summary(stud_survey_df)
```

```
##   TimeReading        TimeTV        Happiness         Gender
##  Min.   :1.000   Min.   :50.00   Min.   :45.67   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:67.50   1st Qu.:65.34   1st Qu.:0.0000
##  Median :4.000   Median :75.00   Median :75.92   Median :1.0000
##  Mean   :3.636   Mean   :74.09   Mean   :73.31   Mean   :0.5455
##  3rd Qu.:5.000   3rd Qu.:82.50   3rd Qu.:83.83   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :95.00   Max.   :89.52   Max.   :1.0000
```

Answer: The student survey data has 4 variables and 11 observations.3 variables are of datatype Integer and one variable(Happiness) of datatype Number.Also, Happiness variable has fractional data upto 2 decimal places.By looking at the data and summary, I can see the min and max values for each variables but many key information are missing like measurement frequency for time reading, time tv, happiness. Is it daily stats, weekly, monthly, yearly, etc. No such information is provided. Also, from the data it seems time reading is captured in hours whereas time TV in minutes.Both are whole numbers.For Happiness, from the data it seems to be in percentage but not clear. Same goes for gender, we don't know what 0 and 1 represents.So, knowing the correct measurement scale/frequency will change the covariance results as well.

Q3. 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?

```
cor(stud_survey_df,method = "pearson")
```

```
##             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
```

```
cor(stud_survey_df,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
```

```
cor(stud_survey_df,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
```

Answer: All 3 correlation methods gives similar results from negative and positive relationship perspective but I will go with Pearson method since its most widely used and relationship between variables(happiness vs timereading & timetv) seems to be linear.

Q4. Perform a correlation analysis of:

i.All variables

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

```
## Using `cor()` compute correlation for
## Time Reading vs. Time TV
cor(stud_survey_df$TimeReading,stud_survey_df$TimeTV,method = "pearson")
```

```
## [1] -0.8830677
```

```
cor(stud_survey_df$TimeReading,stud_survey_df$TimeTV,method = "spearman")
```

```
## [1] -0.9072536
```

```
## Time Reading vs. Happiness
cor(stud_survey_df$TimeReading,stud_survey_df$Happiness, method = "pearson")
```

```
## [1] -0.4348663
```

```
## Time TV vs. Happiness
cor(stud_survey_df$TimeTV,stud_survey_df$Happiness, method = "pearson")
```

```
## [1] 0.636556
```

```
## Time Reading vs. Gender
cor(stud_survey_df$TimeReading,stud_survey_df$Gender, method = "pearson")
```

```
## [1] -0.08964215
```

```
## Time TV vs. Gender
cor(stud_survey_df$TimeTV,stud_survey_df$Gender, method = "pearson")
```

```
## [1] 0.006596673
```

```
## compute covariance for entire dataset
cor(stud_survey_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
```

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

```
## Time Reading vs. Happiness
cor.test(stud_survey_df$TimeReading,stud_survey_df$Happiness, method = "pearson",conf.level = .99)
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  stud_survey_df$TimeReading and stud_survey_df$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8801821  0.4176242
## sample estimates:
##        cor
## -0.4348663
```

```
## Time TV vs. Happiness
cor.test(stud_survey_df$TimeTV,stud_survey_df$Happiness, method = "pearson",conf.level = .99)
```

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

iv. Describe what the calculations in the correlation matrix suggest about the relationship between the variables.

Answer: Based on the results, I can see happiness and time reading has negative relationship whereas happiness and time watching TV has positive relationship.

Q5. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```
cor_timeread_df <- cor(stud_survey_df$TimeReading,stud_survey_df$Happiness) ^ 2

print(cor_timeread_df)
```

```
## [1] 0.1891087
```

```
cor_timetv_df <- cor(stud_survey_df$TimeTV,stud_survey_df$Happiness) ^ 2

print(cor_timetv_df)
```

```
## [1] 0.4052035
```

Answer: The r-squared value for both cases are closer to 0. It means the variables selected doesnt have much correlation.

Q6. Based on your analysis can you say that watching more TV caused students to read less? Explain.

```
## Time Reading vs. Time TV
cor_time_rd_tv_df <- cor(stud_survey_df$TimeReading,stud_survey_df$TimeTV,method = "pearson") ^ 2

print(cor_time_rd_tv_df)
```

```
## [1] 0.7798085
```

Answer: The r-squared value is closer to 1. It means the variables selected has a correlation and we can say that watching more TV caused students to read less.

Q7. 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.

```
pcor.test(stud_survey_df$TimeTV,stud_survey_df$Happiness, stud_survey_df$Gender)
```

```
##    estimate    p.value statistic  n gp  Method
## 1 0.6435158 0.04469059  2.377919 11  1 pearson
```

Answer: Based on initial correlation results, time spend watching TV shows higher correlation with Happiness when compared to time reading.But, now adding 3rd variable "gender" has changed the partial correlation between TimeTV & Happiness and has increased it further from 0.40 to 0.64.