```
---
title: "ASSIGNMENT 4.1_Student_Survey"
author: "Bhushan Suryawanshi"
date: '2020-06-22'

---
```

As a data science intern with newly learned knowledge in skills in statistical correlation and R programming, you will analyze the results of a survey recently given to college students. You learn that the research question being investigated is: "Is there a significant relationship between the amount of time spent reading and the time spent watching television?" You are also interested if there are other significant relationships that can be discovered? The survey data is located in this StudentSurvey.csv file.

```
    student_df <- read.csv("Student-Survey.csv", stringsAsFactors = FALSE)
    head(student_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
```

```
    summary(student_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
```

```
    str(student_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 ...
```

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_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-

> Calculating the covariance is a good way to assess whether two variables are related to each other. A positive covariance indicates that as one variable deviates from the mean, the other variable deviates in the same direction. On the other hand, a negative covariance indicates that as one variable deviates from the mean (e.g., increases), the other deviates from the mean in the opposite direction (e.g., decreases). (Ref - Discovering Statistics Using R (Field, Miles, and Field 2012, 316))

Covariance shows the variability of two variables. In the student dataset we can see that time of reading "TimeReading" is negatively impacting time watching TV ("TimeTv"). It has the covariance of -20.3637. Which means - if you read more then you get less time to watch TV or vice versa. Where as watching TV ("TimeTV") is positively impacting happiness quotient ("Happiness"). It means students watching more TV are more happy.

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.

Answer-

> There is, however, one problem with covariance as a measure of the relationship between variables and that is that it depends upon the scales of measurement used. So, covariance is not a standardized measure. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 316))

Based on the dataset details we can see that the reading time "TimeTV" is represented in the minute format where as reading time "TimeReading" is in hour format. This makes our covariance calculation a non-standard approach.

```
modified_df <- student_df
modified_df$TimeReading <- modified_df$TimeReading * 60
head(modified_df)
```

```
##   TimeReading TimeTV Happiness Gender
## 1          60     90     86.20      1
## 2         120     95     88.70      0
## 3         120     85     70.17      0
## 4         120     80     61.31      1
## 5         180     75     89.52      1
## 6         240     70     60.50      1
```

```
modified_cov <- cov(modified_df)
round(modified_cov, 2)
```

```
##             TimeReading   TimeTV Happiness Gender
## TimeReading    10996.36 -1221.82   -621.01  -4.91
## TimeTV         -1221.82   174.09    114.38   0.05
## Happiness       -621.01   114.38    185.45   1.12
## Gender            -4.91     0.05      1.12   0.27
```

Now changing measurement of "TimeReading" to minutes shows that the covariance has changed to very different values. Again if we change dataset we don't know if it will remain same. Hence we need to standardize the covariance. The covariance standardization can be done using standard deviation. Here we are looking at 2 variables and hence we have two standard deviations. So to calculate standard covariance which is also known as **Correlation Coefficient** we use -

$r = \frac{COV_{xy}}{s_x s_y} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{(N-1) s_x s_y}$

The coefficient equation above is known as **Pearson product-moment correlation coefficient** or **Pearson correlation coefficient**

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?

Answer- We are using **Pearson's Correlation** test here as we want to use confidence intervals as well as our data is interval data and it makes more sense to use Pearson's Correlation test. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 329))

d. Perform a correlation analysis of:

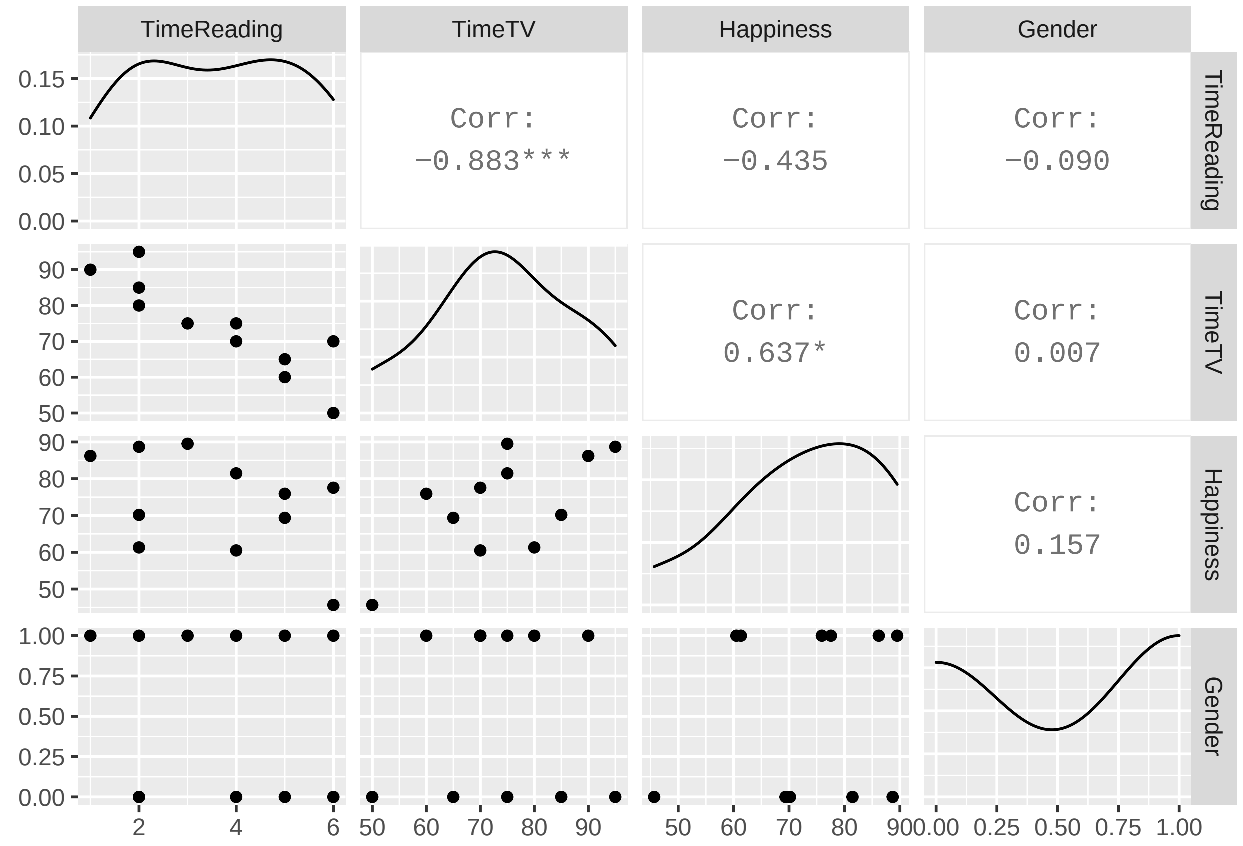
```
1. All variables
```

```
cor(student_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
```

```
GGally::ggpairs(student_df)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```



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

```r
with(student_df, cor.test(Happiness, TimeReading,
  alternative="two.sided", method="pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Happiness and TimeReading
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates:
##        cor
## -0.4348663
```

```r
with(student_df, cor.test(Happiness, TimeTV, alternative="two.sided",
  method="pearson"))
```

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

```
with(student_df, cor.test(TimeReading, TimeTV, alternative="two.sided",
   method="pearson"))
```

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

```
with(student_df, cor.test(Happiness, TimeReading,
  alternative="two.sided", method="pearson", conf.level = 0.99))
```

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

```
with(student_df, cor.test(Happiness, TimeTV, alternative="two.sided",
  method="pearson", conf.level = 0.99))
```

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

```
with(student_df, cor.test(TimeReading, TimeTV, alternative="two.sided",
   method="pearson", conf.level = 0.99))
```

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

Answer- Based on the above correlation matrix and correlation tests we can say that Happiness and Time reading are negatively related which means if students read more they are less happy. In case of Time watching TV is positively related to Happiness, where student watching TV are more happy. With reference to reading time and watching TV we see negative relation. Students watching more TV are getting less time to read.

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

```
#Correlation coefficient
cor(student_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
```

```
#Coefficient of Determination
cor(student_df)^2 * 100
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

Answer-

> Although we cannot make direct conclusions about causality from a correlation, we can take the correlation coefficient a step further by squaring it. The correlation coefficient squared (known as the coefficient of determination, R^2) is a measure of the amount of variability in one variable that is shared by the other. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 334))

In our student survey example the correlation coefficient tells us that the watching TV is negatively related to reading. However we don't know how much percent of affected reading time is because of watching TV. This is where **Coefficient of Determination** comes handy. It shows us what percent of reading is affected by watching TV. So above R^2 matrix shows that the 77% of the time the reading is affected by watching TV.

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

Answer - Based on correlation test of student survey attributes we can say reading is affected by watching TV. Also as we have seen coefficient of determination also shows as much as 77% of the time reading time is affected by watching TV.

g. 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.

```r
library(ggm)
```

```
## Warning: package 'ggm' was built under R version 4.0.2
```

```r
pcor(c( "TimeTV", "TimeReading", "Happiness"), var(student_df))
```

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

Answer- Partial correlation analysis using TimeTv, TimeReading and Happiness shows that the time watching TV is negatively affecting reading time. Also when we keep Happiness constant doesn't affect much the relation between watching TV and reading time. With correlation test we had r = -0.88 where as with partial test we get partial correlation of -0.87.

References:

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.