```
---
title: "Assignment05_Exercise9_StudentSurvey"
author: "Sumbaraju Aditya"
date: "1/16/2021"
---
```

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.

```
      setwd("C:/BU/DSC520/assignment_repo/dsc520/completed/assignment05")
      getwd()
```

```
## [1] "C:/BU/DSC520/assignment_repo/dsc520/completed/assignment05"
```

```
      student_df <- read.csv("../../data/Student-Survey.csv", stringsAsFactors = FALSE)
      head(student_df,n=5)
```

```
##   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
```

```
      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<- cov(student_df)
    round(cov, 2)
```

```
##             TimeReading TimeTV Happiness Gender
## TimeReading        3.05 -20.36    -10.35  -0.08
## TimeTV           -20.36 174.09    114.38   0.05
## Happiness        -10.35 114.38    185.45   1.12
## Gender            -0.08   0.05      1.12   0.27
```

Answer-

Covariance is a measure of two random variables that vary together. It's similar to variance, but where variance tells you how a single variable varies, Covariance tells you how two variables vary together. A positive covariance indicates that as one variable deviates from the mean and the other variable deviates in the same direction. On the other hand, a negative covariance indicates that one variable deviates from the mean and the other deviates from the mean in the opposite direction w.r.t first variable. (Ref - Discovering Statistics Using R (Field, Miles, and Field 2012, 208)) In my opinion, Covariance shows the variability of the two variables. Observation1: In the student dataset, If we observe, the quotient "TimeReading" is negatively impacting time watching TV ("TimeTv") and "Happiness." It has the Covariance of -20.36 and -10.35, respectively. If we read more, we get less time watching TV and comparatively less happy or vice versa. Observation2: watching TV ("TimeTV") is positively impacting the happiness quotient ("Happiness"). It means students watching more TV are happier and spends less time reading.

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-

- In my view, Covariance values are not standardized. The value for a perfect linear relationship depends on the scale of data. Because the input data are not standardized, it would be difficult to determine the relationship between the variables. Based on the dataset student_df1, we can observe that "TimeTV" is scaled in the minute format, whereas "TimeReading" is in hour format. And this is undoubtedly a case of a non-standard approach covariance calculation.

```
student_df1 <- student_df
head(student_df1,n=5)
```

```
##   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
```

```
student_df2 <- student_df
student_df2$TimeReading <- student_df2$TimeReading * 60
head(student_df2,n=5)
```

```
##   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
```

```
cov2 <- cov(student_df2)
round(cov2, 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
```

Let's change the scale of measurement "TimeReading" to minutes. A converted dataset is student_df2. If we observe the Covariance, it has changed to different values. If we try to apply the standardization on other measures, we cannot justify that Covariance remains constant after the scale change is done. Hence there is always a need to proceed with standardization. If

we want to express the Covariance in a standard unit of measurement, we can divide it as per Pearson correlation coefficient . In our case, there are two variables and, hence, two standard deviations. When we calculate the Covariance, we calculate two deviations (one for each variable). Therefore,we do the same for the standard deviations: we multiply them and divide by the product of this multiplication. The standardized Covariance is known as a correlation coefficient (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 209)) 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- >I am opting for **Pearson's Correlation** test, and below are the facts: Pearson's correlation requires only that data are interval for it to be an accurate measure of the linear relationship between two variables.conf.level allows you to specify the width of the confidence interval computed for the correlation. Confidence intervals are produced only for Pearson's correlation coefficient (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 218))

d. Perform a correlation analysis of:

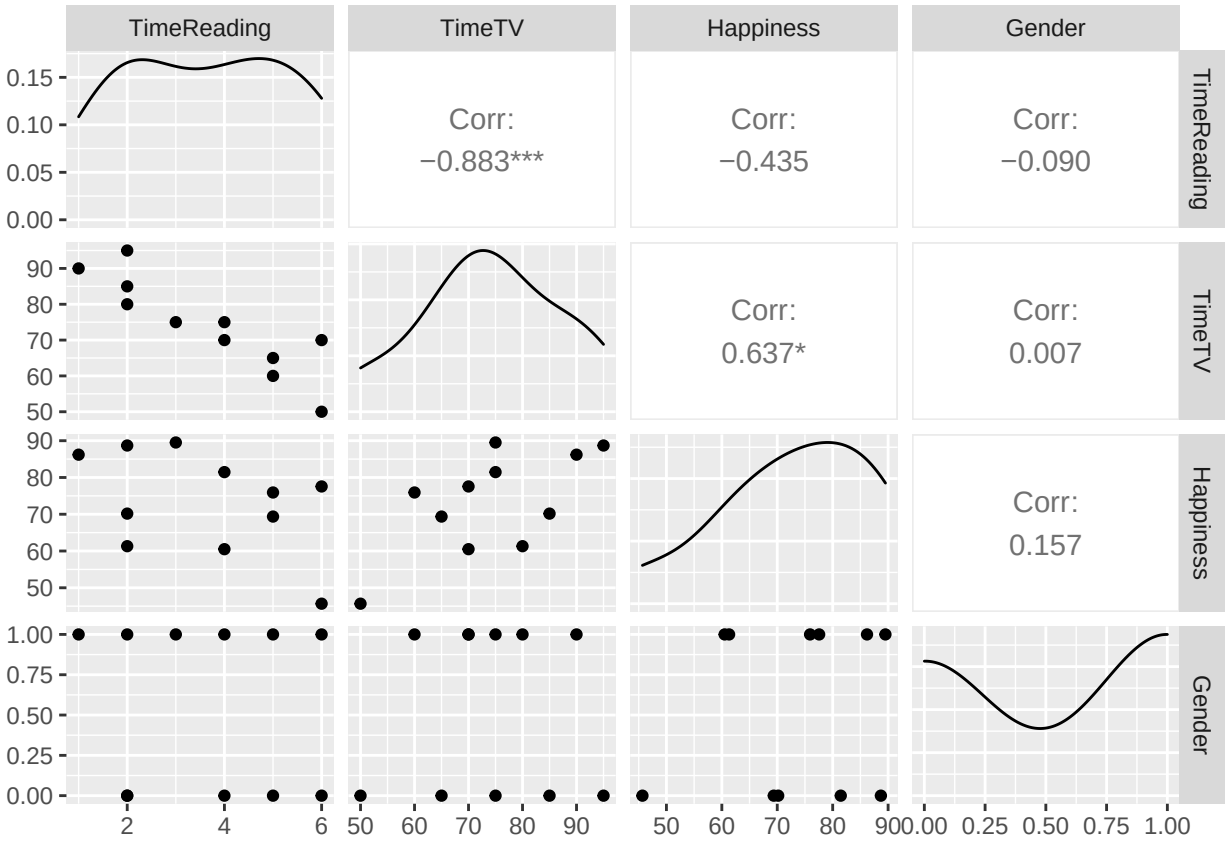
1. All variables

```
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

```
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
```

```
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- >As per the above correlation matrix and correlation test observations, we can justify that Happiness and Time reading are negatively related, which means if students read more, they are less happy. In the case of Time watching TV; positively correlates with Happiness, where students watching TV are happier. Regarding reading Time and watching TV, we see a 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
round(cor(student_df),2)
```

```
##             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
```

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

```
##             TimeReading TimeTV Happiness Gender
## TimeReading      100.00  77.98     18.91   0.80
## TimeTV            77.98 100.00     40.52   0.00
## Happiness         18.91  40.52    100.00   2.47
## Gender             0.80   0.00      2.47 100.00
```

Answer-

> In our student survey example the correlation coefficient tells us that the watching TV is negatively related to reading. However we cannot be able to complie the percent of affected reading time is because of watching TV. to solve this gap **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. We cannot make any direct conclusions about causality from a correlation, If we take the apprach of 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, 222))

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

Answer - >correlation test of student survey measures justifies that reading is affected by watching TV. As well as 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)
pcor(c( "TimeTV", "TimeReading", "Happiness"), var(student_df))
```

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

Answer- >The correlation test results states that r = -0.88 and where as partial correlation stand as -0.87.Partial correlation analysis using TimeTv, TimeReading and Happiness depicts that the time watching TV is negatively affecting reading time. And the Happiness constant doesn't affect much the relation between watching TV and reading time.

References:

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