ASSIGNMENT 7

Janine Par

2022-04-30

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/janin/OneDrive/Documents/R_repo/dsc520/")
## Load the `data/student-surver.csv
studentsurvey_df <- read.csv("data/student-survey.csv")
studentsurvey_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
```

```
#Use R to calculate the covariance of the Survey variables and provide an explanation of why you would

#Covariance of Survey Variables
cov(studentsurvey_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
```

Covariance is used to measure the average relationship between two variables. Calculating covariance is a good way to verify if variables are related to each other and to see if changes on one variable are similar in the other variable.

1. Positive Covariance indicates that as one variable deviates from the mean and the other deviates in the same direction.
2. Negative Covariance indicates one variable deviates from the mean the other deviates from the opposite direction

#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.

```
str(studentsurvey_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 ...
```

The following variables are in the student survey dataframe:

- TimeReading is a continuous ratio variable which is an integer and assumed to be time spent by student reading in hours
- TimeTV is a continuous ratio variable which is an integer and assumed to be time spent by student watching TV in minutes
- Happiness is a continuous ratio variable which is a number and assumed to be percentage rate of student happiness
- Gender is a categorical binary variable which is an integer representing Male and Female

The two variables:TimeReading and TimeTV seem to have a different scale which could be a problem when comparing because of covariance of dependence on the measurement scale. Better alternative is to standardize the variables to same unit.

#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 and doing a two pearson correlation test to see if there are linear correlation between:

1. TimeReading and Happiness
2. TimeTV and Happiness.

I am more entertained watching tv so my prediction is that the TimeTV and Happiness will yield to positive correlation

```
cor.test(studentsurvey_df$TimeReading, studentsurvey_df$Happiness,method = "pearson")
```

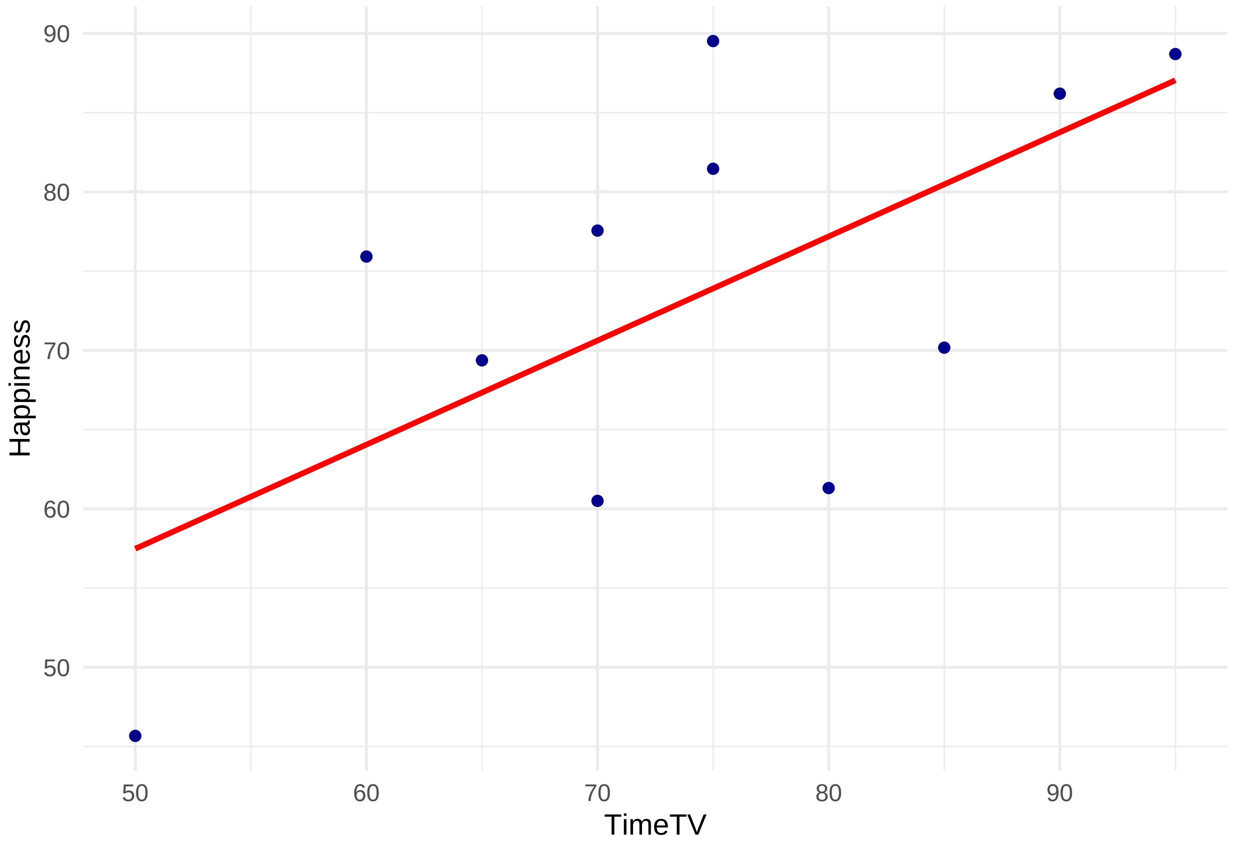
```
##
##  Pearson's product-moment correlation
##
## data:  studentsurvey_df$TimeReading and studentsurvey_df$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates: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

```
cor.test(studentsurvey_df$TimeTV, studentsurvey_df$Happiness,method = "pearson")
```

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

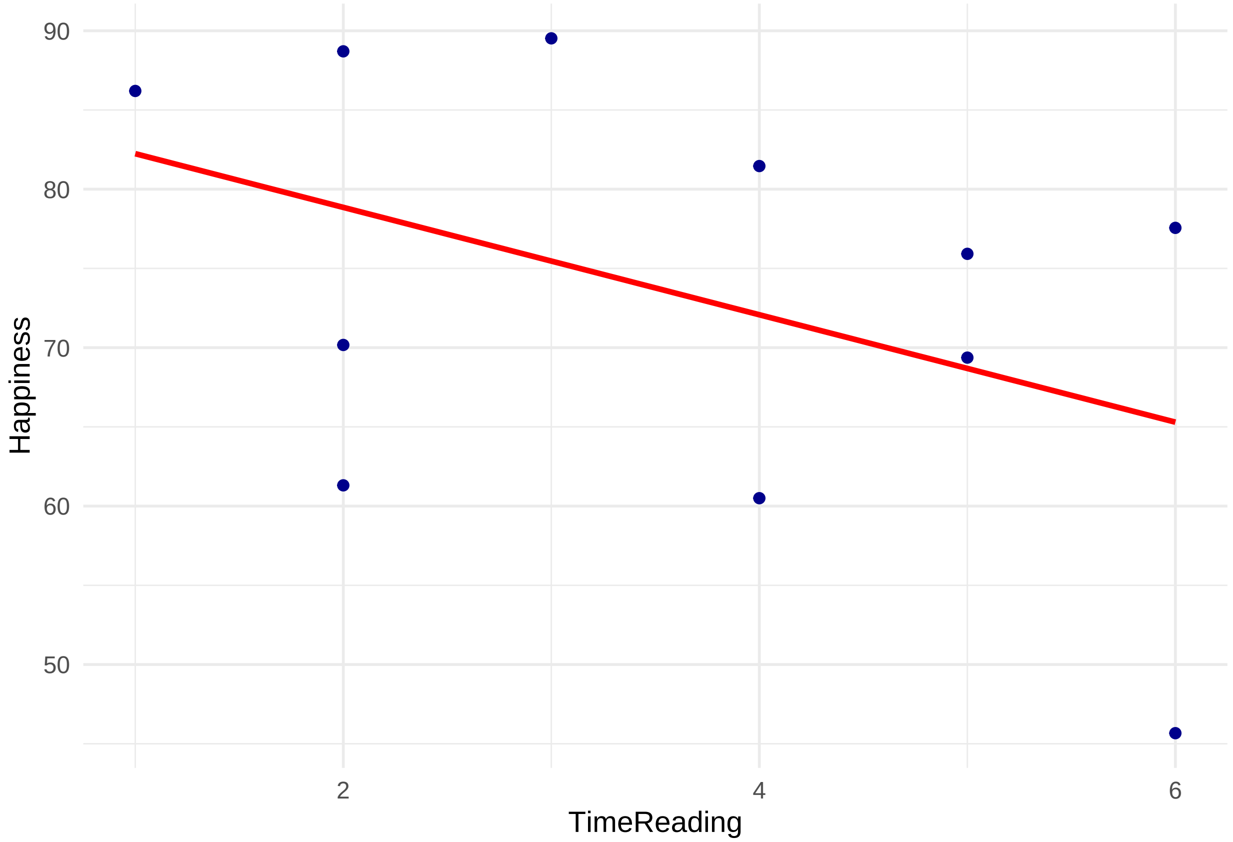
```
#Check scatter plot
ggplot(studentsurvey_df, aes(x=TimeTV, y=Happiness))+geom_point(color="darkblue")+stat_smooth(method=lm
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(studentsurvey_df, aes(x=TimeReading, y=Happiness))+geom_point(color="darkblue")+stat_smooth(meth
```

```
## `geom_smooth()` using formula 'y ~ x'
```



#Perform a correlation analysis of:

```
#All variables
cor(studentsurvey_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
```

```
#All Continous variable correlation
cor(studentsurvey_df[,c("TimeReading", "TimeTV","Happiness")])
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

```
#A single correlation between two a pair of the variables
cor.test(studentsurvey_df$TimeTV, studentsurvey_df$Happiness,method = "pearson")
```

```
## 
##  Pearson's product-moment correlation
```

```
##
## data:  studentsurvey_df$TimeTV and studentsurvey_df$Happiness
## 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
```

```
#Repeat your correlation test in step 2 but set the confidence interval at 99%
cor.test(studentsurvey_df$TimeTV, studentsurvey_df$Happiness, method = "pearson",conf.level = 0.99)
```

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

#Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

The pearson correlation coefficient of 0.636556 indicate positive association between TimeTv and Happiness which means that student spending more time watching TV have greater percentage rate happiness. The r=.636556 indicate a large effect because it is greater than .50. The p-value = 0.03521 which is considered statistically significant.

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

```
#correlation coefficient
cor (studentsurvey_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(studentsurvey_df)^2
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143
## TimeTV      0.779808529 1.0000000000 0.40520352 0.0000435161
## Happiness   0.189108726 0.4052035234 1.00000000 0.0246527174
## Gender      0.008035714 0.0000435161 0.02465272 1.0000000000
```

As determined above that there is a positive correlation between the Time spent watching TV and Happiness but the coefficient of determination only shows .405(40.5%) of the variability in happiness is explained by or shared by the Time spent watching TV but there is 59.5% of variance that can explain by other factors.

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

Correlation between Time Reading and Time TV is showing negative correlation which means that as the Time watching TV increases the Time spent Reading decreases. The coefficient of determination also shows that .7789(78%) of variability is shared by the two variable and only 22% of variable are from other factors.

#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.

```
library(ppcor)
```

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

```
pcor(studentsurvey_df)
```

```
## $estimate
##             TimeReading     TimeTV Happiness     Gender
## TimeReading   1.0000000 -0.8827973 0.4013124 -0.2706036
## TimeTV       -0.8827973  1.0000000 0.6311611 -0.2943135
## Happiness     0.4013124  0.6311611 1.0000000  0.2833152
## Gender       -0.2706036 -0.2943135 0.2833152  1.0000000
## 
## $p.value
##             TimeReading      TimeTV  Happiness    Gender
## TimeReading 0.000000000 0.001615344 0.28437887 0.4812716
## TimeTV      0.001615344 0.000000000 0.06832112 0.4420392
## Happiness   0.284378868 0.068321119 0.00000000 0.4600603
## Gender      0.481271572 0.442039185 0.46006033 0.0000000
## 
## $statistic
##             TimeReading     TimeTV Happiness     Gender
## TimeReading   0.0000000 -4.9720962 1.1592148 -0.7436966
## TimeTV       -4.9720962  0.0000000 2.1528933 -0.8147673
## Happiness     1.1592148  2.1528933 0.0000000  0.7816064
## Gender       -0.7436966 -0.8147673 0.7816064  0.0000000
## 
## $n
## [1] 11
## 
## $gp
## [1] 2
## 
## $method
## [1] "pearson"
```

```
#TimeTv and Happiness while controlling Gender
pcor.test(studentsurvey_df$TimeTV, studentsurvey_df$Happiness, studentsurvey_df$Gender)
```

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

I have picked TimeTV and Happiness varibles while the controlling variable is Gender. I noticed that the partial correlation between TimeTV and Happiness is 0.6311611 which is close to the correlation when the effect of Gender is not controlled (r=0.636556) however the p-value when Gender is controlled increases to 0.06832112 which mean that is less statistically significant compare to when effect of Gender is not controlled(p-value = 0.03521) which is considered to be statistically significant since it's value is <.05