Student Survey: Time spent reading vs Time spent Watching TV

Rajdeep Biswas

April 11th 2020

Libraries included

```
library(data.table)
library(dplyr)
library(ggplot2)
library(ggm)
```

Read the provided StudentSurvey.csv from current directory

```
raw_input_file_survey <- fread('student-survey.csv')
```

Quick verification

```
class(raw_input_file_survey)
summary(raw_input_file_survey)
str(raw_input_file_survey)
head(raw_input_file_survey)
```

Synopsis

We will analyze the results of a survey recently given to college students. 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 StudentSurvey.csv file.

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

Comparing multiple covariance

```
student_survey_covariance <- cov(raw_input_file_survey)
student_survey_covariance
```

```
##              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
```

```
student_survey_covariance_test_p <- cov(raw_input_file_survey$TimeReading, raw_input_f
ile_survey$TimeTV,method = "pearson")
sprintf("Covariance between TimeReading and TimeTV is: %f", student_survey_covariance_
test_p)
```

```
## [1] "Covariance between TimeReading and TimeTV is: -20.363636"
```

So the covariance between TimeReading and TimeTV is -20.35

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). 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.

This dependence on the scale of measurement is a problem because it means that we cannot compare covariances in an objective way – so, we cannot say whether a covariance is particularly large or small relative to another data set unless both data sets were measured in the same units.

The standardized covariance is known as a correlation coefficient. By standardizing the covariance we end up with a value that has to lie between −1 and +1

Hence from the covarance of TimeReading and TimeTV we can predict that they as the Time to watch TV increases, the time spent in reading decreases and vice versa. But to what degress or how strong the relationship is, is unknown to us just by measuring covariance.

For example if the TimeReading is converted into minutes the covariance would have been even bigger negative number, what would have happened if we have calculated in seconds or milliseconds? However even if the covariance would have increased (decreased in this case), we should not be saying it has built a stronger negative relationship since the variables are the same and we are just changing the units of measurement.

Summarizing covariance is a measure used to determine how much two variables change in tandem. The unit of covariance is a product of the units of the two variables. Covariance is affected by a change in scale. The value of covariance lies between -∞ and +∞.

Ref: Field, Andy. Discovering Statistics Using R. SAGE Publications.

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.

Examine the Survey data variables

```
glimpse(raw_input_file_survey)
```

```
## Observations: 11
## Variables: 4
## $ TimeReading <int> 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6
## $ TimeTV      <int> 90, 95, 85, 80, 75, 70, 75, 60, 65, 50, 70
## $ Happiness   <dbl> 86.20, 88.70, 70.17, 61.31, 89.52, 60.50, 81.46, 75.92,...
## $ Gender      <int> 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1
```

What measurement is being used for the variables

TimeReading, TimeTV, Happiness are continuous variables measured on different scales.

Without documentation on the collection method of the dataset an unbiased guess would be

TimeReading is getting measured in hours of a day

TimeTV is measured in minutes of a day

Happiness is measured in a scale of 100 points or percentage. (High meaning more)

Gender got addressed as 0 and 1. This is a categorical variable.

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.

```
student_survey_covariance_test_p <- cov(raw_input_file_survey$TimeReading, raw_input_f
ile_survey$TimeTV,method = "pearson")

sprintf("Covariance between TimeReading and TimeTV is: %f", student_survey_covariance_
test_p)
```

```
## [1] "Covariance between TimeReading and TimeTV is: -20.363636"
```

So, the covariance between TimeReading and TimeTV is -20.35

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). 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.

This dependence on the scale of measurement is a problem because it means that we cannot compare covariances in an objective way – so, we cannot say whether a covariance is particularly large or small relative to another data set unless both data sets were measured in the same units.

The standardized covariance is known as a correlation coefficient. By standardizing the covariance we end up with a value that has to lie between −1 and +1

Hence from the covarance of TimeReading and TimeTV we can predict that they as the Time to watch TV increases, the time spent in reading decreases and vice versa. But to what degress or how strong the relationship is, is unknown to us just by measuring covariance.

For example if the TimeReading is converted into minutes the covariance would have been even bigger negative number, what would have happened if we have calculated in seconds or milliseconds? However even if the covariance would have increased (decreased in this case), we should not be saying it has built a stronger negative relationship since the variables are the same and we are just changing the units of measurement.

Summarizing covariance is a measure used to determine how much two variables change in tandem. The unit of covariance is a product of the units of the two variables. Covariance is affected by a change in scale. The value of covariance lies between -∞ and +∞.

Few changes we can do to the measurements used

1. Standardize the unit of time for both TimeReading and TimeTV

2. Convert Happiness into an ordinal variable with a standard rule. For example <50 -> Unhappy, >=50 to <=65 -> Neutral, >65 to <=80 Happy, >80 Very happy

Regardless the solution as mentioned above is to standardized covariance which is known as correlation coefficient. To calculate the correlation, we have to divide the covariance of TimeReading and TimeTV by the product of the standard deviation of TimeReading and the standard deviation of TimeTV.

```
student_survey_covariance_test_p <- cor(raw_input_file_survey$TimeReading, raw_input_f
ile_survey$TimeTV)

sprintf("Correlation between TimeReading and TimeTV is: %f", student_survey_covariance
_test_p)
```

```
## [1] "Correlation between TimeReading and TimeTV is: -0.883068"
```

Since we know corrleation has a value that has to lie between −1 and +1, we can safely assume TimeReading and TimeTV are strongly negatively correlated with a value of -0.88. Note that by that statement I am not inferring a causal relationship.

Assumed Scale

Correlation coefficients between .10 and .29 represent a small association, coefficients between .30 and .49 represent a medium association, and coefficients of .50 and above represent a large association or relationship.

Ref: Field, Andy. Discovering Statistics Using R. SAGE Publications.

Ref: https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/ (https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/)

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?

Coorelation quantifies the strength (Correlation coefcient between -1 and 1) and direction (+/-) of a bivariate relationship.

Three commonly used correlation tests are Pearson(default in cor {stats}), Kendall and Spearman

Correlation for all variables

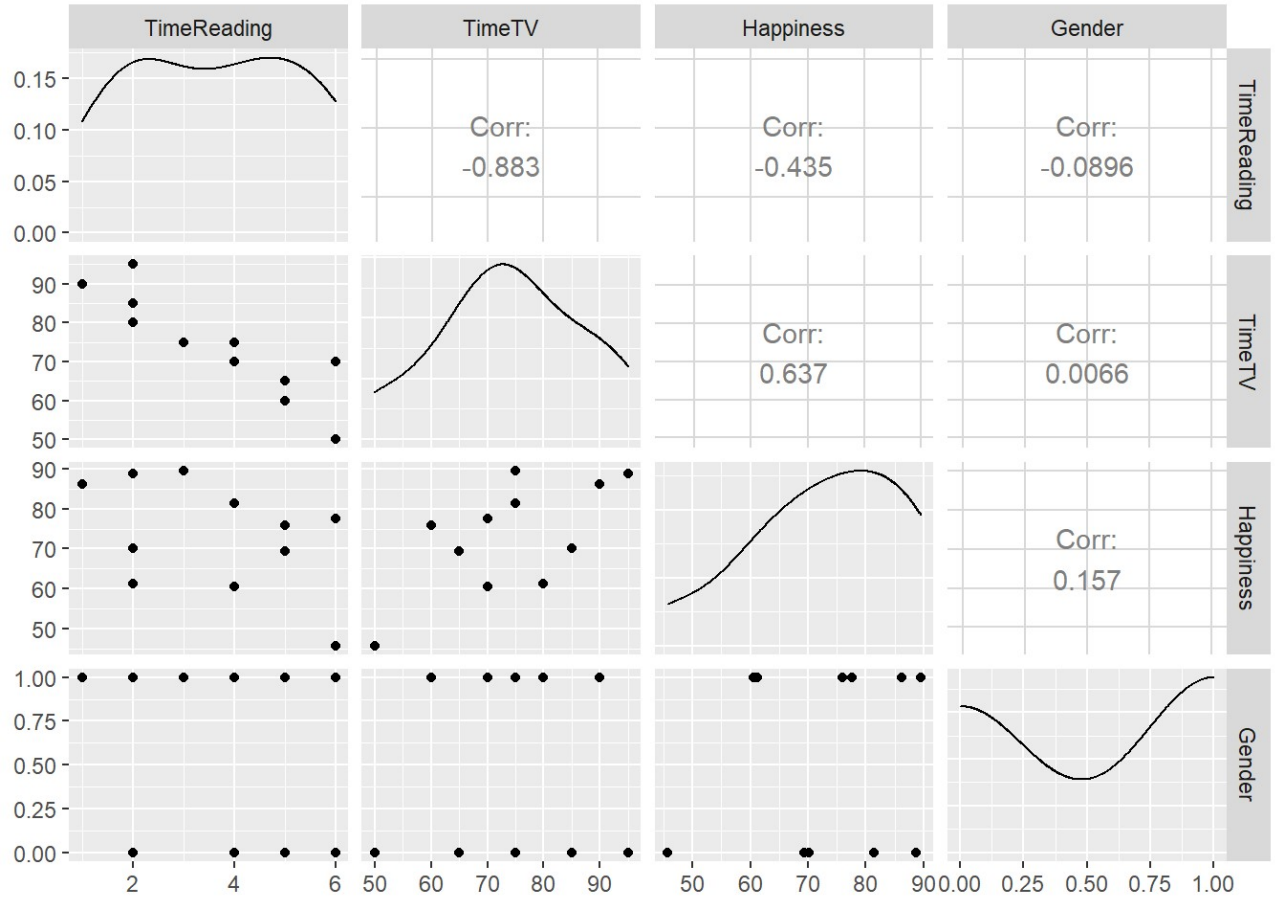
```
student_survey_correlation_all <- cor(raw_input_file_survey)
student_survey_correlation_all
```

```
##             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
```

Pairs plot of student survey data showing the relationship between each pair of variables as a scatterplot with the correlations printed as numbers.

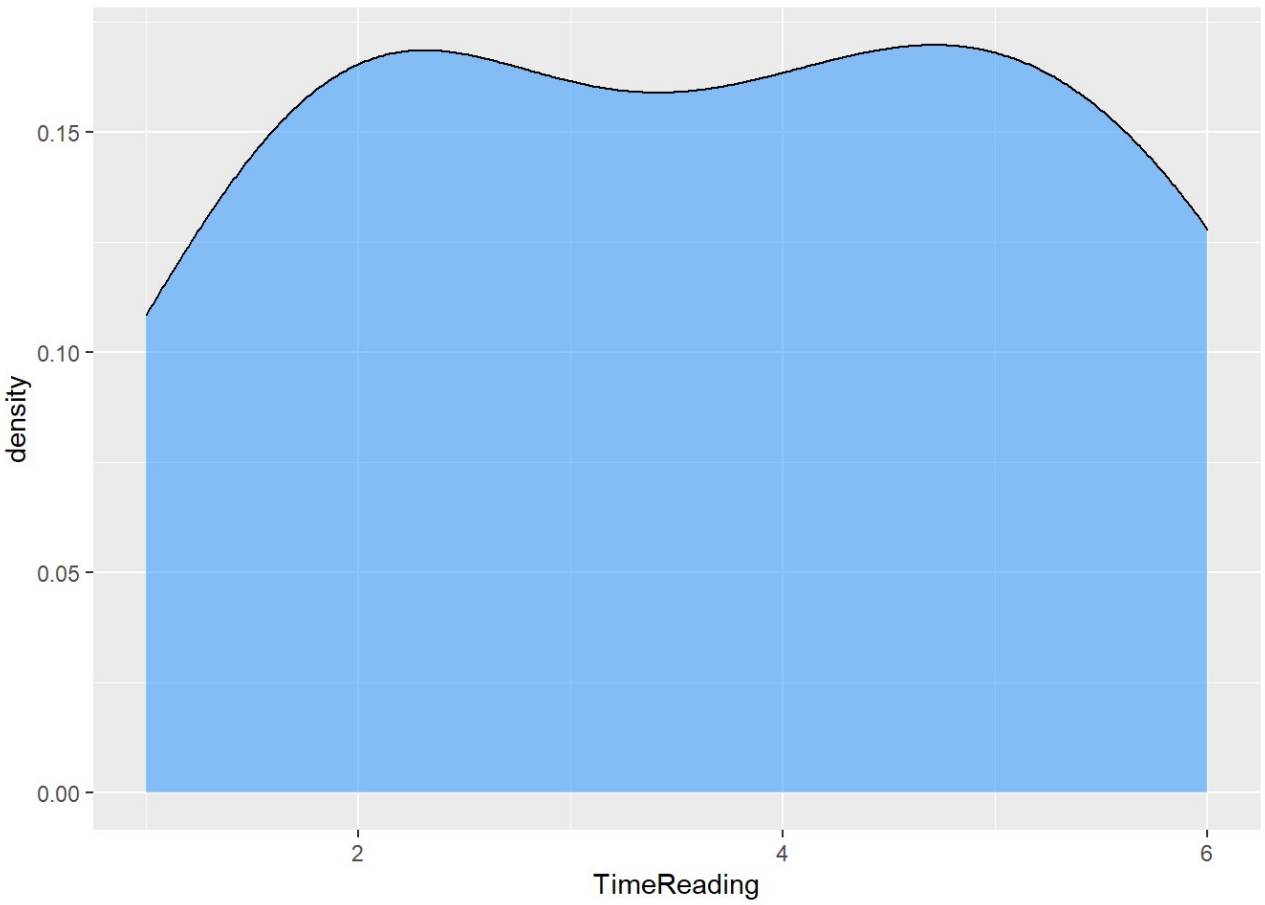
```
GGally::ggpairs(raw_input_file_survey[, c(1:4)])
```

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

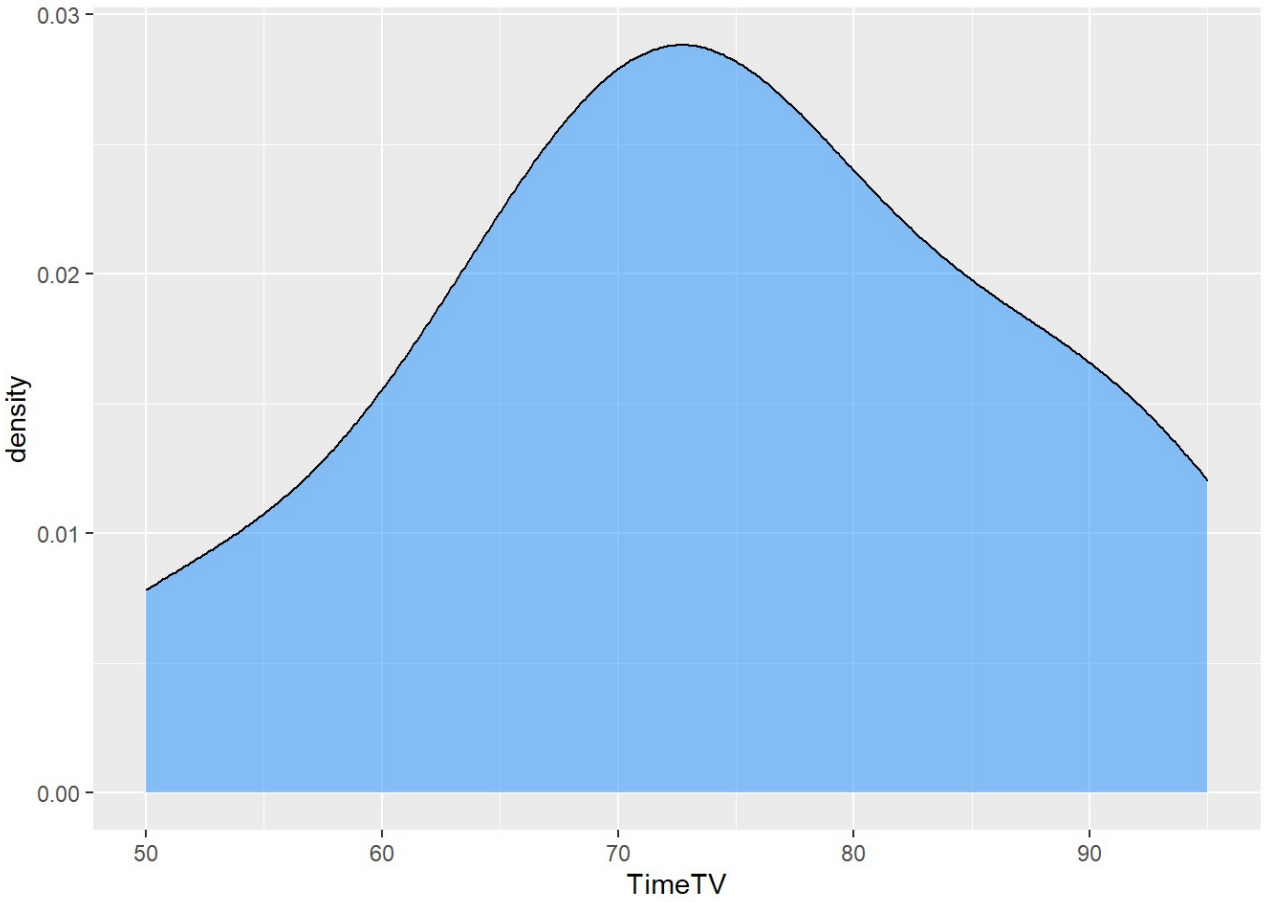


Distribution of variables

```
ggplot(raw_input_file_survey, aes(TimeReading)) +
  geom_density(fill="dodgerblue", alpha=0.5)
```




```
ggplot(raw_input_file_survey, aes(TimeTV)) +
  geom_density(fill="dodgerblue", alpha=0.5)
```



I am choosing to use Pearson correlation test for the below reasons:

Both variables are normally distributed.

Since the data provided is continuous (interval level) in nature, for the time spent reading and the time spent watching television .

The Non-parametric correlations are less powerful because they use less information in their calculations. In the case of Pearson's correlation, it uses information about the mean and deviation while non-parametric correlations like Kendall and Spearman use only the ordinal information and scores of pairs.

The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

The two variables have a linear relationship.

There is a strong Homoscedascity or equal variances as can be inferred from the earlier scatterplot.

```
student_survey_correlation_test_p <- cor.test(raw_input_file_survey$TimeReading, raw_i
nput_file_survey$TimeTV,method = "pearson")
student_survey_correlation_test_p
```

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

```
#student_survey_correlation_test_k <- cor.test(raw_input_file_survey$TimeReading, raw_
input_file_survey$TimeTV,method = "kendall")
#student_survey_correlation_test_k

#student_survey_correlation_test_s <- cor.test(raw_input_file_survey$TimeReading, raw_
input_file_survey$TimeTV,method = "spearman")
#student_survey_correlation_test_s
```

Perform a correlation analysis of:

All variables

```
student_survey_correlation_all <- cor(raw_input_file_survey)
student_survey_correlation_all
```

```
##             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
```

A single correlation between two of the variables

```
student_survey_correlation_test_p_95 <- cor.test(raw_input_file_survey$TimeReading, ra
w_input_file_survey$TimeTV, method = "pearson")
student_survey_correlation_test_p_95
```

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

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

```
student_survey_correlation_test_p_99 <- cor.test(raw_input_file_survey$TimeReading, ra
w_input_file_survey$TimeTV, method = "pearson", conf.level = 0.99)
student_survey_correlation_test_p_99
```

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

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

Correlation coefficient calculation at both 99% and 95% confidence intervals time spent on Tv and time spent on reading have a strong and significant negative correlation of -0.88. This means that increase in time spent on TV shows a decline in the time that students spend reading. Note that by that statement I am not inferring a causal relationship.

Both the 95% and 99% confidence does not cross zero. This tells us that in all likelihood, the population or actual value of the correlation is negative, so we can be pretty content that time spent on Tv and time spent on reading are, in reality, negatively related.

Each variable is perfectly correlated with itself (obviously) and so r = 1 along the diagonal of the table.

The Gender variable should have been excluded. So the Gender row and Gender column does not give us much in this context.

Interestingly Happiness is more strongly correlated with TimeTV (.63) than TimeReading (.43)

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

```
TimeReading_TimeTV_r = cor(raw_input_file_survey$TimeReading, raw_input_file_survey$Ti
meTV)
lm.fit = lm(raw_input_file_survey$TimeReading~raw_input_file_survey$TimeTV)
TimeReading_TimeTV_R2 = summary(lm.fit)$r.squared

# Print results
sprintf("Correlation coefficient = %f", TimeReading_TimeTV_r)
```

```
## [1] "Correlation coefficient = -0.883068"
```

```
## Correlation coefficient =  -0.883068

sprintf("Coefficient of determination= %f", TimeReading_TimeTV_R2)
```

```
## [1] "Coefficient of determination= 0.779809"
```

```
## Coefficient of determination =  0.779809
```

Since we know corrleation has a value that has to lie between −1 and +1, we can safely assume TimeReading and TimeTV are strongly negatively correlated with a value of -0.88. This means that increase in time spent on TV shows a decline in the time that students spend reading. Note that by that statement I am not inferring a causal relationship.

The Coefficient of determination at .78 or if we convert in terms of percentage 78% says that Time spent in reading accounts for 78% of the variability in the time spent in watching tv. However, although R2 is an extremely useful measure of the substantive importance of an effect, it cannot be used to infer causal relationships. TimeReading might well share 78% of the variation in TimeTV, but it does not necessarily cause this variation.

Another way to explain is the coefficient of determination represents the percent of the data that is the closest to the line of best fit. In this case 78% of the total variation in TimeTV can be explained by the linear relationship between TimeReading and TimeTV (as described by the regression equation). The other 22% of the total variation in TimeTV remains unexplained.

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

We can safely assume TimeReading and TimeTV are strongly negatively correlated with a value of correlation at -0.88. This means that increase in time spent on TV shows a decline in the time that students had spend reading. Note that by that statement I am not inferring a causal relationship.

The Coefficient of determination at .78 or if we convert in terms of percentage 78% says that Time spent in reading accounts for 78% of the variability in the time spent in watching tv. However, although R2 is an extremely useful measure of the substantive importance of an effect, it cannot be used to infer causal relationships. TimeReading might well share 78% of the variation in TimeTV, but it does not necessarily cause this variation.

In conclusion “correlation does not imply causation”

Thus there can be no conclusion made regarding the existence or the direction of a cause-and-effect relationship only from the fact that TimeReading and TimeTV are correlated. Determining whether there is an actual cause-and-effect relationship requires further investigation, even

when the relationship between TimeReading and TimeTV is statistically significant, a large effect size is observed, or a large part of the variance is explained.

Use TV Time and Happiness while controlling for Gender and perform a partial correlation. Explain how this changes your interpretation and explanation of the results.

```
pc_input_file_survey <- raw_input_file_survey[,c("TimeTV","Happiness","Gender") ]
glimpse(pc_input_file_survey)
```

```
## Observations: 11
## Variables: 3
## $ TimeTV    <int> 90, 95, 85, 80, 75, 70, 75, 60, 65, 50, 70
## $ Happiness <dbl> 86.20, 88.70, 70.17, 61.31, 89.52, 60.50, 81.46, 75.92, 6...
## $ Gender    <int> 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1
```

```
cor <- cor(pc_input_file_survey$TimeTV, pc_input_file_survey$Happiness)
cor #0.636556
```

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

```
R2_cor <- cor^2
R2_cor #0.4052035
```

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

To compute a partial correlation and its significance we will use the pcor() and pcor.test() functions respectively. These are part of the ggm package,

The general form of pcor() is: pcor(c(“var1”, “var2”, “control1”, “control2” etc.), var(dataframe))

Computing a first-order partial correlation:

```
pc <- pcor(c("TimeTV","Happiness","Gender"),var(pc_input_file_survey))
pc #0.6435158
```

```
## [1] 0.6435158
```

```
R2_pc <- pc^2
R2_pc #0.4141125
```

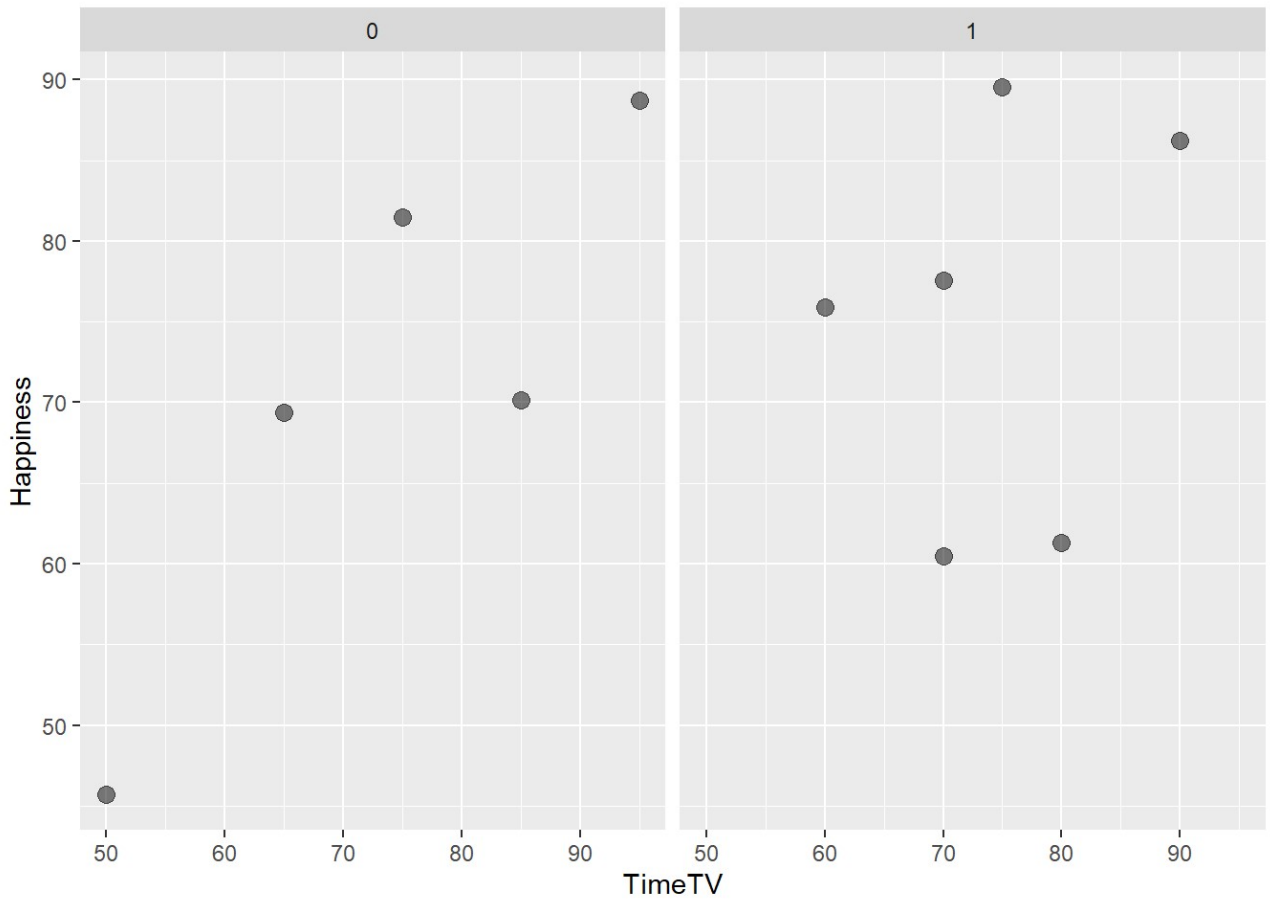
```
## [1] 0.4141125
```

The general form of pcor.test() is: pcor.test(pcor object, number of control variables, sample size)

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

```
## $tval
## [1] 2.377919
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.04469059
```

```
ggplot(pc_input_file_survey, aes(x = TimeTV,y=Happiness)) +
  geom_point( size = 3, alpha = 0.5) +
  #geom_smooth(aes(method="lm")) + #span too small. fewer data values than degrees of freedom.
  facet_grid(. ~ Gender)
```



One way to look at the correlation(.63) and the partial correlation(.64) is they remain almost the same. Hence controlling the Gender variable has insignificant effect on the strong positive relationship between TV Time and Happiness. Even the value of coefficient of determination in the case including the effect of Gender (0.41) and without it (0.40) almost remains the same. Hence we can safely say that roughly 40% of the variability in happiness is accounted by the time spent in watching tv irrespective of the gender.