
Title:
"Student
Survey
Assignment
Week
7
DSC520"
author:
"Saima
Rahmanzai"
date:
April
29,
2021
output:
pdf_document:
default

ASSIGNMENT OBJECTIVE

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

i. 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.

Answer: Covariance is a measure of correlation. Covariance can be used to measure the linear relationship between two variables in a dataset. A positive covariance value indicates a positive linear relationship between variables, a negative value represents the negative linear relationship. Zero

The results of my covariance tests show that there is negative but strong correlation between reading and watching TV (-.883), followed by another negative but a little less strong relationship between reading and happiness (-0.434), a positive close to strong (0.636) between happiness and watch-

Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeReadingandStdnt_Srvy_df$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

Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeReadingandStdnt_Srvy_df$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 Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeTVandStdnt_Srvy_df$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 Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeReadingandStdnt_Srvy_df$Gender t = -0.27001, df = 9, p-value = 0.7932 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.6543311 0.5392294 sample estimates: cor -0.08964215 Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeTVandStdnt_Srvy_df$Gender t = 0.01979, df = 9, p-value = 0.9846 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.5956354 0.6040812 sample estimates: cor 0.006596673

ii. 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.

Variables can be categorical, continuous or ordinal, etc. We need to select the correlation measurement based on the type of data. Usually Pearson correlation is used for parametric linear relationships and continuous variables. There are others like Spearman and Kendall correlations that

---

iii. 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?

As I did with my testing/analysis, before we look at the type of correlations to use, we should also look at the plots of our variables to get an idea of what to expect. In particular, we need to determine if it's reasonable to assume that our variables have linear relationships. I ran the scatterplot tests

Shapiro-Wilk normality test

data: Stdnt_Srvy_df$TimeTV

W = 0.98681, p-value = 0.9923

Shapiro-Wilk normality test

data: Stdnt_Srvy_df$TimeReading

W = 0.92093, p-value = 0.3265

iv. Perform a correlation analysis of:

1. All variables

Answer: I created the Correlation matrix to analyze the correlation between multiple variables at the same time. The command I used was as follows: cor(Stdnt_Srvy_df, method = “pearson”, use = “complete.obs”) The results are below: Table: Table with kable

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1.0000000	-0.8830677	-0.4348663	-0.0896421
TimeTV	-0.8830677	1.0000000	0.6365560	0.0065967
Happiness	-0.4348663	0.6365560	1.0000000	0.1570118
Gender	-0.0896421	0.0065967	0.1570118	1.0000000

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

Answer:
I ran the correlation between the following pairs using the following commands. The results were provided in the section i. above:
###TimeReading vs. TimeTV
cor.test(Stdnt_Srvy_df$TimeReading, Stdnt_Srvy_df$TimeTV, method = ("pearson"), use = "complete.obs")
###TimeReading vs. Happiness
cor.test(Stdnt_Srvy_df$TimeReading, Stdnt_Srvy_df$Happiness, method = ("pearson"), use = "complete.obs")
###TimeTV vs. Happiness

cor.test(Stdnt_Srvy_df$TimeTV$, $Stdnt_Srvy_df$Happiness, method = (“pearson”), use = “complete.obs”)

###TimeReading vs. Gender

cor.test(Stdnt_Srvy_df$TimeReading$, $Stdnt_Srvy_df$Gender, method = (“pearson”), use = “complete.obs”)

###TimeTV vs. Gender

cor.test(Stdnt_Srvy_df$TimeTV$, $Stdnt_Srvy_df$Gender, method = (“pearson”), use = “complete.obs”)

Please note that my analysis of gender correlation with other variables was not a good relationship. I used scatterplots as well as Spearman correlation and there does not seem to be a relationship and I will not consider this variable in my model.

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

Answer: The confidence interval is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or resample the population in the same way. The confidence level is the percentage of times you expect to reproduce an estimate between the

I repeated the code using 99% confidence interval on items 2 above. For the sake of illustration, I will show the first pair (TimeReading vs. TimeTV) and the code used to do this: The result was:

```
r
cor.test(Stdnt_Srvy_df$TimeReading,
Stdnt_Srvy_df$TimeTV,
alternative
=
c("greater"),
method
=
c("pearson"),
exact
=
NULL,
conf.level
=
0.99,
continuity
=
FALSE)
```

```
Pearson's
product-
moment
corre-
lation
```

data: Stdnt_Srvy_df$TimeReadingandStdnt_Srvy_df$TimeTV t = -5.6457, df = 9, p-value = 0.9998 alternative hypothesis: true correlation is greater than 0 99 percent confidence interval: -0.9763125 1.0000000 sample estimates: cor -0.8830677 The 95% Confidence Interval results showed the upper and lower limits of -.962 and -.602 compared to the 99% CI

The 99% results for the other variable pairs I analyzed are provided below: Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeReadingandStdnt_Srvy_df$Happiness t = -1.4488, df = 9, p-value = 0.9093 alternative hypothesis: true correlation is greater than 0 99 percent confidence interval: -0.8586992 1.0000000 sample estimates: cor -0.4348663 Pearson's product-moment correlation

data: Stdnt_Srvy_df$TimeTVandStdnt_Srvy_df$Happiness t = 2.4761, df = 9, p-value = 0.01761 alternative hypothesis: true correlation is greater than 0 99 percent confidence interval: -0.07001143 1.00000000 sample estimates: cor 0.636556

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

Answer: The correlation matrix is provided under section iv 1. above. A correlation matrix is a matrix that represents the pair correlation of all the variables. The cor () function returns a correlation matrix. The only difference with the bivariate correlation is we don't need to specify which

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

Correlation coefficients help quantify mutual relationships or connections between two things. How close is the data to the line of best fit? If points are far away, r (correlation coefficient) is close to 0. If very close to the line and moving upwards, it is close to +1, and if it is close to the line and sloping downwards, r is

Coefficient of Determination R2 tells how good is the model. It measures how well the predicted values match the observed values. +1 indicates that the predictions match the observations perfectly. R2=0, indicates that the predictions are as good as random guesses around the mean of the observed values. Nega-

The correlation coefficients are calculated for our paired variable under section i above. The coefficient of determination is calculated with the following formula:
model <- lm(TimeReading~TimeTV+Happiness, data=Stdnt_Srvy_df)
summary(model)
The result is below:
Call:
lm(formula = TimeReading ~ TimeTV + Happiness, data = Stdnt_Srvy_df)

Residuals:				
Min	1Q	Median	3Q	Max
-0.95879	-0.55984	-0.07737	0.25344	1.69455
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.62659	1.67194	6.954	0.000118
TimeTV	***-0.13501***	***0.02667***	***-5.061***	***0.000975***
Happiness	0.02746	0.02584	1.062	0.319059

Signif. codes: 0 ‘***’ ***0.001*** ‘**’ *0.01* ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8584 on 8 degrees of freedom Multiple R-squared: 0.807, Adjusted R-squared: 0.7588 F-statistic: 16.73 on 2 and 8 DF, p-value: 0.001386

This means that 80.7% of the variation in the TimeReading can be explained by the number of TimeTV and happiness.

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

Correct. As seen in section v above, the coefficient of determination validates that 80.7% of the variation in TimeReading can be explained by the two variables. Additionally, the other statistical measures we

performed above demonstrates there is a strong negative correlation between TimeTV and TimeReading. The scatterplots also shows this relationship between them (See 'Including Plots' section below)

vii. 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.

Answer: I used the following command to explaining the relationship between the three variables I picked: TimeTV, TimeReading and Happiness.

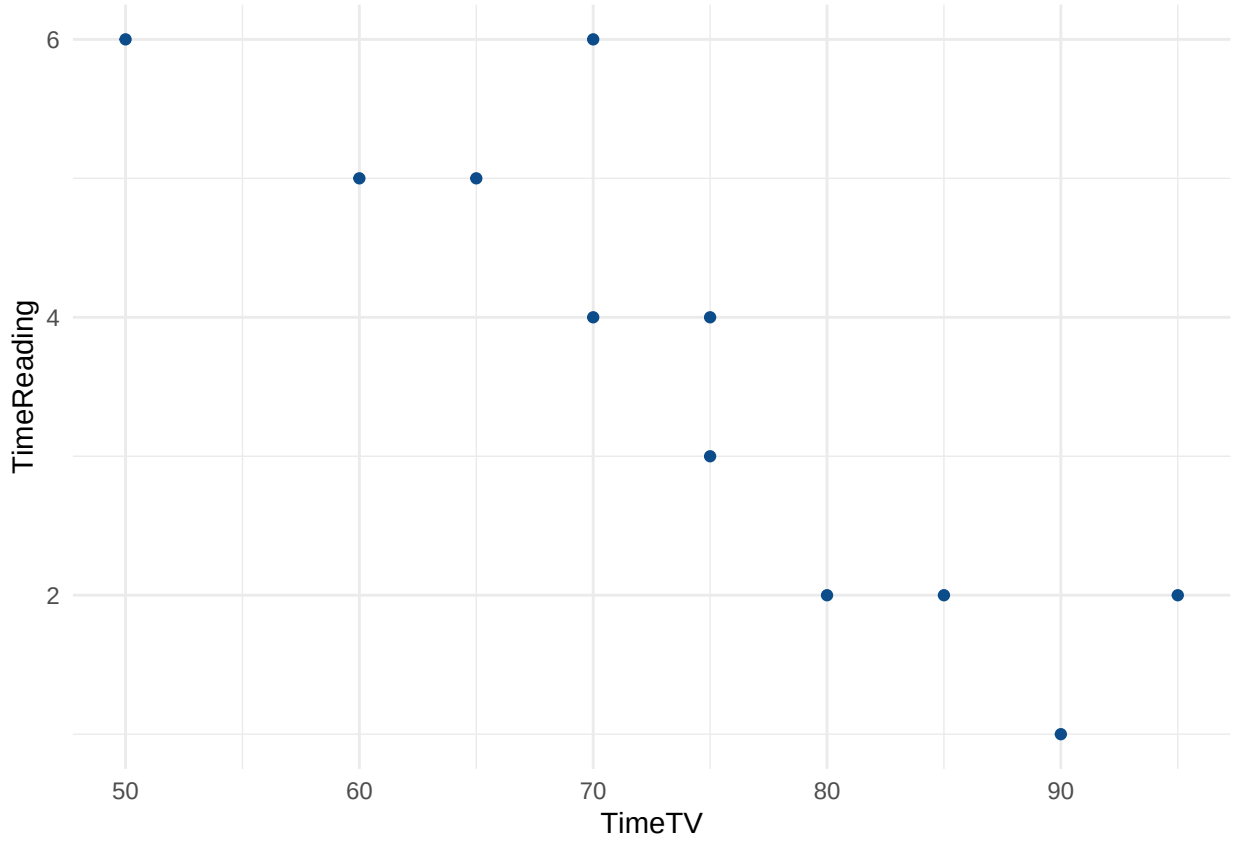
```
pcor.test(x=Stdnt_Srvy_df$TimeReading, y=Stdnt_Srvy_df$TimeTV, z=Stdnt_Srvy_df$Happiness)
```

estimate p.value statistic n gp Method 1 -0.872945 0.0009753126 -5.061434 11 1 pearson The results show that the p value is low (0.000975) that means the two variables (TimeReading and TimeTV) are partially correlated. Control variable is Happiness. The results show that the estimate value of -0.8729 Partial Correlation shows a strong but opposite direction correlation and the pValue being small suggests the relationship between them highly statistically significant. Happiness is a mediating variable and partially explains the correlation between the TimeReading and TimeTV variables.

Including Plots

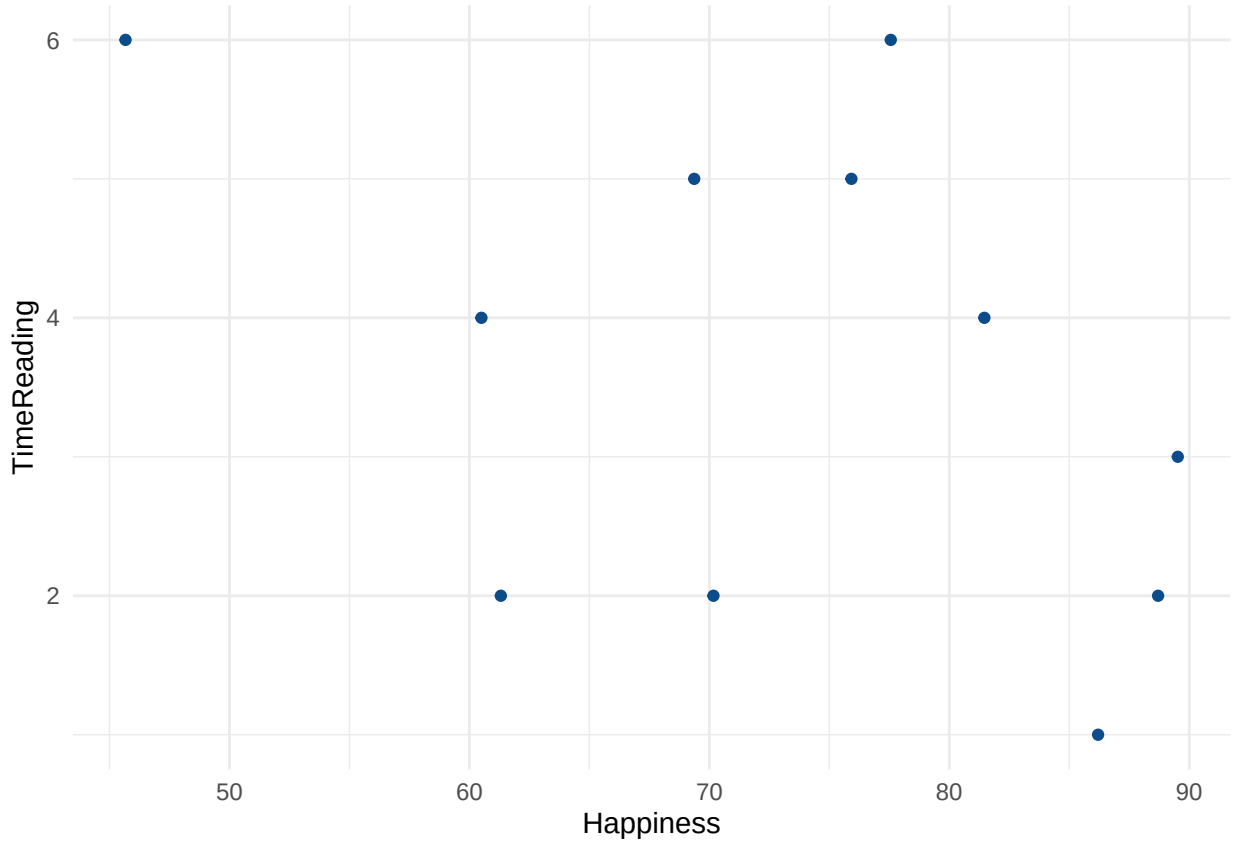
The scatterplots I generated to show relationships between two variables are provided below:

TimeTV vs. TimeReading



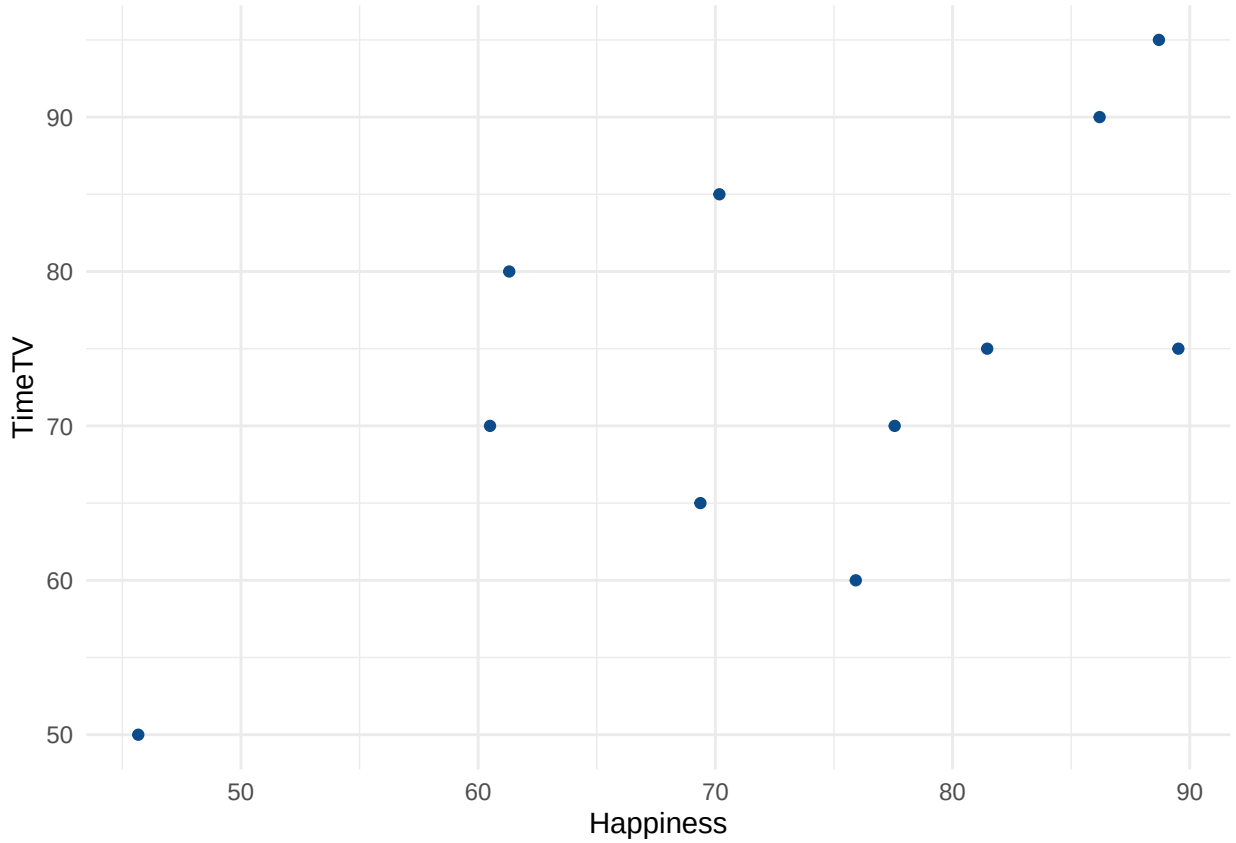
The above scatterplot between TimeTV and TimeReading shows a fairly strong negative relationship as sloping top left to bottom right.

Happiness vs. TimeReading



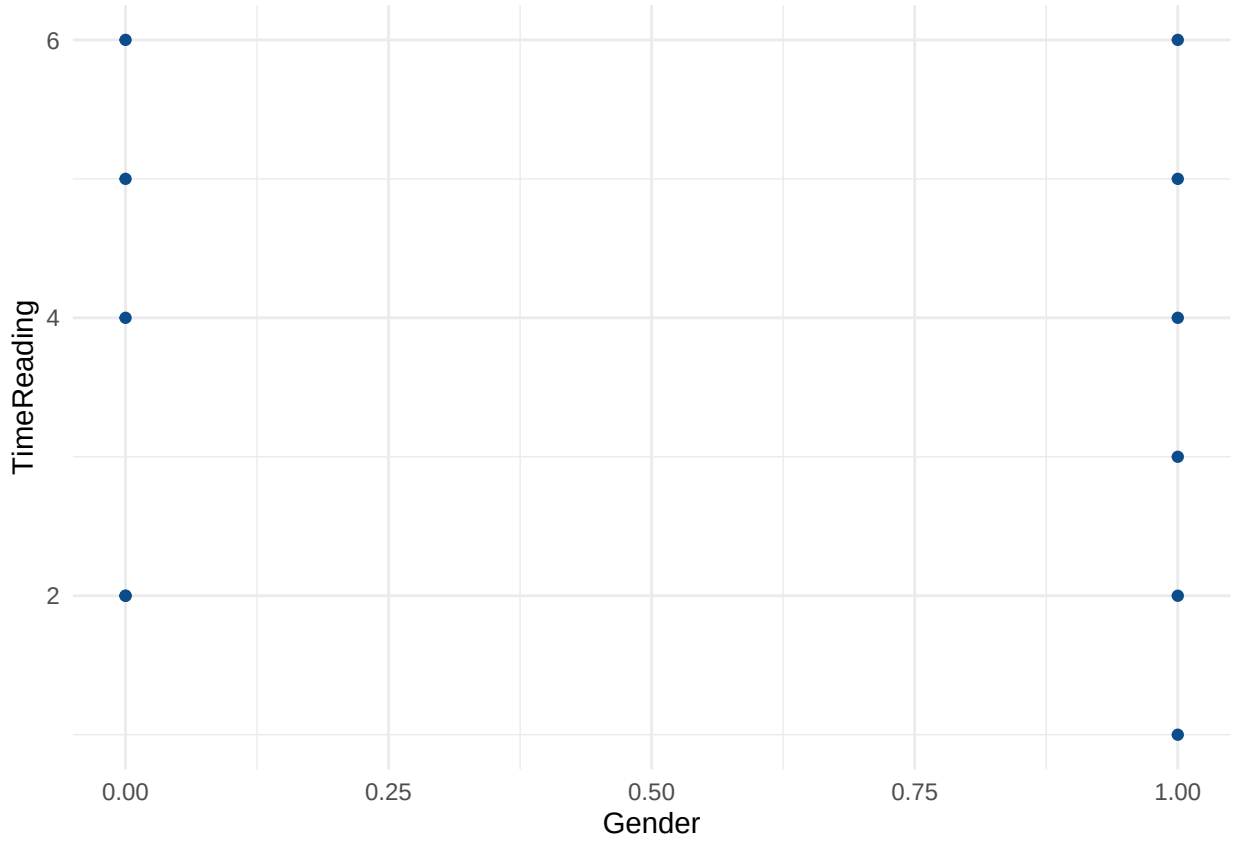
The above scatterplot between Happiness and TimeReading also shows somewhat a negative relationship as sloping top left to bottom right but data points are more scattered. Makes sense as the correlation coefficient above for these two variables is -0.434 compared to -0.883 which is much stronger and data more in closer to the straight line between TimeTV and TimeReading.

Happiness vs. TimeTV



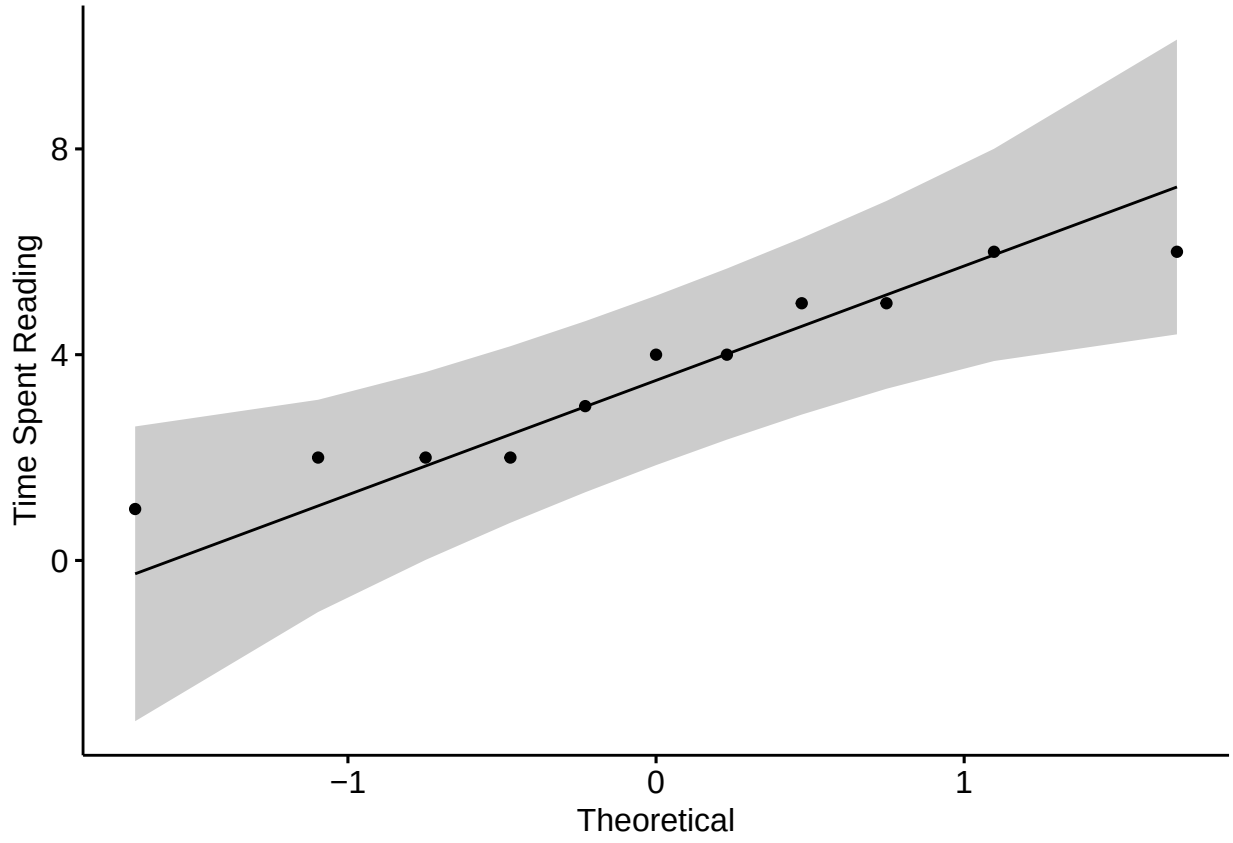
The above scatterplot between Happiness and TimeTV shows a positive relationship and data sloping up from lower left to top right. The correlation coefficient also suggests this positive relationship with 0.636. However not very strong. The scatterplot also shows spread of data around the path.

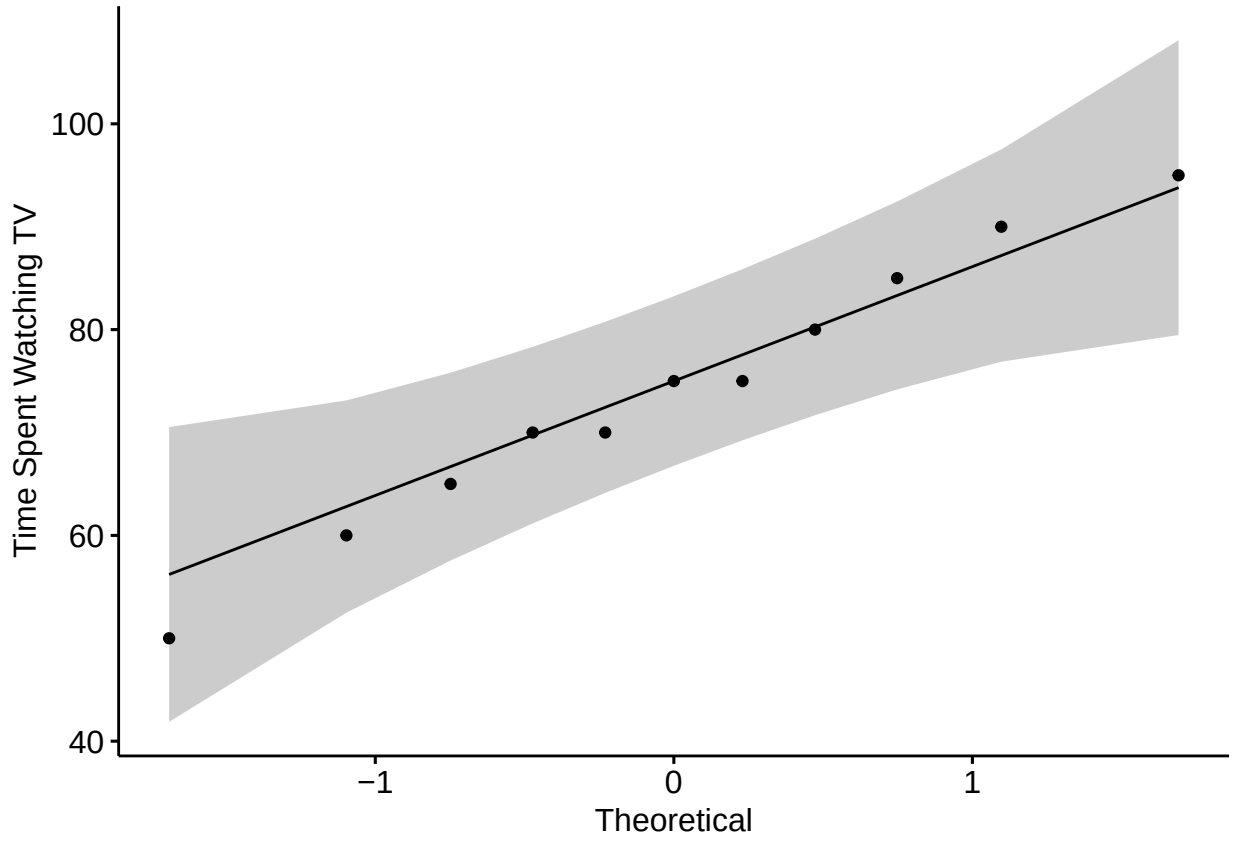
Gender vs. TimeReading

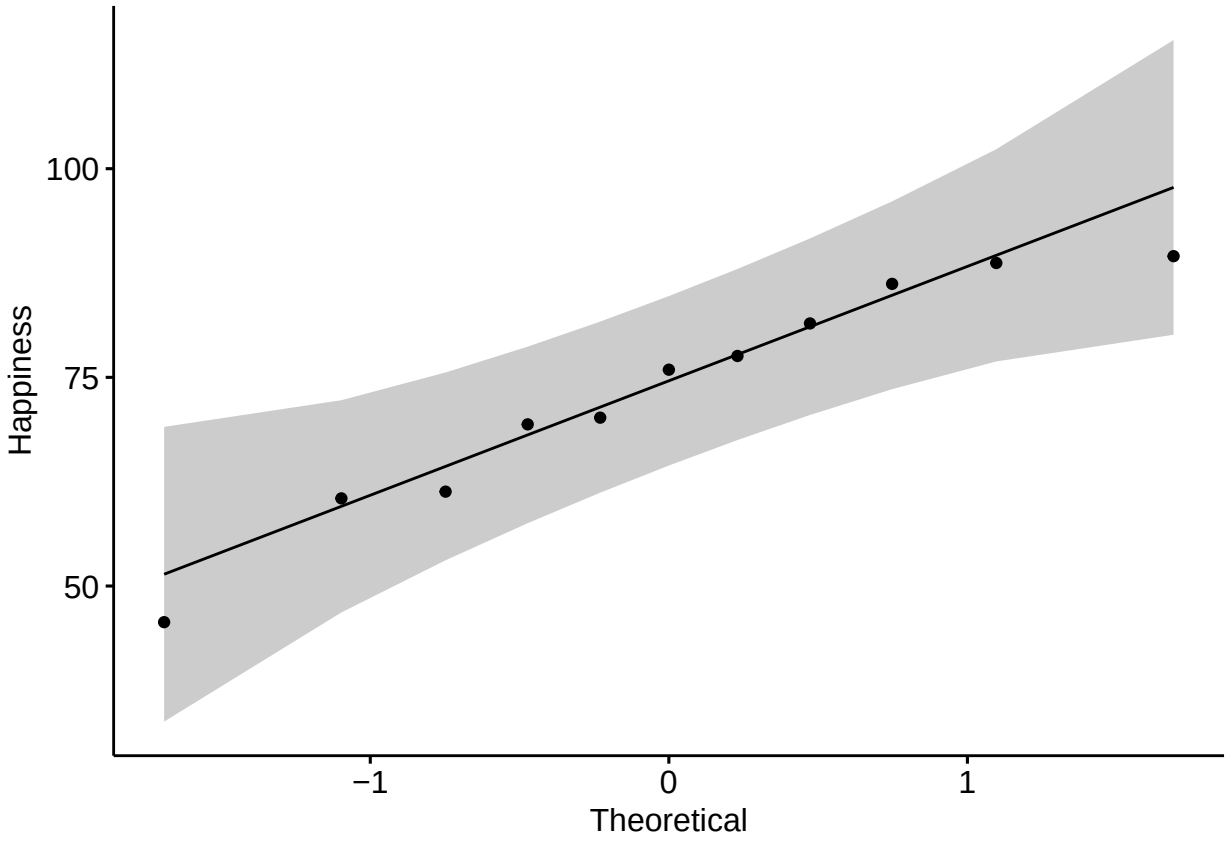


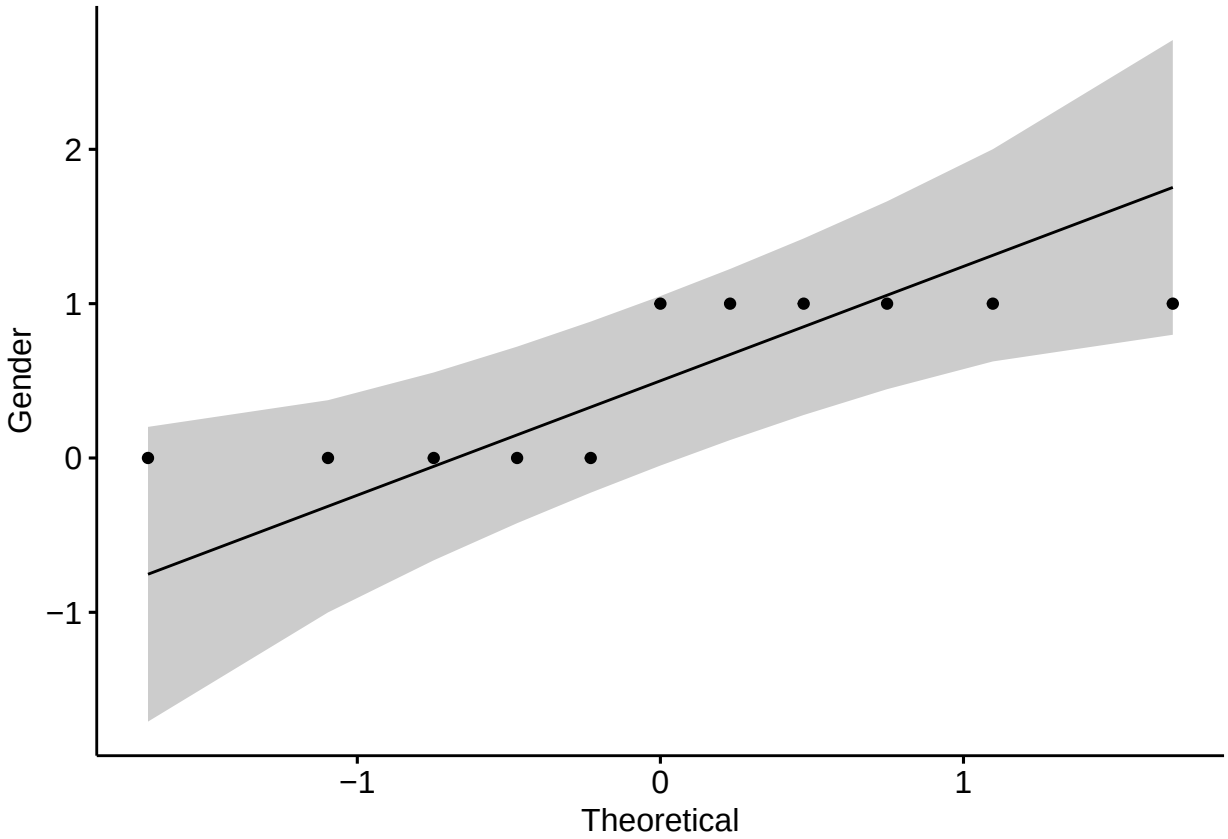
I created a scatterplot between Gender and TimeReading. There does not appear to be any linearity to the data. the coefficient correlation (Pearson method) shows -0.08 so hints towards not having a relationship. However, Gender appears to be ordinal data. We will run Spearman test as well as Pearson is not a good measure for nonlinear or ordinal data.

ggqqplots


Time Spent Reading
8
4
0
−1
0
1
Theoretical


Time Spent Watching TV
100
80
60
40
−1
0
1
Theoretical


Happiness
100
75
50
−1
0
1
Theoretical



The ggqqplots above show that the TimeReading, TimeTV and Happiness data are linear.