ASSIGNMENT

Yograj Karki

20121-04-29

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?"

For this analysis, I am going to use the student survey data. Let's have a quick look at the data.

```
##   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
```

I wanted to see the direction of the relationship between the variables, so I've performed the covariance calculation. Table below shows the calculated the covariance of all the variables in it.

```
##              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
```

From the covariance table above, it is evident that TimeReading and TimeTV has covariance score of -20.3636 which indicate the relationship is negative. Similarly, TimeReading and Happiness has the covariance of -10.35 which also indicates a negative relationship. On the contrary, here we can observe a positive relationship between TimeTV and Happiness with the covariance of 114.3772.

Examining the survey data variables

I'm going to create scatter plots matrix for each variables here excluding gender variable since it's a discrete categorical data.

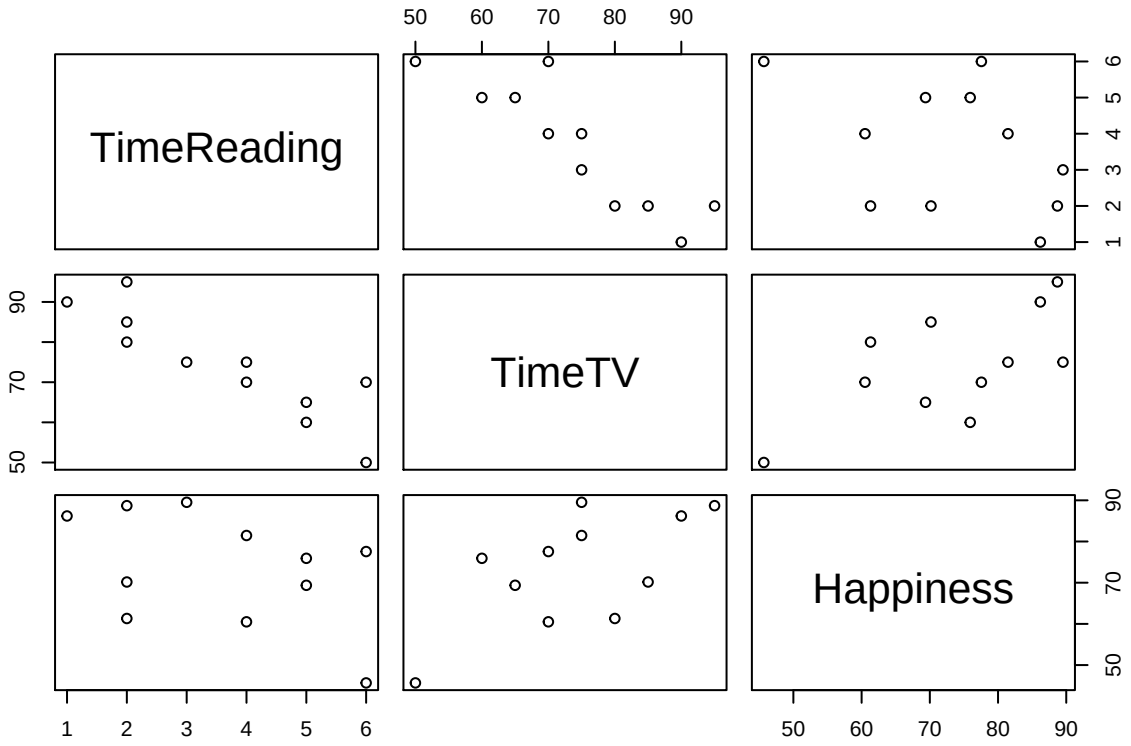

Figure 1: Scatter plot matrix

Correlation Test

Here with this data, I'm going to test the correlation with Spearman's rank correlation method because it seems there is monotonic relationship between variables according to covariances. The Spearman rank correlation test does not carry any assumptions about the distribution of the data. Since our sample size is very small, we cannot determine the normality of the distribution, Spearman's rank correlation test seems to do the just. Another reason for using Spearman's correlation test is because observations are paired, meaning data was collected from individual participants.

In terms of variable "Gender", it is a discrete categorical data with dichotomous values, so I will have to perform Point-Biserial correlation test with it.

Correlation Matrix

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.9072536 -0.4065196
## TimeTV       -0.9072536  1.0000000  0.5662159
## Happiness    -0.4065196  0.5662159  1.0000000
```

From the correlation matrix above, we can see that there is strong negative relationship between time spent watching TV and time spent reading. We can see moderately negative relationship between Reading Time and Happiness. Whereas there is moderately positive relationship between Happiness and time spent watching TV.

Single correlation test with a pair of variables

```
##
##  Spearman's rank correlation rho
##
## data:  data$TimeTV and data$TimeReading
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9072536
```

Single correlation test with a pair of variables and Confidence level of 99%

```
##
##  Spearman's rank correlation rho
##
## data:  data$TimeTV and data$TimeReading
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.9072536
```

Note: Spearman's rank correlation test does not show the confidence interval in R. Confidence Interval is shown only in Pearson product moment correlation coefficient calculation.

Calculating the coefficient of determination

```
# creating linear regression model
timereading_lm = lm(TimeReading ~ TimeTV, data = data)

# extracting r squared coefficient from the summary of regression model
summary(timereading_lm)$r.squared
```

```
## [1] 0.7798085
```

Here, We've already calculated that correlation coefficient of Reading time and TV time is -0.9072, which is a very strong negative relationship. And R-squared or coefficient of determination is found to be 0.7798 or 78%. This means that 77.98% of data fits the regression model or in other words, 77.98% of variation is Reading time is explained by the time spent on TV.

Conclusion

Based on the analyses above, we can conclude that there is a significant and negative relationship between the amount of time spent reading and the time spent watching television.

Partial Correlation Test

Partial correlation between TV watching time and Reading time while controlling Happiness score.

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

```
## Warning: package 'MASS' was built under R version 3.6.2
```

```
##     estimate      p.value statistic  n gp   Method
## 1 -0.8990805 0.0004011345 -5.808771 11  1 spearman
```

Partial correlation between Reading time and TV time is found to be -0.8990 which is not very different from actual correlation, so it can be concluded that happiness score has nothing to do with reading time and TV watching time.

Thank you.