Exercise 9 - Students Surevy Example

Ninad Patkhedkar

2020-10-01

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.

```
##   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
```

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.

```
pander(cov(student_survey_df), caption = "Covariance for Student Survey Attributes")
```

Table 1: Covariance for Student Survey Attributes

	TimeReading	TimeTV	Happiness	Gender
TimeReading	3.055	-20.36	-10.35	-0.08182
TimeTV	-20.36	174.1	114.4	0.04545
Happiness	-10.35	114.4	185.5	1.117
Gender	-0.08182	0.04545	1.117	0.2727

Covariance is a measurement of how closely related two variables are based on a linear relationship. In this example, we received a covariance value of -20 between `TimeTV` and `TimeReading`, meaning that for every hours of reading time a student adds to their daily routine, their daily consumption of television is reduced, indicating a negative correlation.

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.

`TimeReading`: This measurement seems to represent the number of hours each student spends reading a day.

TimeTV: This measurement represents the number of minutes each student spends watching TV a day.

Happiness: This seems to be some sort of measurement of each students happiness on some unknown scale. Having four siginificant figures would imply a high level of accuracy even though happiness isn't an easily measureable attribute.

Gender: This statistic represents the gender of each student but We don't which gender is represented by 0 or 1. This attribute should be converted to a factor

Changing the `TimeReading` and `TimeTV` attributes so they would both represent time in hours would reduce the covariance value.

```
cov(student_survey_df$TimeReading, student_survey_df$TimeTV / 60)
```

```
## [1] -0.3393939
```

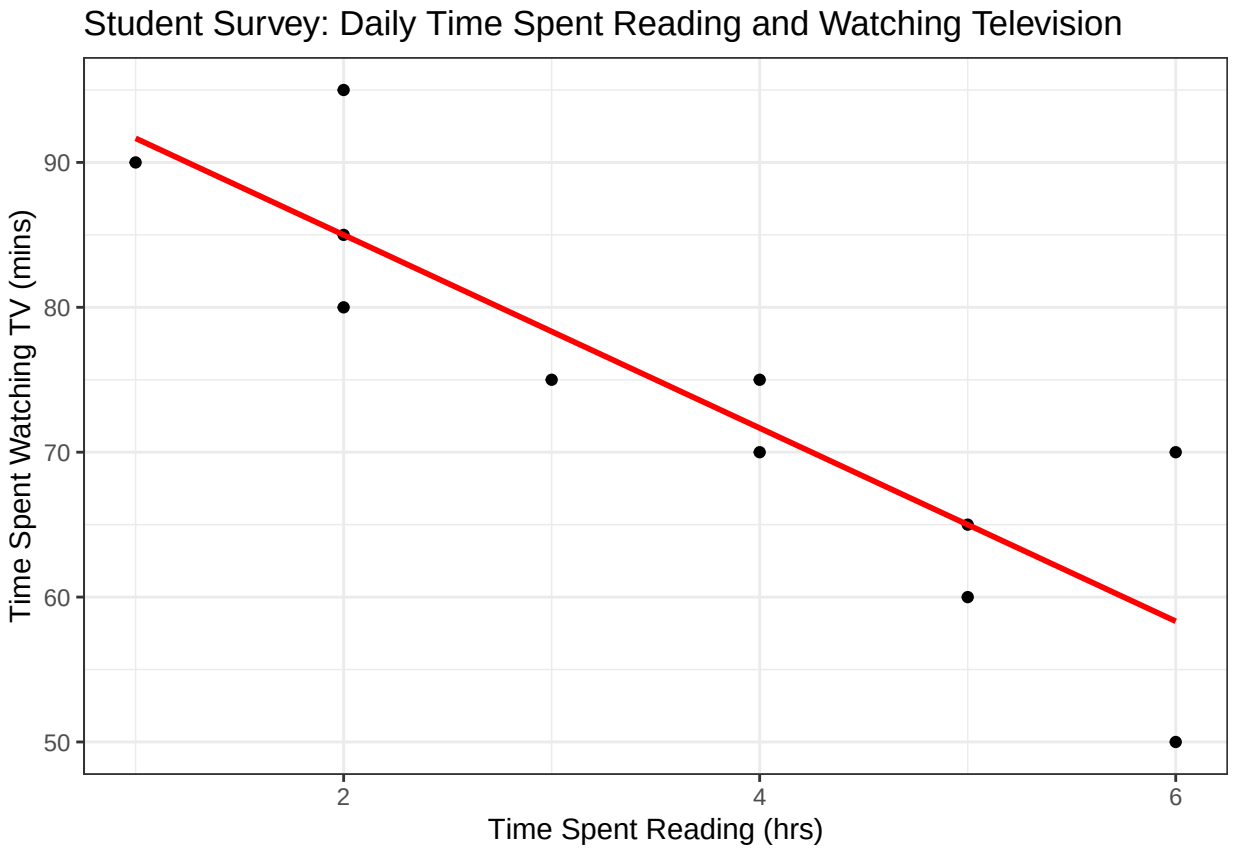
c) Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

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

```
##
##  Kendall's rank correlation tau
##
## data:  student_survey_df$TimeReading and student_survey_df$TimeTV
## z = -3.2768, p-value = 0.00105
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## -0.8045404
```

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

Pearson, Kendall, and Spearman correlation tests all return p-values less than 5%, indicating a high level of correlation. The correlation values, cor, tau, and rho are also very close to -1, indicating a high negative correlation. We can confirm this visually by plotting the data.



d) Perform a correlation analysis of:

1) All variables

```
pander(cor(student_survey_df),
       caption = "Student Survey Correlation Matrix")
```

Table 2: Student Survey Correlation Matrix

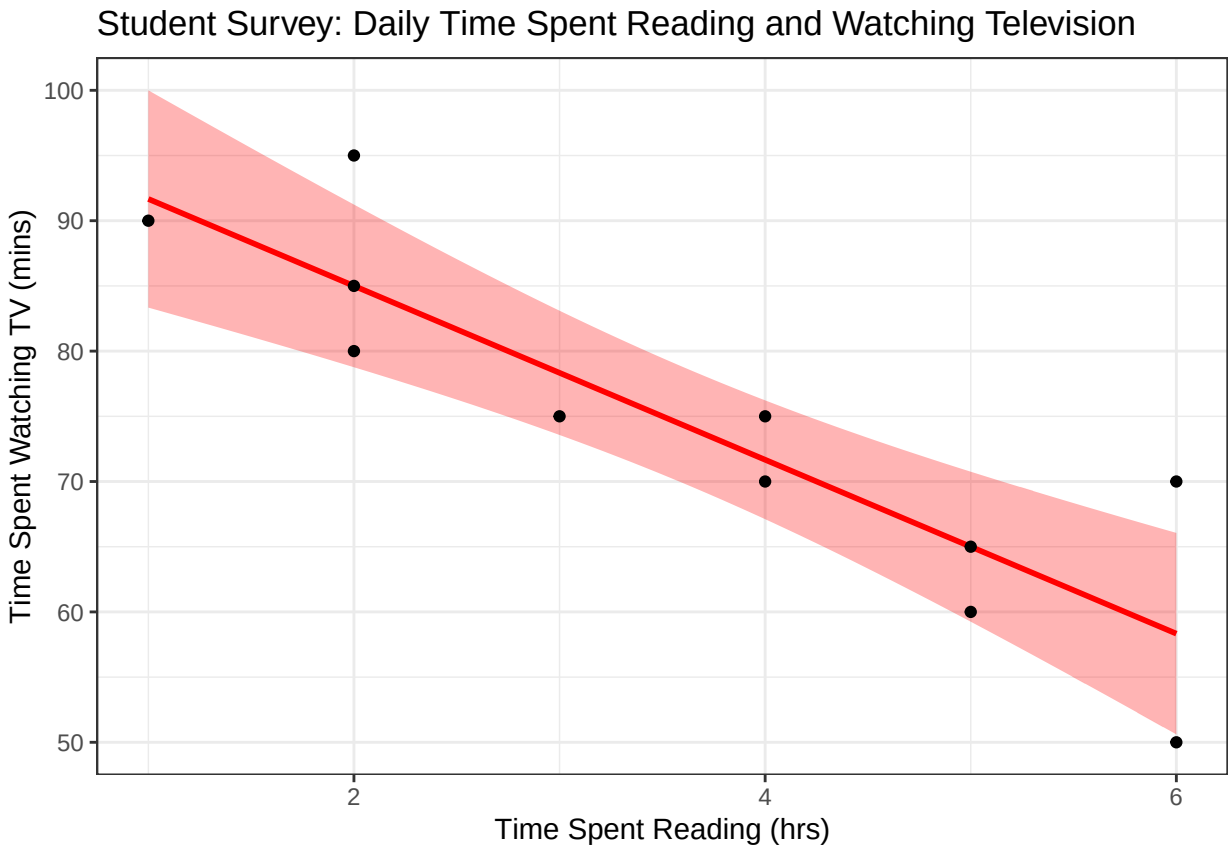
	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.8831	-0.4349	-0.08964
TimeTV	-0.8831	1	0.6366	0.006597
Happiness	-0.4349	0.6366	1	0.157
Gender	-0.08964	0.006597	0.157	1

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

```
cor.test(formula = ~ student_survey_df$TimeReading + student_survey_df$TimeTV,
         data = student_survey_df)
```

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

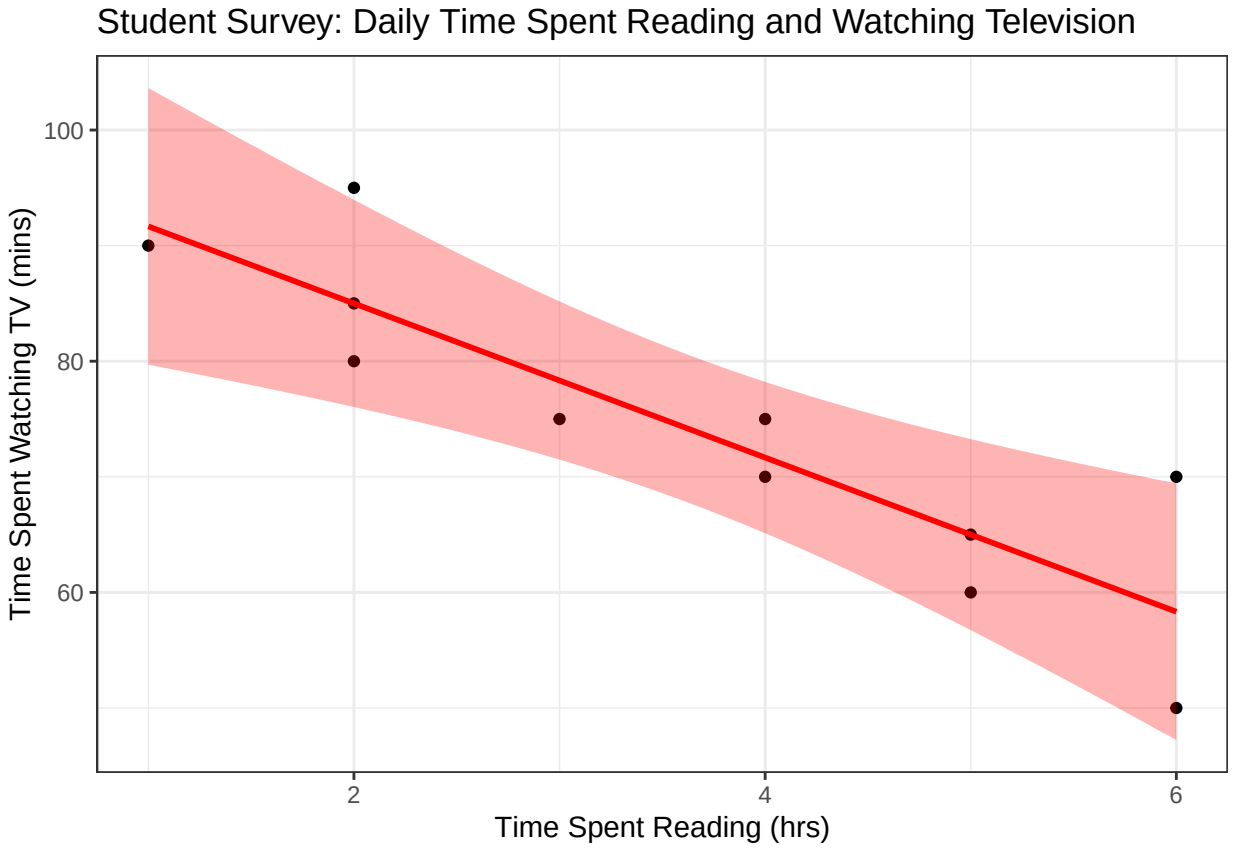
```
## data:  student_survey_df$TimeReading and student_survey_df$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9694145 -0.6021920
## sample estimates:
##        cor
## -0.8830677
```



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

```
cor.test(formula = ~ student_survey_df$TimeReading + student_survey_df$TimeTV,
         data = student_survey_df, conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey_df$TimeReading and student_survey_df$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 variables. Be specific with your explanation.

For correlaton, the closer the value is to 1 or -1, the stronger the variables are correlated. We see a correlation score of -0.88 between `TimeTV` and `TimeReading`, indicating that these two attributes are highly negatively correlated. `Gender` isn't strongly correlated with any other attributes. `Happiness` has a slight positive correlation with `TimeTV` and a slight negative correlation with `TimeReading`. Perhaps they should pick better books?

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

```
cor(student_survey_df$TimeReading, student_survey_df$TimeTV)
```

```
## [1] -0.8830677
```

```
lm_fit <- lm(TimeReading ~ TimeTV , data = student_survey_df)
summary(lm_fit)$r.squared
```

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

The correlation coefficient returned is -0.883, indicating a high level of negative correlation. The coefficident of determination, which is the correlation coefficeint squared, is 0.779, meaning that 77.9% of our data falls into our expected variance.

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

Based on the p-values we received from the Pearson, Spearman, and Kendall tests plus the correlation values we calculated, we can say with a high confidence that students who watch more TV spend less time reading.