Week 7 Assignment

Kimberly Adams

2022-07-21

Using `cor()` compute correlation coefficients for

height vs. earn:

```
## [1] 0.2418481
```

age vs. earn:

```
## [1] 0.08100297
```

ed vs. earn:

```
## [1] 0.3399765
```

Spurious correlation

The following is data on US spending on science, space, and technology in millions of today's dollars and Suicides by hanging strangulation and suffocation for the years 1999 to 2009.

Compute the correlation between these variables:

```
## [1] 0.9920817
```

This is a VERY strong correlation, but correlation does not imply causation. In this instance, the categorical spending did not necessarily (I won't rule it out as a possibility without further study) cause the population to want to commit suicide. Both values could be increasing for separate reasons (e.g. inflation and increasing social isolation).

Student Survey

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.

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.

```
##              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 represents the strength and direction of the relationship between two variables. In this data, [TimeReading and TimeTV] and [TimeReading and Happiness] have moderately strong negative relationships (as one increases the other decreases). On the flip side [TimeTV and Happiness] have a very strong positive relationship indicating that the more TV watches is often reflected in greater happiness. Gender had no real definable relationships with any other variable (perhaps a very weak positive relationship with happiness?).

Examine the Survey data variables.

What measurement is being used for the variables?

- TimeReading = interval number of days/hours/minutes(?) reading. Units unclear
- TimeTV = interval number of minutes or percentage(?) watching TV in 5 min increments
- Happiness = ratio contiguous value (percentage?)
- Gender = binary categorical variable (male/female) converted to a number (0/1). Value assignment unknown.

Explain what effect changing the measurement being used for the variables would have on the covariance calculation.

Changing the unit of measurement could have an impact on covariance if the scale of the values of either variable is altered.

Would this be a problem?

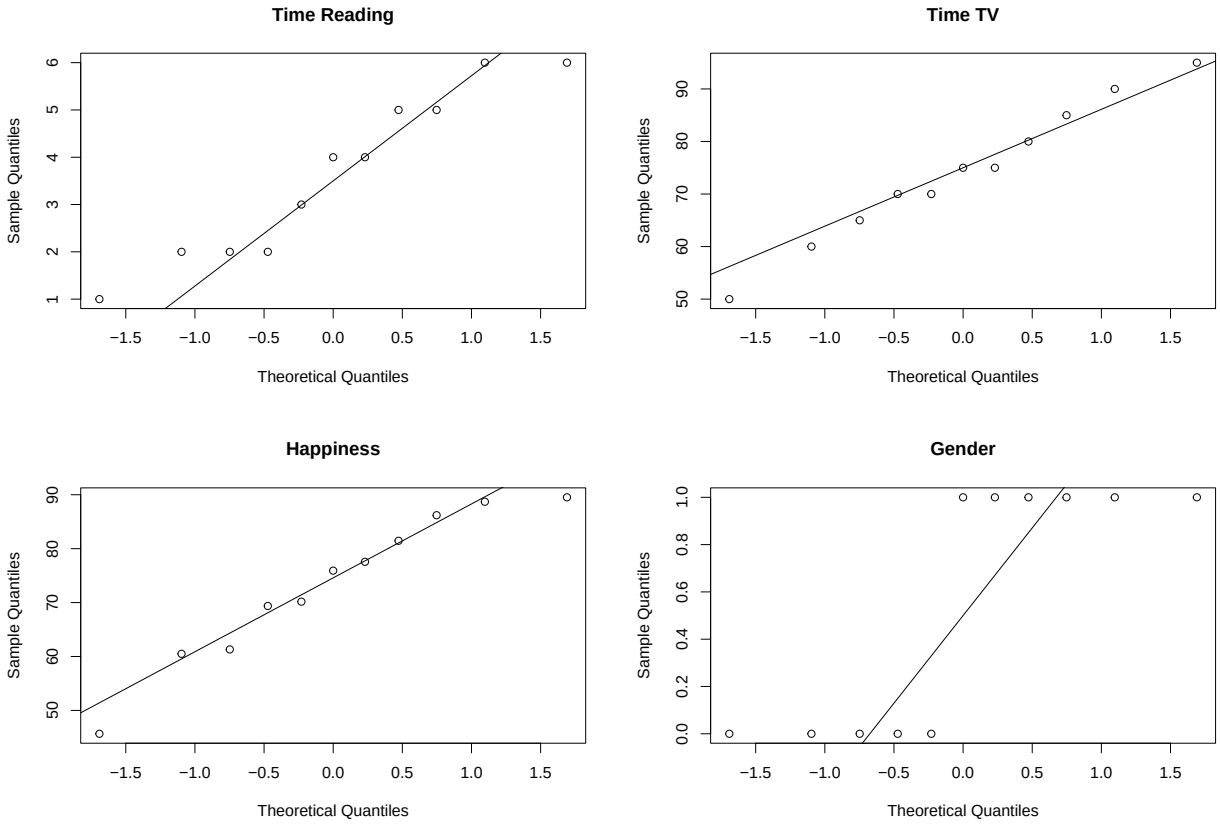
This is definitely a problem in the current dataset because we are unclear on what the units are for several variables such as TimeReading. It could be the number of hours per day or minutes per week. We do not have any idea without any sort of metadata explaining the measurement units.

Explain and provide a better alternative if needed.

A way to avoid the issue of units experienced by the covariance calculation is to use the unit-less correlation coefficient. This type of calculation standardizes each variable by dividing the covariance by the standard deviation of the variables thus removing the units. The resulting value falls between -1 and 1.

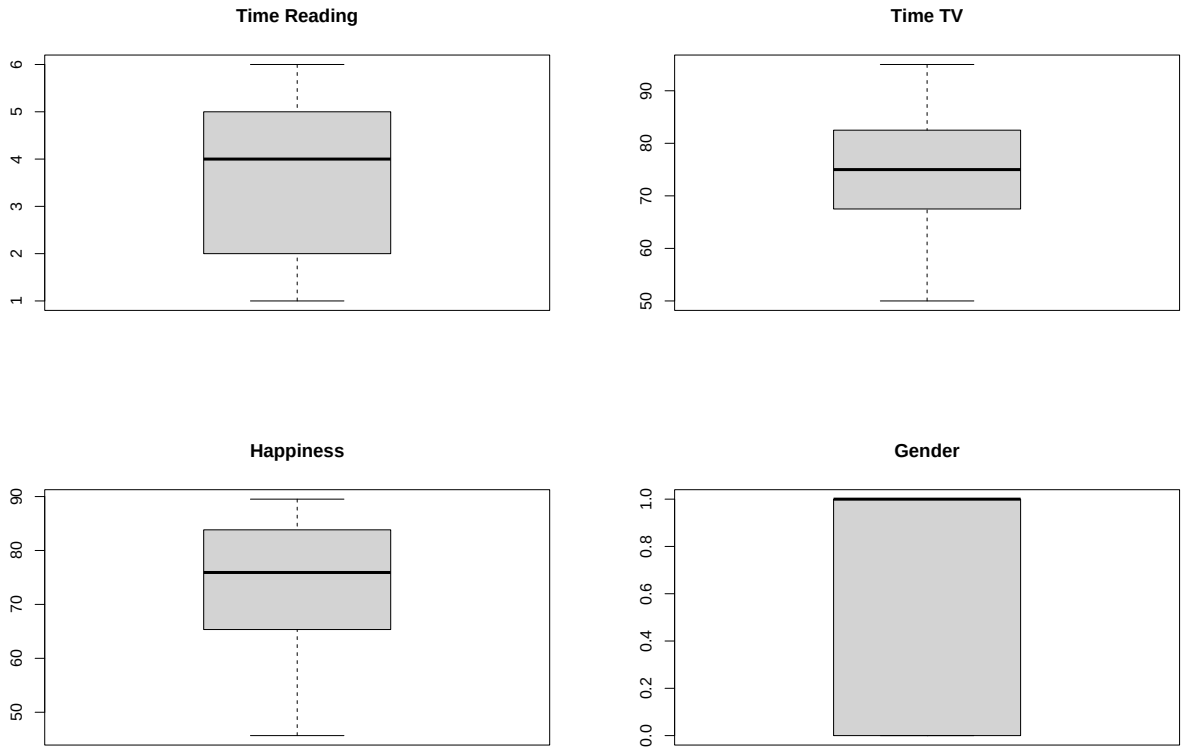
Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

Check for normal distribution of each variable:



Looks like, based on all the Q-Q plots, all variables are normally distributed except for Gender.

Check for outliers:



No outliers evident.

Types of data:

Excluding Gender (binary), all other variables of interest are interval (TimeReading, TimeTV) or ratio (Happiness).

Test Selection:

The data meets all the assumptions for the Pearson's Test:

- (Assumed random sampling)
- Normal distributions
- No Outliers
- Interval and ratio data (excluding Gender which has already been coded for point-biserial correlation)
- Predicted linear relationship between variables

Predictions:

I predict a positive correlation between TV watching and happiness and a negative correlation between TV watching and reading.

Perform a correlation analysis of:

All variables:

```
##             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
```

Time Watching TV and Time Reading:

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

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

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

The test returned a correlation value of -.88 between TV watching and Reading. As this value is below 0, it means there is negative linear correlation (one variable decreases as the other increases) and because it is closer to the lower limit (limit = -1), this is a very strong negative linear relationship between the variables. The relationship is also statistically significant at a 99% confidence interval (p = .0003).

The test returned a correlation value of .64 between TV watching and Happiness. As this value is above 0, it means there is positive linear correlation (both variables increase) and because it is closer to the upper limit (limit = 1), this is a strong positive linear relationship between the variables.

The test returned a correlation value of -.43 between TimeReading and Happiness. As this value is below 0, it means there is negative linear correlation (one variable decreases as the other increases) and because it is halfway to the lower limit (limit = -1), this is a moderately strong negative linear relationship between the variables.

The test returned a correlation value of -.09 between Gender and TimeReading. There basically no relationship between the variables.

The test returned a correlation value of .01 between Gender and TimeTV. There is basically no relationship between the variables.

The test returned a correlation value of .156 between Gender and Happiness. There is a slight linear relationship between the variables.

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

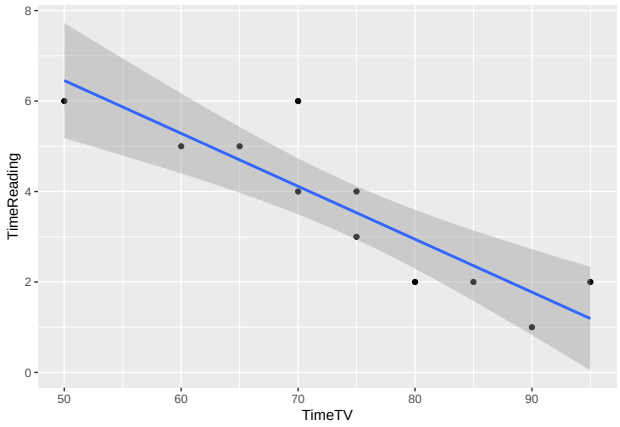
The correlation coefficient has already been calculated to be equal to -.88 between Time TV watching and Time Reading. This indicates a strong negative linear relationship between the two variables.

The coefficient of determination is equal to:

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

This indicates that the data points fall very close to the line of best fit, meaning that a linear model can predict it fairly well, but not perfectly as is reinforced by this graph of the data:

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



Notice that a few of the points are outside even the gray shaded area showing that the linear model isn't perfect in its prediction. The model can roughly explain 78% of the data, but leaves 22% unaccounted for.

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

Although there is a strong negative relationship between the two variables (-0.88), we can never directly say what causes something else to happen using correlation. We can only say that watching more TV is strongly associated with reading less. The actual cause may be something else such as limited amounts of time available or not having found a book that week that the student wants to read - perhaps the student might have binge-read several books the previous week and was taking a break. In some cases, the two activities could be done simultaneously. I have both watched TV and read a book at the same time.

Once again, the saying "Correlation does not imply causation" reminds us that even though there is a certain type of trend between two variables, that does not mean that the value of one directly caused or influenced the value of the other. The two variables could be reflecting a third unknown factor or just by chance follow a certain trend.

Pick three variables and perform a partial correlation, documenting which variable you are "controlling".

Looking at the correlation of TimeTV and TimeReading, controlling for Happiness:

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

```
##    estimate      p.value statistic  n gp  Method
## 1 -0.872945 0.0009753126 -5.061434 11  1 pearson
```

Explain how this changes your interpretation and explanation of the results.

The partial correlation value is still roughly -.88 so the initial analysis doesn't change. There is still a strong negative linear relationship between TimeTV and TimeReading and the strength of that relationship does not change even when controlling for Happiness value. This means that Happiness does not influence the direct relationship between TimeTV and TimeReading. The relationship is still statistically significant at a 99% confidence interval (p = .001).