Assignment 4.1 - Student Survey

Ramizuddin Mohammed Shabuddin

6/24/2020

# set the working directory

## [1] "E:/Bellevue/Semester3Assignment/dsc520/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

## 4.1.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.

cov(Student\_Survey)

## 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 is a measure to find the relationship between 2 variables or how much 2 random variables are linear variable. This helps in understanding, How one variable is related to other. Like, with increase in x value, the Y value increases or decreases. 1. Positive covariance indicates a positive linear relationship between the variables. 2. Negative covariance indicate a negative linear relationship between the variables.

From the above table, we can notice that TimeReading and TimeTV, TimeReading and Happiness are negative covariance.Whereas, TimeTV and Happiness are positive covariance. This concludes that students are more happy when watching TV which is affecting their study time.

##4.1.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.

## 'data.frame': 11 obs. of 4 variables:  
## $ TimeReading: int 1 2 2 2 3 4 4 5 5 6 ...  
## $ TimeTV : int 90 95 85 80 75 70 75 60 65 50 ...  
## $ Happiness : num 86.2 88.7 70.2 61.3 89.5 ...  
## $ Gender : int 1 0 0 1 1 1 0 1 0 0 ...

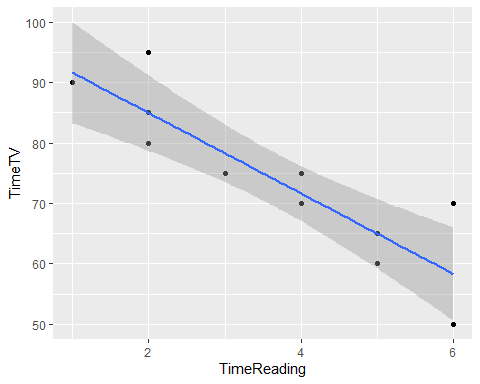
Based on the information above, the data are as below. 1. TimeReading is in integer format and values are in range of 1 till 6. I would think that these are in units of Hours. 2. TimeTV is in integer format and values are in range of 50 to 95. I would think that the people spend like 50 to 95 minutes watching TV. 3. Happiness is a number format. i think, the happiness is calculated to percentage. 4. Gender is integer with categorical field with values of either Male or Female (0 or 1)

So, The values are in different scale of units.Changing the scale of the measurement will change the covariance calculations. We need to make the values of same unit scale or standard unit scale. This process is called standardization

##4.1.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?

I will prefer the correlation method of ‘pearson’. As the pearson correlation measures a linear dependence between 2 variables. From the plot below, The TimeReading and TimeTV are negatively correlated.

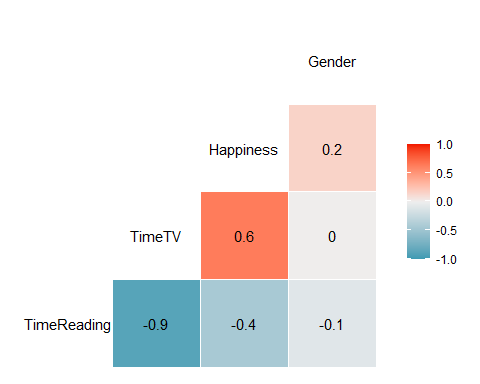
## `geom\_smooth()` using formula 'y ~ x'



##4.1.d Perform a correlation analysis of:

##4.1.d.1 All variables

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



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

##   
## Pearson's product-moment correlation  
##   
## data: Student\_Survey$TimeTV and Student\_Survey$TimeReading  
## 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

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

cor.test(Student\_Survey$TimeTV, Student\_Survey$TimeReading, conf.level = 0.99)

##   
## Pearson's product-moment correlation  
##   
## data: Student\_Survey$TimeTV and Student\_Survey$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

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

Based on the results of above question, we notice the following information. t = -5.6457, df = 9, p-value = 0.0003153 As the p-value is less than alpha. TimeReading and TimeTV have a strong negative correlation.

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

## 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

## TimeReading TimeTV Happiness Gender  
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143  
## TimeTV 0.779808529 1.0000000000 0.40520352 0.0000435161  
## Happiness 0.189108726 0.4052035234 1.00000000 0.0246527174  
## Gender 0.008035714 0.0000435161 0.02465272 1.0000000000

The TimeTV coefficient of determination Is 0.7798. The value indicates that TimeReading to TimeTV is significantly correlated with 77% variability.

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

The correlation coefficient of TimeTV to reading TimeReading is -0.88. The negative correlation indicates that relationship is in opposite directions.

#4.1.g 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.

## [1] 0.6435158

## [1] 0.4141125

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

The partial correlation between the Happiness and TV Time with Gender as controlled parameter. The partial correlation is 0.64. and square of the partial correlation is show as 0.41. This indicates around 41% of TV are happy. The Gender is impacting the happiness on TimeTV. As P-value(0.04) is less than 0.5.