|  |
| --- |
| 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 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. |
| ## 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 or around zero shows no relationship between them. |
| 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 watching TV. There appears to be no or close to no correlation between reading and gender (-0.089). I used pearson correlation as the data appears to be linear for most of these variables (except Gender) and is normally distributed. I did that by running additional statistical techiques including looking at scatterplots, performing ggqqplot and Shapiro tests. See the results of teh covariance of paired variables below: |
| Pearson’s product-moment correlation |
| data: Stdnt\_Srvy\_dfTimeTV 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\_dfHappiness 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\_dfHappiness 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\_dfGender 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\_dfGender 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 are nonparametric alternatives to Pearson’s correlation and for ordinal data. Gender appears to be ordinal data while others are continuous. Pearson’s is an excellent choice when you have data for a pair of variables and the relationship follows a straight line. If the data do not meet both of those requirements, it’s time to find a different correlation measure. Spearman’s rho is an excellent choice when we have ordinal data because Pearson’s is not appropriate correlation measure for that. Choosing a correlation test that is not fitting will not identify a strong relationship the two variables may have. In addition to ordinal data, if the data is curvilinear, it confuses Pearson and it underestimates the relationship’s strength. |
| ## 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 (See the “Including Plots” section below), the ggqqplot (See the “Including Plots” section below) and Shapiro tests below that confirms that TimeTV and TimeReading, TimeTV and Happiness, Happiness and Time Reading have linear relationships and hence Pearsons correlation fits better. There appears to be no relationship when comparing with Gender so we will not use that variable for relationship with other variables. |
| 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\_dfTimeTV, method = (“pearson”), use = “complete.obs”) |
| ###TimeReading vs. Happiness |
| cor.test(Stdnt\_Srvy\_dfHappiness, method = (“pearson”), use = “complete.obs”) |
| ###TimeTV vs. Happiness |
| cor.test(Stdnt\_Srvy\_dfHappiness, method = (“pearson”), use = “complete.obs”) |
| ###TimeReading vs. Gender |
| cor.test(Stdnt\_Srvy\_dfGender, method = (“pearson”), use = “complete.obs”) |
| ###TimeTV vs. Gender |
| cor.test(Stdnt\_Srvy\_dfGender, 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 re-sample the population in the same way. The confidence level is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval, and is set by the alpha value. The alpha value, or the threshold for statistical significance, is arbitrary – which value you use depends on your field of study. In most cases, researchers use an alpha of 0.05, which means that there is a less than 5% chance that the data being tested could have occurred under the null hypothesis. As well as the larger is the sample, the narrower is the confidence interval.Intuitively, the more observations we have, the better our estimates will be. |
| 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 correlation |
| data: Stdnt\_Srvy\_dfTimeTV 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 values of -.976 and 1.00, much more range. |
| The 99% results for the other variable pairs I analyzed are provided below: |
| Pearson’s product-moment correlation |
| data: Stdnt\_Srvy\_dfHappiness 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\_dfHappiness 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 variables. By default, R computes the correlation between all the variables. The bivariate Pearson Correlation produces a sample correlation coefficient, r, which measures the strength and direction of linear relationships between pairs of continuous variables. You can see perfect correlation of 1 when a variable is compared with itself.e.g. TimeReading compared to TimeReading is 1.00 etc. At a glance, we can see strong negative correlation between TimeReading and TimeTV, Next stronger negative correlation is TimeTV and Happiness which is a positive correlation of 0.636, followed by a negative correlation of Happiness with timeReading (-0.434). Gender as you can see does not have a great correlation with other variables but with itself. |
| ## 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 close to -1. In other words, This correlation coefficient is a single number that measures both the strength and direction of the linear relationship between two variables. Values can range from -1 to +1. The greater the absolute value of the correlation coefficient, the stronger the relationship. The extreme values of -1 and 1 indicate a perfectly linear relationship where a change in one variable is accompanied by a perfectly consistent change in the other. A coefficient of zero represents no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease. The sign of the correlation coefficient represents the direction of the relationship. Positive coefficients indicate that when the value of one variable increases, the value of the other variable also tends to increase. Positive relationships produce an upward slope on a scatterplot. Negative coefficients represent cases when the value of one variable increases, the value of the other variable tends to decrease. Negative relationships produce a downward slope. |
| 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. Negative R2 indicates that the predictions are worse than random. Since R2 indicates the distance of points from 1:1 line, it does depend on the magnitude of the numbers (unlike r2). |
| The correlation coefficients are calculated for our paired variable under section i above. The coefficient of determination is calculated below: |
| r model <- lm(TimeReading~TimeTV+Happiness, data=Stdnt\_Srvy\_df) summary(model) |
| 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.

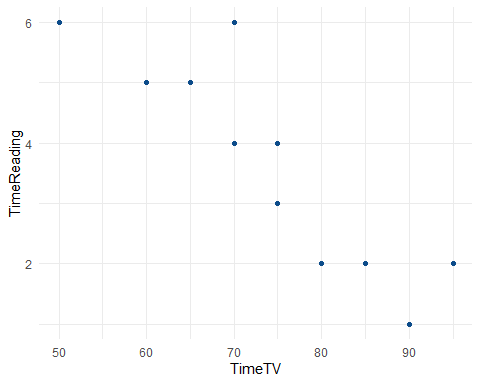
pcor.test(x=Stdnt\_Srvy\_df$TimeReading, y=Stdnt\_Srvy\_df$TimeTV, z=Stdnt\_Srvy\_df$Happiness)

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.

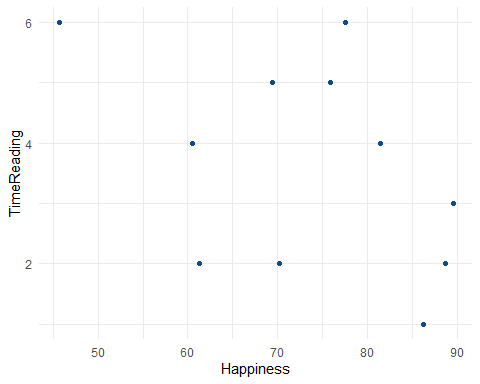
## Including Plots

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

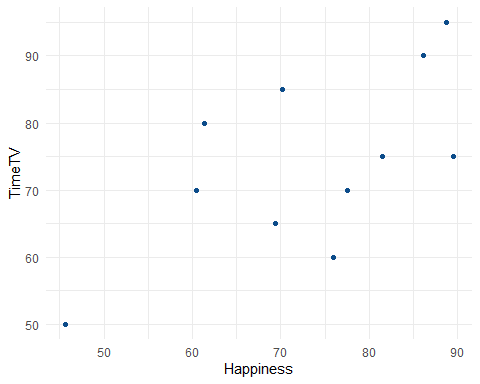
### TimeTV vs. TimeReading

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

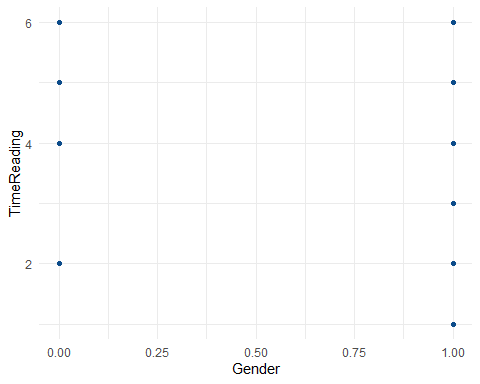
### Happiness vs. TimeReading

 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.

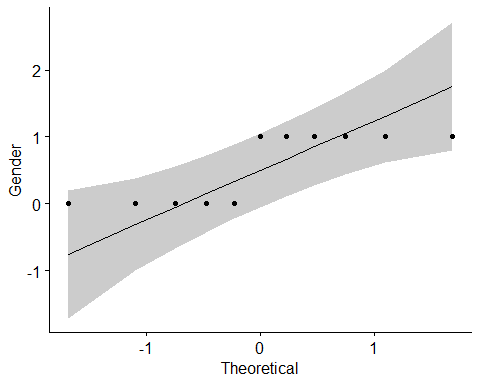
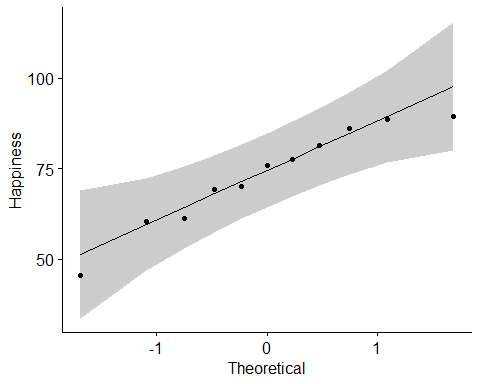
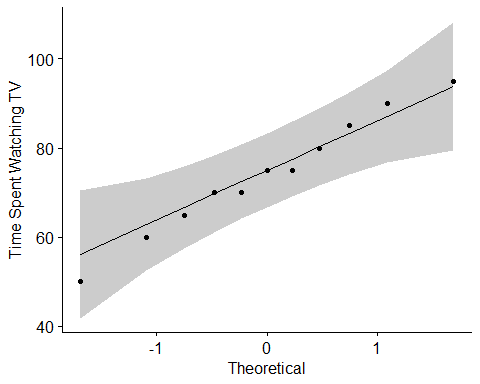
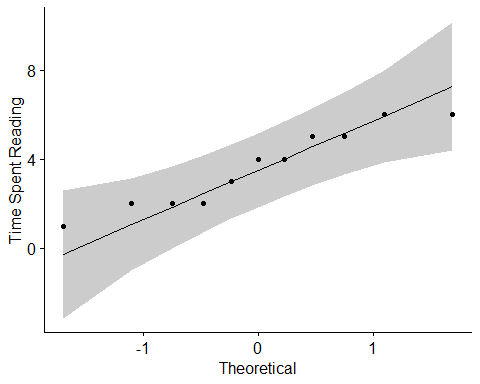
### Happiness vs. TimeTV

 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

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