Brooke Woods

DSCI5180

I gathered a dataset from Source UCI Learning machine. It is called “student performance data” and contains information about two Portuguese schools. There are 30 attributes which consist of student grades, demographic, social and education related features. This survey was collected by using school reports and questionnaires. All results shown in excel.

***What is the distribution of total grades? What proportion of students passed the exam?***

**\*\*\*Check Appendix A\*\*\***

The table on the left of Appendix A shows the percentile for any given score (ex-a score of 19 falls in the 99.80percentile while a score of 17 falls in the 93% percentile etc.) The table on the right of Appendix A counts the number of students who qualify for each letter grade.

*The proportion of students who had a passing final grade was .2994.* (number of students who passed with 70 or higher/ total number of students) = 194/648.

* ***What is the 95 % Confidence Interval for the mean age of a student?***

A picture containing text, black, orange

Description automatically generated

**Interpretation**: I am 95% confident the mean student age for any student is between 16.650504 and 16.83794

* ***The next question I asked was is there a significant difference in the average total grade between male and female students? \*\*\*\*Check Appendix B\*\*\*\****

***Step 1:***

H0: There is not a significant difference in the average total score between male and female students.

Ha: There is a significant difference in the average total score between male and female students

***Step 2:***

Selected µ as the appropriate statistical measure since the interest was in comparing the average total score for both genders. Appropriate statistical measure->

µ₁ = the true mean average total score for females

µ₂= the true mean average total score for males

***Step 3:***

Since comparing difference between 2 groups this is a 2-sided test t test

.

***Step 4:***

µ₁- µ₂=0 There is no difference in the average total scores between male and female students

µ₁- µ₂≠0 There is difference in average total scores between males and female students.

***Step 5:***

Use a .05 level for test.

**Step 6:**

*Select Appropriate Test Statistic to be 1.64707*

***Step 7:***

*Select Appropriate test statistic as 1.96431*

***Step 8:***

Concluded that the test statistic 1.647644 is < 1.96431 (the critical value) therefore fail to reject null hypothesis.

***Step 9:***

There is insufficient evidence at the 0.05 level to conclude that the average total scores are significantly different between male and female students.

* ***Next, I asked, which variables besides the first and second test scores are good predictors of total grade? \*\*\*CHECK Appendix C\*\*\*\****

I performed multiple linear Regression in excel and Minitab for analysis. \*\*Module 5\*\*

* Trasformed the ‘yes’ and ‘no’ binary variables in the data set to 1’s for yes and 0’s for no with the find and replace function in excel. I created two tables to display regression outfirst. The first tab includes 13 control variables as well as G1 and G2 as predictor variables and the other regression includes solely the 13 control variables.
* G3(final grade) is established as the dependent y variable. G1 is the first period grade and G2 is the second period grade.
* Without the G1 and G2 predictor variables, there was a multiple R (correlation coefficient) of 0.923131376567648. The R squared 0.852171538403681 represents the proportion of variance in grade that can be explained by all the predictor variables, including G1 and G2.
* For a better analysis, I removed the G1 and G2 predictor variables and held 13 control variables to observe how the other variables besides prior period grades affect their final grade.
* This test proves that G1 and G2 are good predictor variables for the total score.
* However, to see which predictor variables other than G1 and G2 were useful, I removed them for a better analysis. After removing the first and second grade periods (G1 and G2) as predictor variables, find multiple R (the correlation coefficient) has decreased to 0.431524591932914. R square (the coefficient of multiple determination) explains that 0.186213473442868 is the proportion of the variance in the grade that can be explained by the predictor variables without including prior scores.
* Health, workday alcohol consumption, free time, romance (whether they have a significant partner), and internet access at home are predictor variables with p<.05 that appear to be statistically significant.

***Graphical user interface, chart, application, table

Description automatically generated***Appendix A:

Appendix B:

*Graphical user interface, table

Description automatically generated*

Appendix C:

Graphical user interface, table, Excel

Description automatically generated

Y= 10.47906b₀ + 0.183492x₁-0.92337x₂+0.394751x₃-0.00932x₄+2.9345₅+1.067476x₆-0.41099x₇+0.130104x₈-0.27921x₉-.02096x₁₀-0.35452x₁₁-0.16508x₁₂-0.18492x₁₃-0.02413x₁₄