LINEAR REGRESSION ANALYSIS MSDS 601
PROJECT REPORT

PREDICTING STUDENT GRADES BASED ON SOCIAL AND ACADEMIC FACTORS

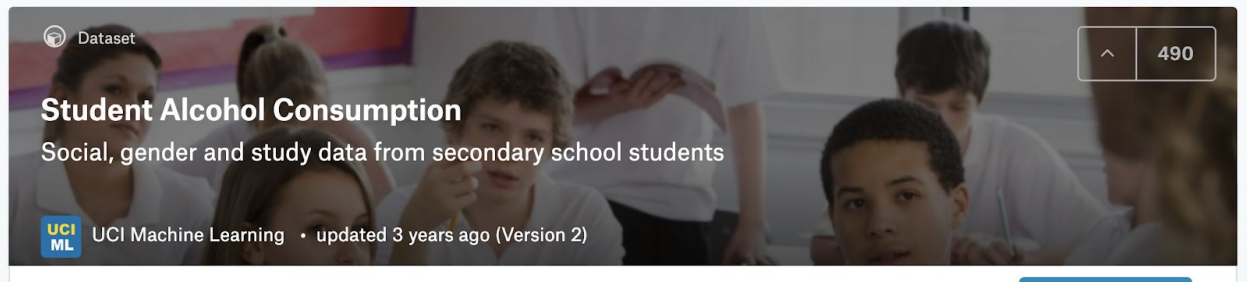
SHREEJAYA BHARATHAN & AAKANKSHA NALLABOTHULLA SURYA

PREDICTING STUDENT GRADES BASED ON SOCIAL AND ACADEMIC FACTORS

1. DESCRIPTION OF DATASET:

This project was done on social, gender and study data from secondary school students, found on Kaggle.
Data source: https://www.kaggle.com/uciml/student-alcohol-consumption

The dataset under consideration has the math grades of students. Below is a detailed description of the data.



Dimensions - This data has 395 rows (sample size) and 33 columns (variables)

Variables description

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, output target)

Research Problem Statement

The purpose of this research is to analyse whether math **grades** of a secondary school student have a linear relationship with **social factors** like sex, family size, parent's cohabitation status, family educational support, going out with friends, having a romantic relationship, workday alcohol consumption and **academic factors** like number of past failures, absences, paid extra classes for the subject.

Response Variable: ***Final grade in mathematics***
Predictors under consideration - social and academic factors: ***sex, family size, parent's cohabitation status, family educational support, going out with friends, having a romantic relationship, workday alcohol consumption, number of past failures, absences, paid extra classes for the subject***

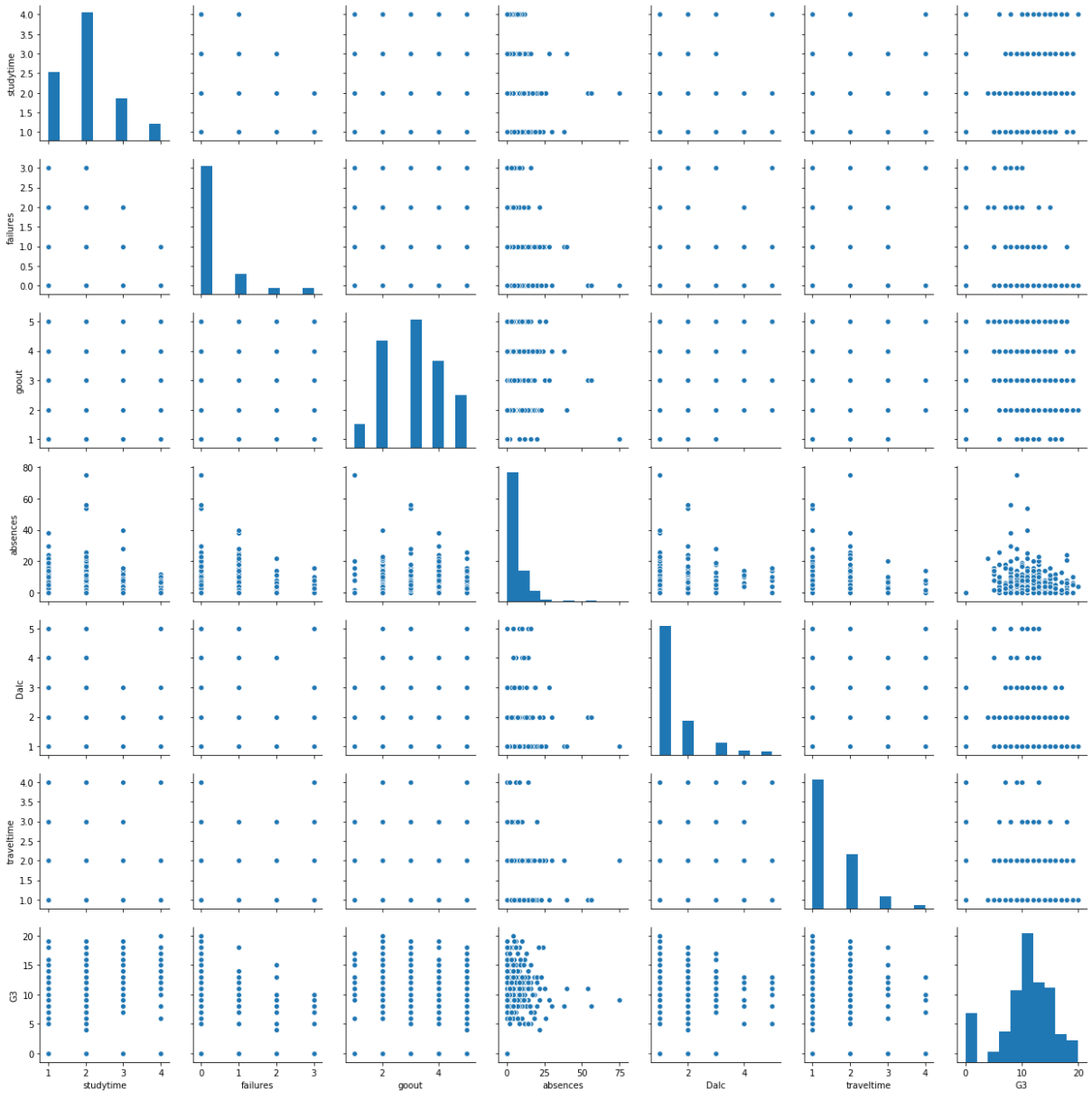
Summary of methods used:

The preliminary analysis included scatterplots, correlation heatmap, and individual t-tests. We created a multiple linear regression model using the best subsets procedure and considered adjusted R squared and Mallow's Cp value to find the best set of models. To compare the models, we used AIC, BIC and F test results. We further diagnosed our model for heteroscedasticity, autocorrelation, multicollinearity, and influential points. We compared our best model with and without the influential points. We also plotted a residuals vs fits plot and qqplot to validate our model assumptions. Finally, we used our model to predict on some test data and used RMSE values to validate the final model.

Exploratory Analysis :

As stated in the research problem, we were curious to identify the effect of social and academic factors on a student's mathematics grades. We took into consideration 13 variables - sex, family size, parent's cohabitation status, family educational support, going out with friends, having a romantic relationship, workday alcohol consumption, number of past failures, absences, paid extra classes for the subject. In order to explore and visualize our data, we generated pair plots, scatterplots, correlation heatmap and performed individual t-tests.

Pair Plots:

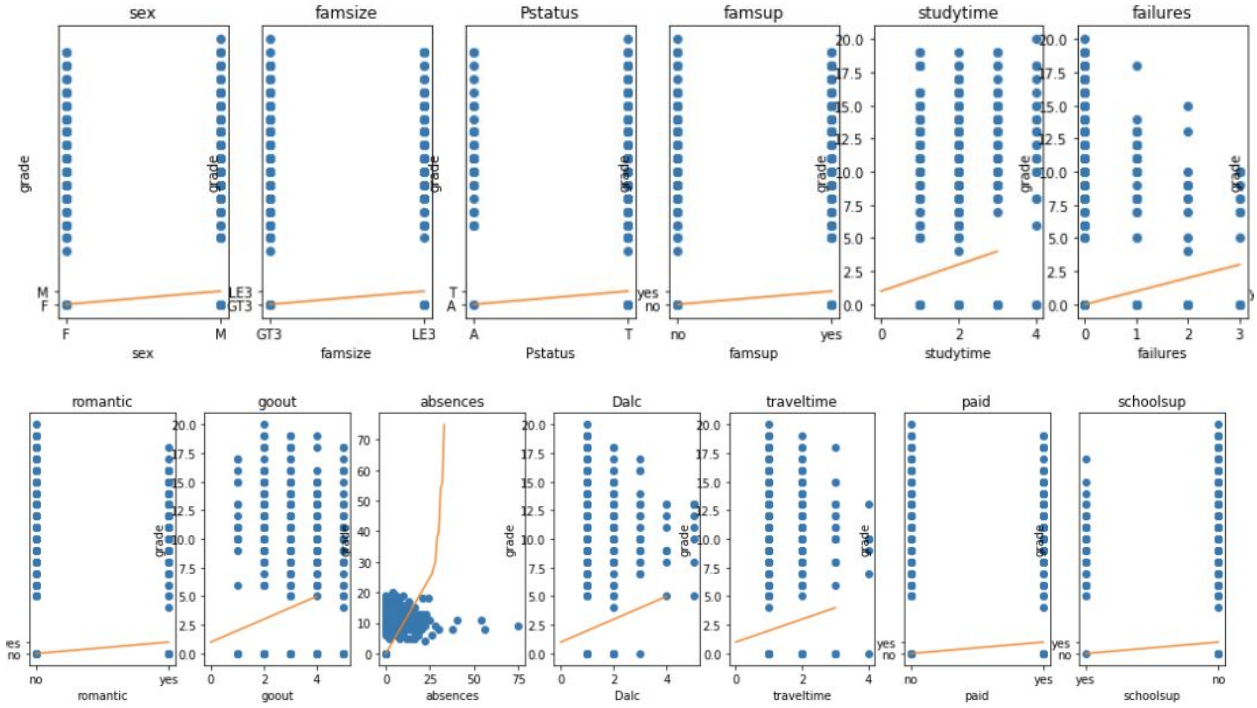


From the pair plots, we observe:

1) Higher number of failures lead to lesser grades
2) As the number of absences increases, grades decrease
3) As travel time increases, grades decrease
4) With higher alcohol consumption, grades get lower

Scatter plots:

As the pair plots did not capture all variables, we created side by side scatter plots for the 13 variables of interest to look for more trends in our data:

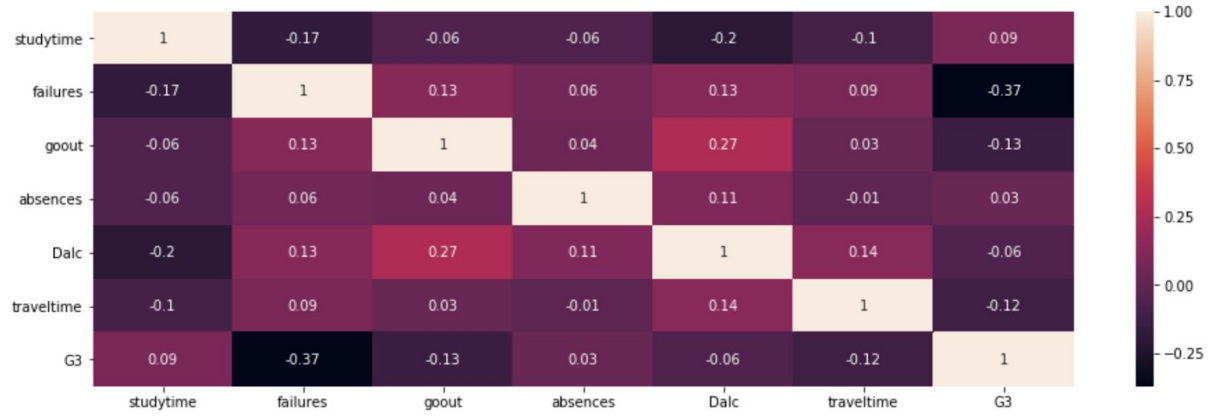


Here we observed:

1) More failures lead to lesser grades
2) Higher alcohol consumption leads to lesser grades
3) As travel time increases, grades seem to decrease

Heatmap:

We then drew a heatmap to look at the amount of correlation between our response and predictor variables. Since the correlation coefficient doesn't give us much information for categorical variables, we only looked for correlation coefficients for numerical variables i.e failures.



We observe that 'failures' is moderately negatively correlated with G3 (-0.37).

Individual t-tests:

Individual t-tests were done on all the variables under consideration:

1. Sex
   p-value = 0.061
   Since p-value was close to our required level of significance, alpha (0.05), we decided to consider this predictor as a significant one

2. Famsize (family size)
   p-value = 0.090
   Since p-value is much greater than alpha in this case, we do not consider famsize to be a significant variable

3. Pstatus (parent's cohabitation status)
   P-value = 0.362
   Since p-value is much greater than alpha in this case, we do not consider Pstatus to be a significant variable

4. Travel time (home to school travel time)
   p-values = 0.116,0.136,0.187
   None of the categories of travel time had p-value less than alpha, so we do not consider traveltime to be a significant variable

5. Study time (weekly study time)
   p-values = 0.817,0.053,0.171
   Since one category has p-value close to alpha, we consider studytime to be a significant variable for our model

6. Failures (number of past class failures)
   p-value = 0.00
   Since the p-value is less than alpha, we consider failures to be a significant variable for our model

7. Schoolsup (extra educational support)
   p-value = 0.082
   p-value > alpha, hence it is not a significant predictor.

8. Famsup (family educational support)
   p-value = 0.452
   p-value >alpha, hence famsup is not significant

9. Paid (extra paid classes within the course subject)
   p-value = 0.064
   Since p-value is close to alpha, we consider paid to be a significant variable

10. Romantic (with a romantic relationship)
    p-value = 0.004
    As p-value < alpha, romantic is a significant variable

11. Goout (going out with friends)
    p-value = -0.705, -0.869, -2.247, -2.965
    None of the categories are significant so goout is not significant for our model

12. Dalc (workday alcohol consumption)
    p-value = 0.024, 0.770, 0.437, 0.945
    One category is significant, so we consider Dalc to be significant for our model

13. Absences (number of school absences)
    p-value = 0.633
    Because p-value <alpha, absences is not a significant variable

From the results of the t-tests, we found that there were 6 significant variables:
sex, failures, paid, romantic, Dalc and studytime

However, ‘Dalc’ and ‘Paid’ turned out to be insignificant variables, when the entire model was considered and the variables we considered for our multiple linear regression model were:

1. Sex
2. Failures (number of past class failures)
3. Romantic (with a romantic relationship)
4. Study time (weekly study time)

Of these variables Sex, Romantic and Studytime are categorical variables which were dummy coded in the models. And failures was a numeric variable.

Multiple linear regression:

The response variable is Grade3 which is the final grade of the student in math. Our predictors were the sex of the student, failures in past classes, whether the student is in a romantic relationship and the weekly study time of the student. The regression was done with the following response and predictor variables:

Grade3 ~ Sex + Failures + Romantic + Study time

To select the best model, we used the best subsets procedure.
We consider a total of 2^4 = 16 models and calculated their adjusted R-squared and Cp values.

Vars	Rsquared	Adj_Rsquared	Mallows_Cp	Predictors	AIC	BIC
0	-2.22045e-16	-2.22045e-16	70.9823	none	2257.31	2261.26
1	0.00711851	0.00452613	69.6794	sex	2256.56	2264.47
1	0.135755	0.133498	9.99456	failures	2203.14	2211.05
1	0.0149667	0.0123948	66.038	romantic	2253.5	2261.41
1	0.0122267	0.00444892	71.3093	studytime	2258.57	2274.39
2	0.145991	0.14152	7.24491	sex+failures	2200.55	2212.41
2	0.0202401	0.0151105	65.5912	sex+romantic	2253.44	2265.3
2	0.0272263	0.0169865	66.3498	sex+studytime	2254.68	2274.45
2	0.143683	0.1392	8.31594	failures+romantic	2201.59	2213.45
2	0.142761	0.133737	12.7438	failures+studytime	2206.01	2225.77
2	0.0288092	0.0185862	65.6153	romantic+studytime	2254.06	2273.82
3	0.152229	0.145554	6.35078	sex+failures+romantic	2199.73	2215.54
3	0.156548	0.145421	8.34657	sex+failues+studytime	2201.76	2225.48
3	0.0419876	0.0293489	61.5008	sex+romantic+studytime	2250.8	2274.51
3	0.151195	0.139997	10.8305	failures+romantic+studytime	2204.2	2227.92
4	0.163761	0.150488	7	sex+failures+romantic+studytime	2200.46	2228.13

From this, we narrowed down our selection of the models with the Cp values closest to the number of parameters and also the ones with higher adjusted R-squared values.

The two models fitting these criteria were:

1. The full model with all 4 variables
2. Model with 3 variables - sex, failures, romantic

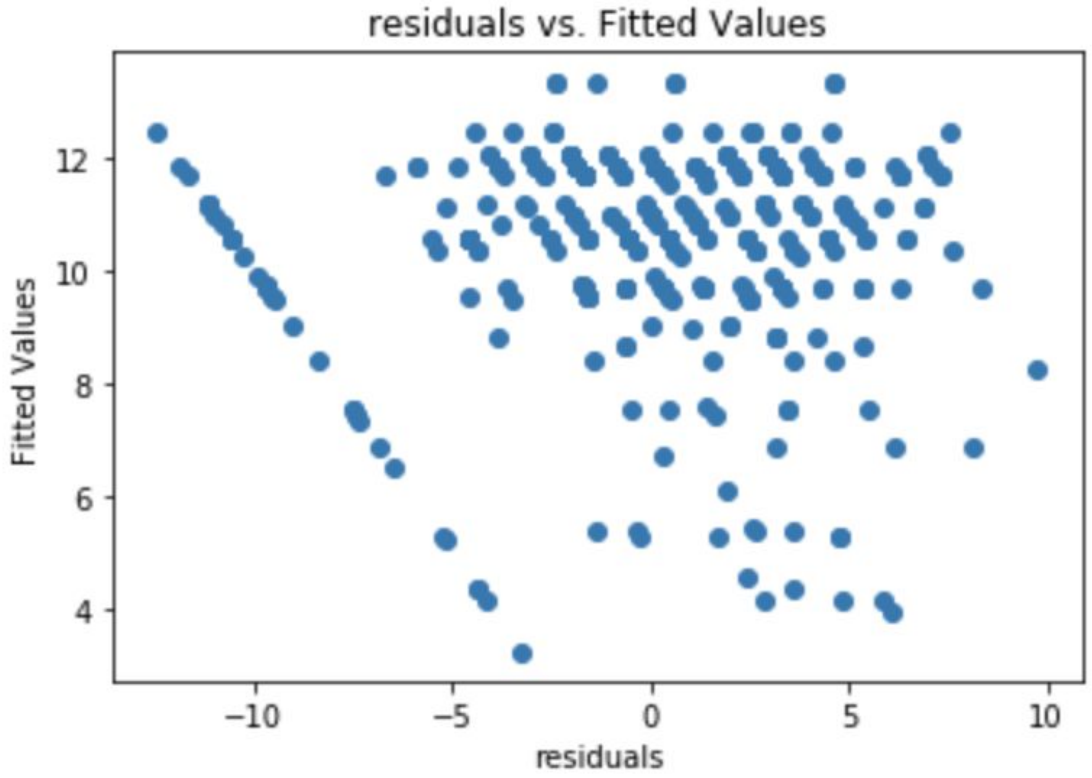
To compare these two models we used AIC and BIC.
Out of these two, full model has lower AIC and higher Adj_Rsquared value so we choose this model although the other model has a lower BIC value

Model Diagnosis:

We performed the following checks for model problems:

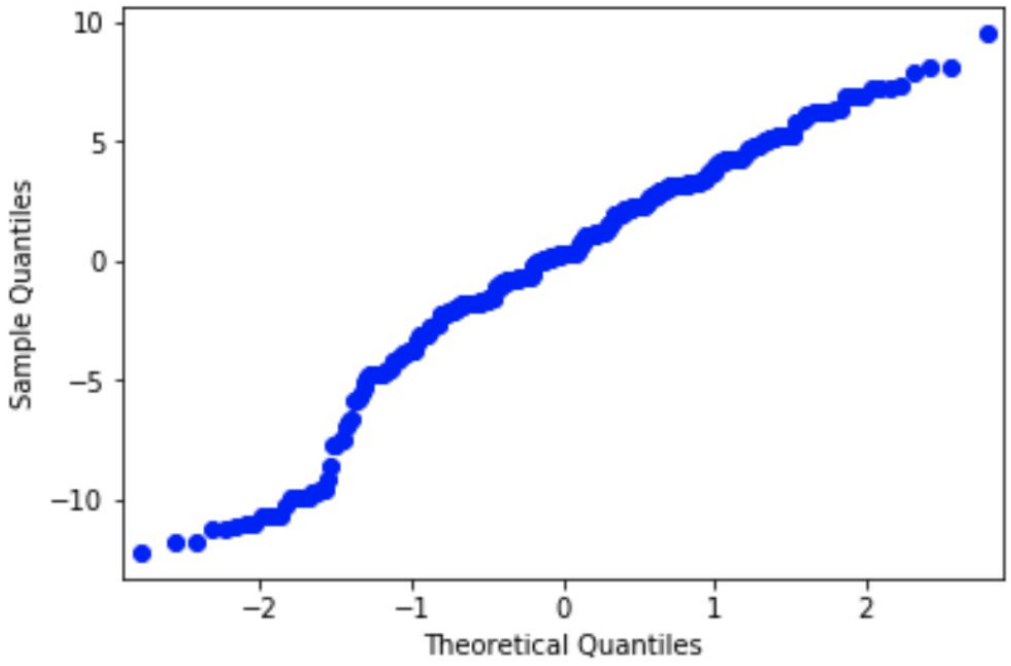
1. **Fitted values vs residuals plot**

As the points bounce randomly around the 0 line, we did not go for any further transformations.



2. **Normality of residuals - qqplot**

   In order to check the validity of the normality of residuals assumption, we plotted a qqplot of the residuals. And from the plot we do not observe a great deviation from our assumption since the plot looks mostly linear.



3. **Heteroscedasticity**

From the Breusch Pagan test for heteroscedsticity, we concluded that there wasn't a non equal variance problem in our model. The p-value was 0.212 > Alpha (0.05) and hence there isn't sufficient evidence to reject the null hypothesis.

4. **Autocorrelation**

From the Breusch Godfrey test we concluded that there was no autocorrelation problem. The p-value was 0.338 > Alpha (0.05) and hence there isn't sufficient evidence to reject the null hypothesis.

5. Influential points

a. Internally studentized residuals

The following indices are flagged as outliers using internally studentized residuals:
[131, 134, 135, 136, 140, 148, 162, 168, 198, 221, 239, 242, 244,259, 260, 264, 269, 296, 316, 332, 333, 334, 337, 341, 383, 387,389]

b. Externally studentized residuals

The following indices are flagged as outliers using externally studentized residuals:
[131, 134, 135, 136, 140, 148, 162, 168, 198, 221, 239, 242, 244, 259, 260, 264, 269, 296, 316, 332, 333, 334, 337, 341, 383, 387,389]

c. Leverage

There were no points in our dataset which were high leverage points.

d. DFFITS

The following points are flagged as outliers using the Difference in fits (DFFITS) criterion:
[2, 47, 130, 131, 134, 135, 136, 140, 148, 157, 198, 221, 242,244, 259, 264, 293, 303, 334, 376, 387]

e. Cook's distance

The following points are flagged as outliers using the Cook's distance criterion:
[2, 47, 130, 131, 134, 135, 136, 140, 148, 157, 198, 221, 242,244, 259, 264, 293, 303, 314, 334, 338, 376, 387]

6. Multicollinearity

To identify whether multicollinearity exists we calculated the VIF scores for our predictor variables. The VIF values for the predictors were:

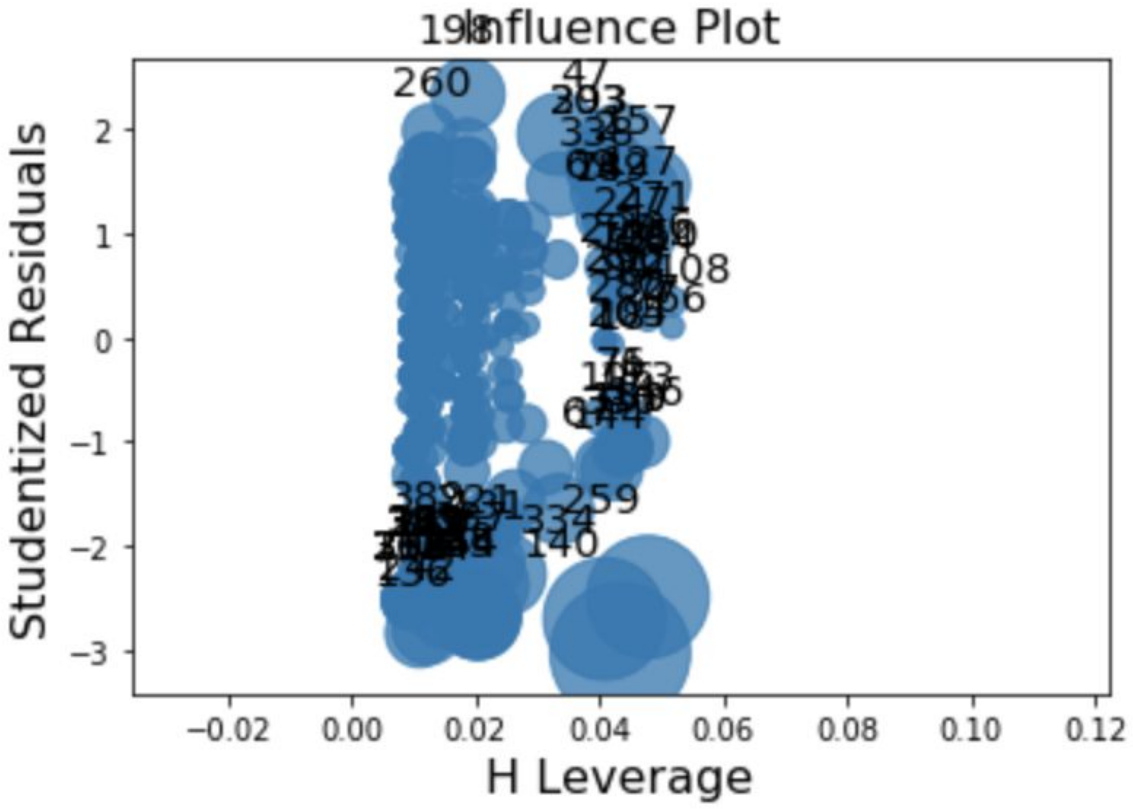
1. Failures - 1.04
2. Sex - 1.16
3. Romantic - 1.046

Since all the VIF are much less than 4, we conclude that there is no multicollinearity problem.

From the model diagnosis, the following were the complete set of influential points:
[2,131,259,260,134,135,136,264,387,389,140, 269,130,148,157,162,293,168,

296,47,303,314,316,198,332,333,334,337,338,341,221,239,242,244,376,383]

Our Influence plot gave us the same values of influential points as the calculated ones:



We further tested a model without any of these influential points, and also one without removing any of them.

Final model selection

The final models with and without influential points are reported below and it is clearly seen that both have significant variables and higher R squared values compared to our previous models. Depending on the industry requirements, we may choose to go with either of these models.

1. Without removing any influential points this is the final model:

Grade(G3) = 10.3769 + 1.3159 * Sex(male) -0.8898 * Romantic(yes) + 0.1794 * Studytime (2) + 1.6694* Studytime (3) + 0.7552* Studytime (4) -2.1436 * failures + Error

- T-test
    - From the t-test we observe that all the predictors have p-values <0.05 and are thus significant (If atleast one of the categorical predictors have a significant p-value, we consider the variable to be significant)
- ANOVA
    - From the ANOVA test, we find that all the variables are significant as the p-values from the f-test are less than 0.05
- R -squared value - 0.165
- Adjusted R-squared value - 0.152

2. After removing all influential points this is the final model:

Grade(G3) = 10.7513 + 1.3763 * Sex(male) -0.4226 * Romantic(yes) + 0.3027* Studytime (2) + 1.9496* Studytime (3) + 0.6249* Studytime (4) -2.4668 * failures + Error

- T-test results:
    - After removing influential points, romantic has become insignificant and the p-values for other predictors have also changed indicating the points were highly influential
- ANOVA
    - From the ANOVA test we find that all the variables are significant as the p-values from the f-test are less than 0.05
- R -squared value - 0.298

- Adjusted R-squared value - 0.286

Predictions on test data:

From our initial dataset, we took a random sample (on which the model was not trained) of 10 observations and tested the model on this sample data.

In order to check the model's performance, we calculated the fitted y-values from the model and calculated the error. Using this we computed the RMSE - root mean squared error terms for the final models - one with all influential points and one without.

The RMSE values were as follows:

1. With the influential points: 1.5954
2. Without the influential points: 0.0266

Interpretations

The model seems like a good fit as the RMSE is only 1.59 (for a grade out of 20) which is not a very huge difference in a practical setting.
Also, on removing the influential points, the RMSE is only 0.026 which is very low for a problem involving practical data. Hence we can conclude that Maths grade of a student can be determined to a good extent by using the chosen predictors.

SUMMARY OF FINDINGS:

Based on our multiple linear regression analysis we identified the most important social and academic factors that predict a student's grade in mathematics.
From our analysis, we observed that the sex of the student, failures in past classes, whether they were in a romantic relationship and their study time were the most important factors. It was interesting to observe that the alcohol consumption of a student did not strongly determine their grades. On a closer look at our data, we observed that majority of the students had low/minimal alcohol consumption and this could be one of the reasons that this variable did not play a significant role in our analysis. Overall, the factors with higher significance also make a lot of sense in the real world setting since academically speaking the time spent on studies and failures in past academic tests do play a major role in how a student performs. From a social perspective, gender and

being involved in a romantic relationship may affect how a student performs academically and this was the case with our dataset. We can extend this research problem for finding relationship between grades and social/academic factors for other subjects as well.