Regression Final Paper

Author: Freeman Chen, Yen Phan, Changqing

Date: 2023-03-29

Stats 530

Abstract

This paper investigates the hypothesis that whether alcohol consumption has any affect over student performance and what are the other influetial factors that hold substantial affect on students grades. For each student observation, data on student performance as the average math grade that calculated in one semester. A multiple linear regression model is proposed and estimated where the key independent variable is the alcohol consumption. other controls to capture the Students' gender, school travel time, study time.... are included in the model. The estimate for the parameters associated with the student's gender, travel time, study time, education support and outside activity are significant. Thus, it is concluded that there is not significant evidence in the data to suggest that the alcohol consumption influence students' performance. diagnostic tests on the model residuals are performed. While model residuals deviate significantly from normality and suggest the presence of heteroscedasticity, diagnostic tests reveal that is no evidence of autocorrelation in the residuals. Furthermore, the multicollinearity among the model predictors is examined and is determined not to be a major problem. Only one outlying observation is present in the data. Potential areas for future research and limitations of the analysis performed in this paper are discussed.

Data

Data is Provided from :
https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?select=student-mat.csv
Through the EDA, we pre filter out the extra data, only kept the data related to student.
Here's the Data Description:
Response Variable :
student performance = (G1 + G2 + G3)/3

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)

Predictors:

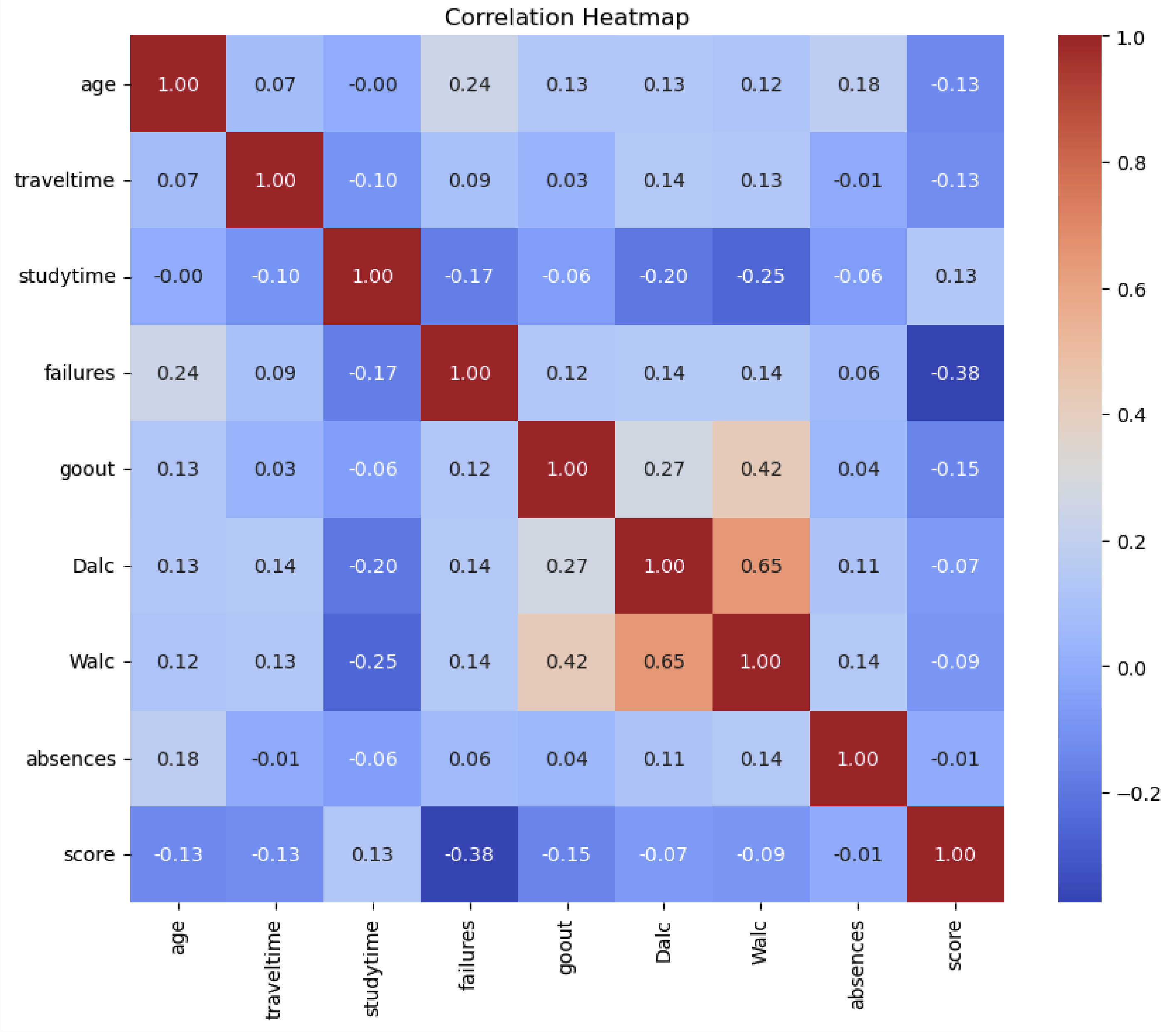
1. sex - student's sex (binary: 'F' - female or 'M' - male)
2. age - student's age (numeric: from 15 to 22)
3. travel time - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
4. study time - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
5. school sup - extra educational support (binary: yes or no)
6. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
7. romantic - with a romantic relationship (binary: yes or no)
8. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
9. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
10. absences - number of school absences (numeric: from 0 to 93)

11. goout - going out with friends (numeric: from 1 - very low to 5 - very high)

    Descriptive statistics based on the 395 observations

EDA

Heatmap



Correlation Heatmap

	age	traveltime	studytime	failures	goout	Dalc	Walc	absences	score
age	1.00	0.07	-0.00	0.24	0.13	0.13	0.12	0.18	-0.13
traveltime	0.07	1.00	-0.10	0.09	0.03	0.14	0.13	-0.01	-0.13
studytime	-0.00	-0.10	1.00	-0.17	-0.06	-0.20	-0.25	-0.06	0.13
failures	0.24	0.09	-0.17	1.00	0.12	0.14	0.14	0.06	-0.38
goout	0.13	0.03	-0.06	0.12	1.00	0.27	0.42	0.04	-0.15
Dalc	0.13	0.14	-0.20	0.14	0.27	1.00	0.65	0.11	-0.07
Walc	0.12	0.13	-0.25	0.14	0.42	0.65	1.00	0.14	-0.09
absences	0.18	-0.01	-0.06	0.06	0.04	0.11	0.14	1.00	-0.01
score	-0.13	-0.13	0.13	-0.38	-0.15	-0.07	-0.09	-0.01	1.00

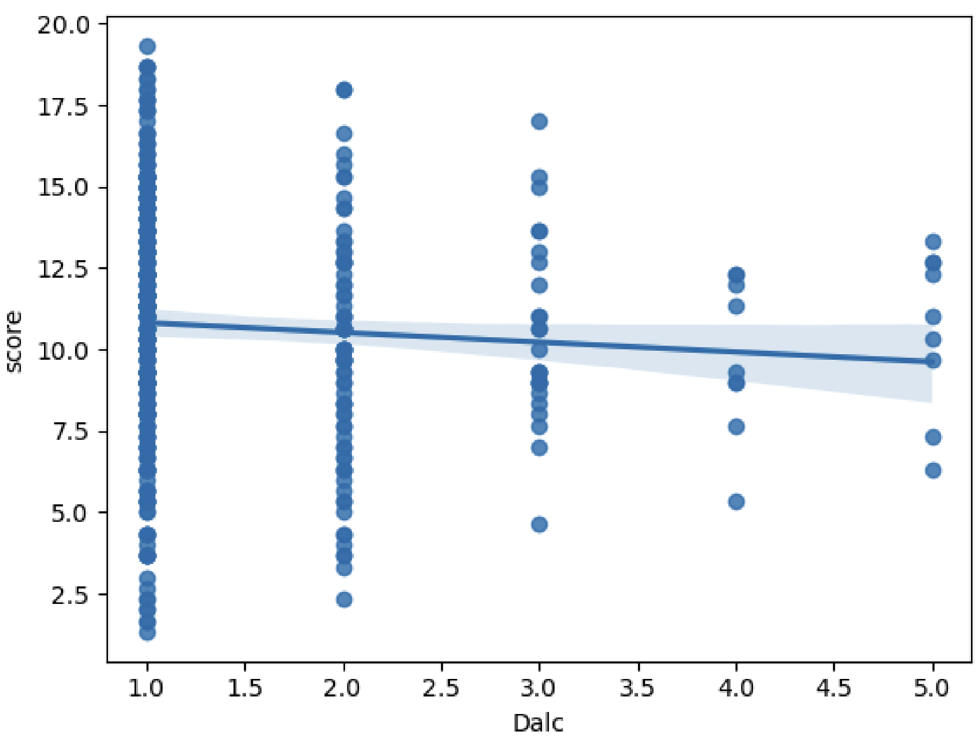
The correlation heatmap displays relationships between various variables, with the most notable being a strong positive correlation of 0.65 between daily and weekend alcohol consumption(possible . Additionally, there's a negative correlation of -0.38 between failures and scores, suggesting students with more failures tend to have lower scores. Most other variables exhibit weak or negligible correlations with each other.

Chart of Descriptive Statistic Table

	age	traveltime	studytime	failures	goout	Dalc	Walc	absences
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
mean	16.696203	1.448101	2.035443	0.334177	3.108861	1.481013	2.291139	5.708861
std	1.276043	0.697505	0.839240	0.743651	1.113278	0.890741	1.287897	8.003096
min	15.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	0.000000
25%	16.000000	1.000000	1.000000	0.000000	2.000000	1.000000	1.000000	0.000000
50%	17.000000	1.000000	2.000000	0.000000	3.000000	1.000000	2.000000	4.000000
75%	18.000000	2.000000	2.000000	0.000000	4.000000	2.000000	3.000000	8.000000
max	22.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	75.000000

It's evident that there might be a notable outlier in the independent variable, "absences". This observation is based on the numerical data: the interquartile range is 8, yet there's a substantial gap of 67 between the maximum value and the 3rd quartile. This gap suggests the presence of a potential outlier.

regplot (scores vs Dalc)



from the regplot, we can see that the (scores vs Dalc) is pretty flat which may indicates that there isn't a strong linear relationship between the daily alcohol consumptions (`Dalc`) This means that changes in `Dalc` are not associated with consistent changes in the scores, at least not in a linear manner.

MODEL STATEMENT:

$$\text{score} = \beta_0 + \beta_1 \times \text{C}(\text{sex}) + \beta_2 \times \text{age} + \beta_3 \times \text{traveltime} + \beta_4 \times \text{studytime} + \beta_5 \times \text{failures} + \beta_6 \times \text{C}(\text{schoolsup}) + \beta_7 \times \text{C}(\text{romant}$$

The multiple linear regression model results are provided below:

Residuals:

	coef	std err	t	$P > \|t\|$
Intercept	15.0922	2.449	6.162	0.000
C(sex)[T.M]	1.0366	0.370	2.798	0.005
C(schoolsup)[T.yes]	−1.6885	0.526	−3.207	0.001
C(romantic)[T.yes]	−0.5715	0.366	−1.563	0.119
age	−0.2001	0.145	−1.382	0.168
traveltime	−0.4526	0.244	−1.853	0.065
studytime	0.5194	0.217	2.396	0.017
failures	−1.5875	0.238	−6.656	0.000
goout	−0.3701	0.167	−2.217	0.027
Dalc	0.0193	0.251	0.077	0.939
Walc	−0.0300	0.187	−0.161	0.872
absences	0.0290	0.022	1.329	0.185

ADJ R SQUARED AND F STATISTIC AND p-value:

$AdjustedR - squared$	$F - statistic$	$p - value$
0.194	9.611 on 11 and 383 DF	$2.25e - 15$

Analysis:

Predictor Variables

1. The model intercept is 15.0922. This represents the expected score when all predictor variables are zero (or at their reference levels for categorical variables).
2. **C(sex)[T.M]**: Males (denoted as "M" in the dataset) are expected to score, on average, 1.0366 points higher than females, holding all else constant. This effect is statistically significant with a p-value of 0.005.
3. **age**: For every year increase in age, there's an expected decrease in score by 0.2001. However, this relationship is not statistically significant at conventional levels, given its p-value of 0.168.
4. **traveltime**: An additional unit increase in travel time results in an expected decrease of 0.4526 in the score, with a borderline p-value of 0.065.
5. **studytime**: For every unit increase in study time, the score is expected to increase by 0.5194. This effect is significant with a p-value of 0.017.
6. **failures**: For each preovis class failure, there's a significant decrease of 1.5875 in the expected score.
7. **goout**: Going out seems to have a negative effect on scores. For each unit increase in going out, there's an expected decrease in score by 0.3701. This is significant with a p-value of 0.027.
8. **Dalc & Walc**: Both weekday and weekend alcohol consumptions (`Dalc` and `Walc`) don't seem to significantly impact the scores, as evidenced by their high p-values.
9. **absences**: The coefficient suggests that for each additional absence, the score increases by 0.0290. However, this is counter-intuitive and also not statistically significant with a p-value of 0.185.
10. **romantic**: Being in a romantic relationship (`C(romantic)[T.yes]`) seems to decrease scores by 0.5715, but this is not statistically significant with a p-value of 0.119.
11. **schoolsup**: Students with school support (`C(schoolsup)[T.yes]`) score 1.6885 points lower on average, which is significant.

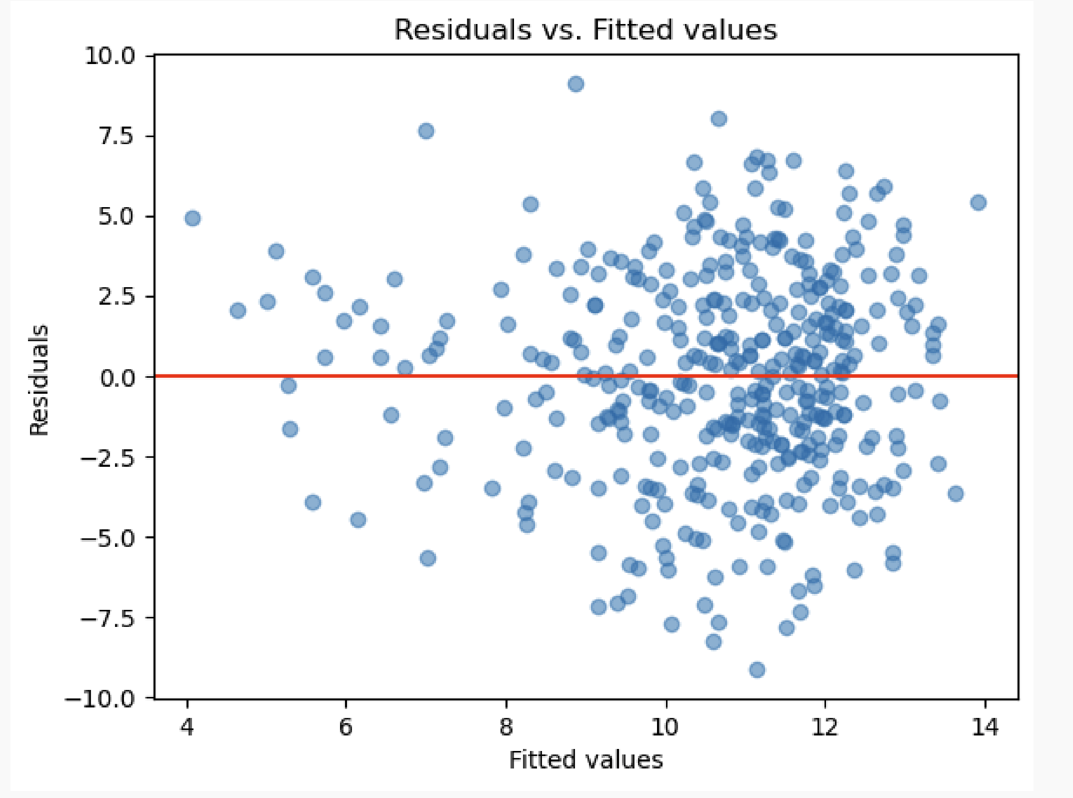
Model Fit:

The Adjusted R^2value is 0.194, which suggests that about 19.4% of the variance in scores is explained by the model. A low R squared value suggests that the proposed model is not particularly effective at capturing variation in Score. This may be due to other omitted variables. However, the F-statistic is statistically significant at the 95% significance level, so we may conclude that there is indeed evidence that at least one of the parameters of the models is not zero.

Model Diagnosis:

Assumption Check:

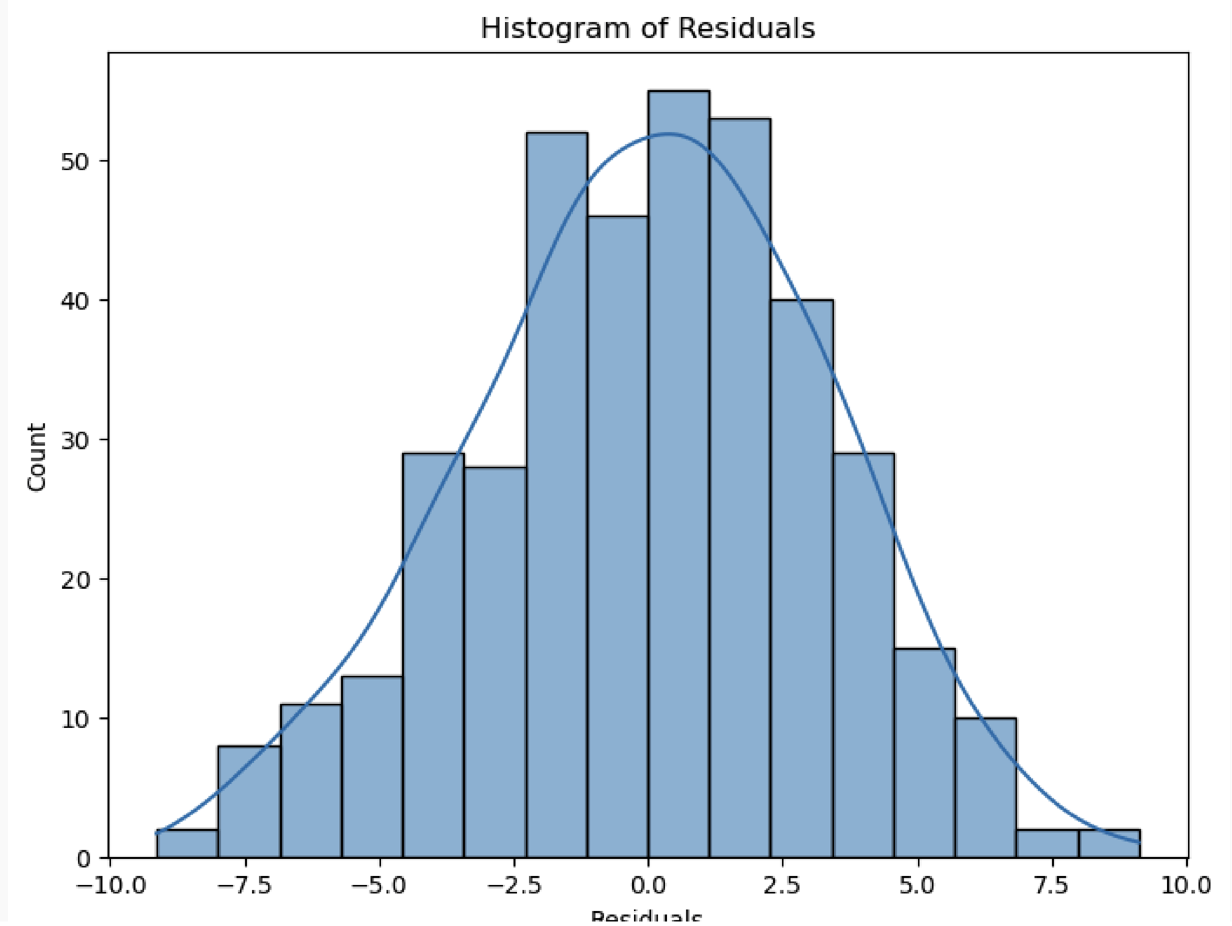
Check Linearity:



From the graph we observe a random scatter of points around the horizontal zero line without any apparent pattern, it might suggests that the assumption of linearity and homoscedasticity hold

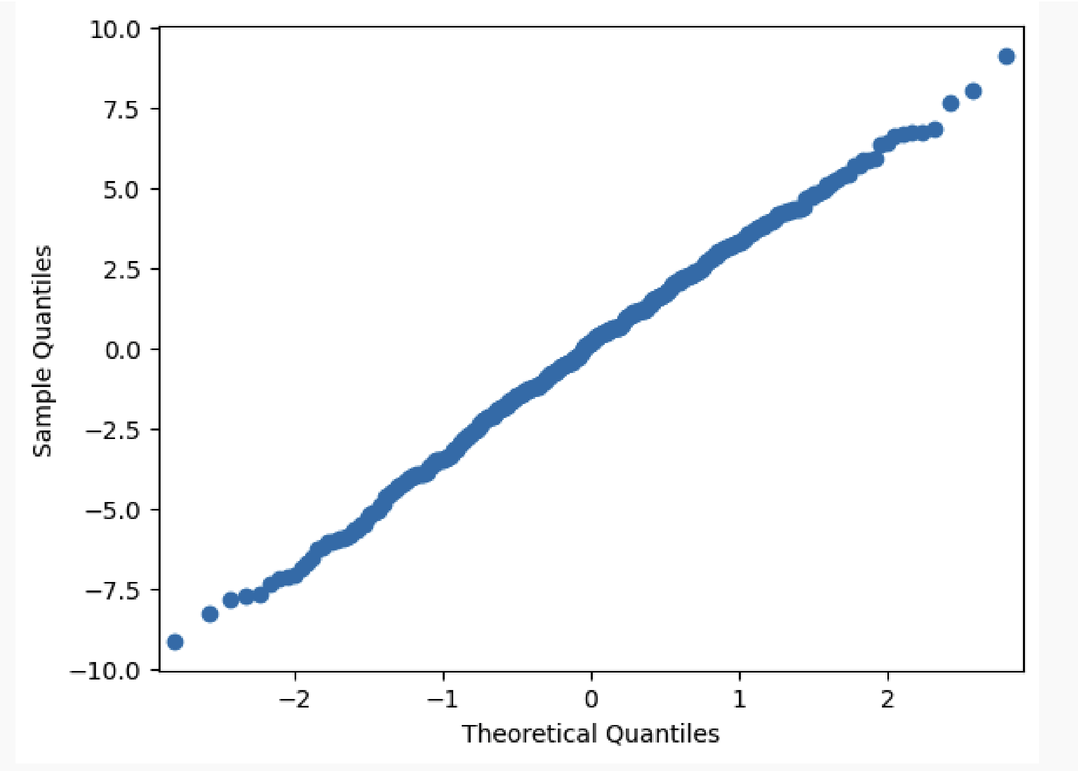
Normality of Residuals:

- Histogram:



Through the histogram plot of the residuals we observed that the histogram of the residuals displays a bell-shaped curve (resembling a normal distribution), it suggests that one of the assumptions of linear regression, which is the normality of residuals, is likely being met

- Q-Q plot:



The Q-Q plot looks reasonably straight suggest that the normality of residuals, is likely being met.

- Kolmogorov-Smirnov test:

Metric	Value
Statistic	0.2743
P-value	$9.20e-27$
Statistic Location	1.5351
Statistic Sign	-1

Through the two-sided Kolmogorov-Smirnov test, it yields a test statistic of D = 0.2743 and a p-value of approximately 0. Thus, it may concluded that there is sufficient evidence to suggest that the residuals are not normally distributed.

Through the plots and the k-s test, we found the K-S test result suggests non-normality, the visual inspection contradicts. since Normality is not seriously violated, therefore we assume the normality

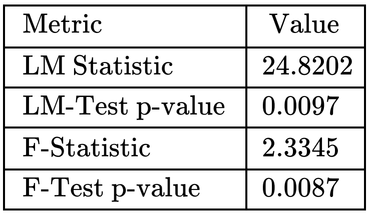
Multicollinearity:

- VIF Score:

VIF Factor	features
215.041468	Intercept
1.226789	C(sex) [T.M]
1.117425	C (schoolsup) [T.yes]
1.066989	C (romantic) [T.yes]
1.220138	age
1.037523	traveltime
1.183921	studytime
1.124897	failures
1.235356	goout
1.784689	Dalc
2.064111	Walc
1.088431	absences

Homoscedasticity:

Here's the Breusch_Pagan test result:

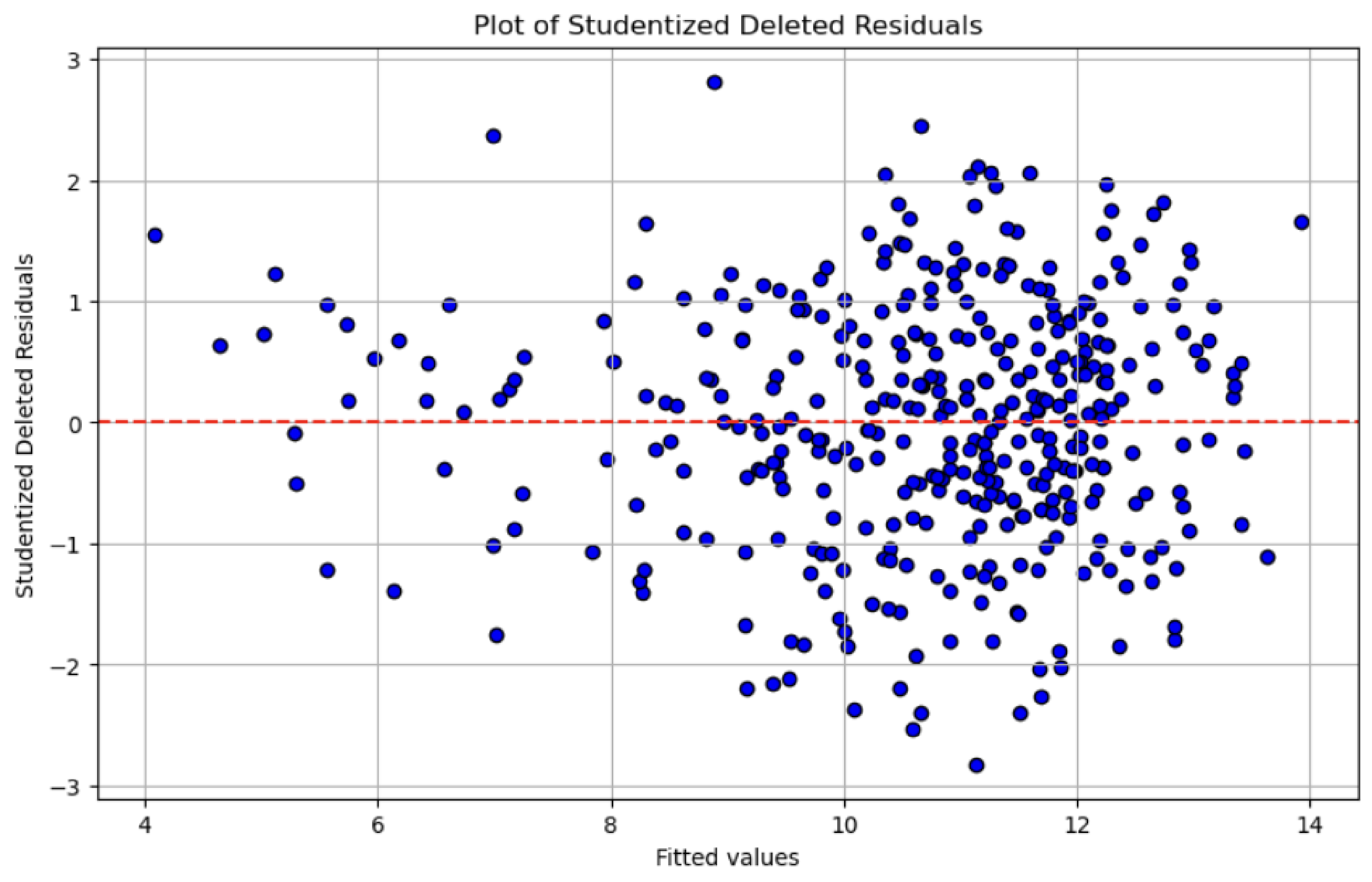


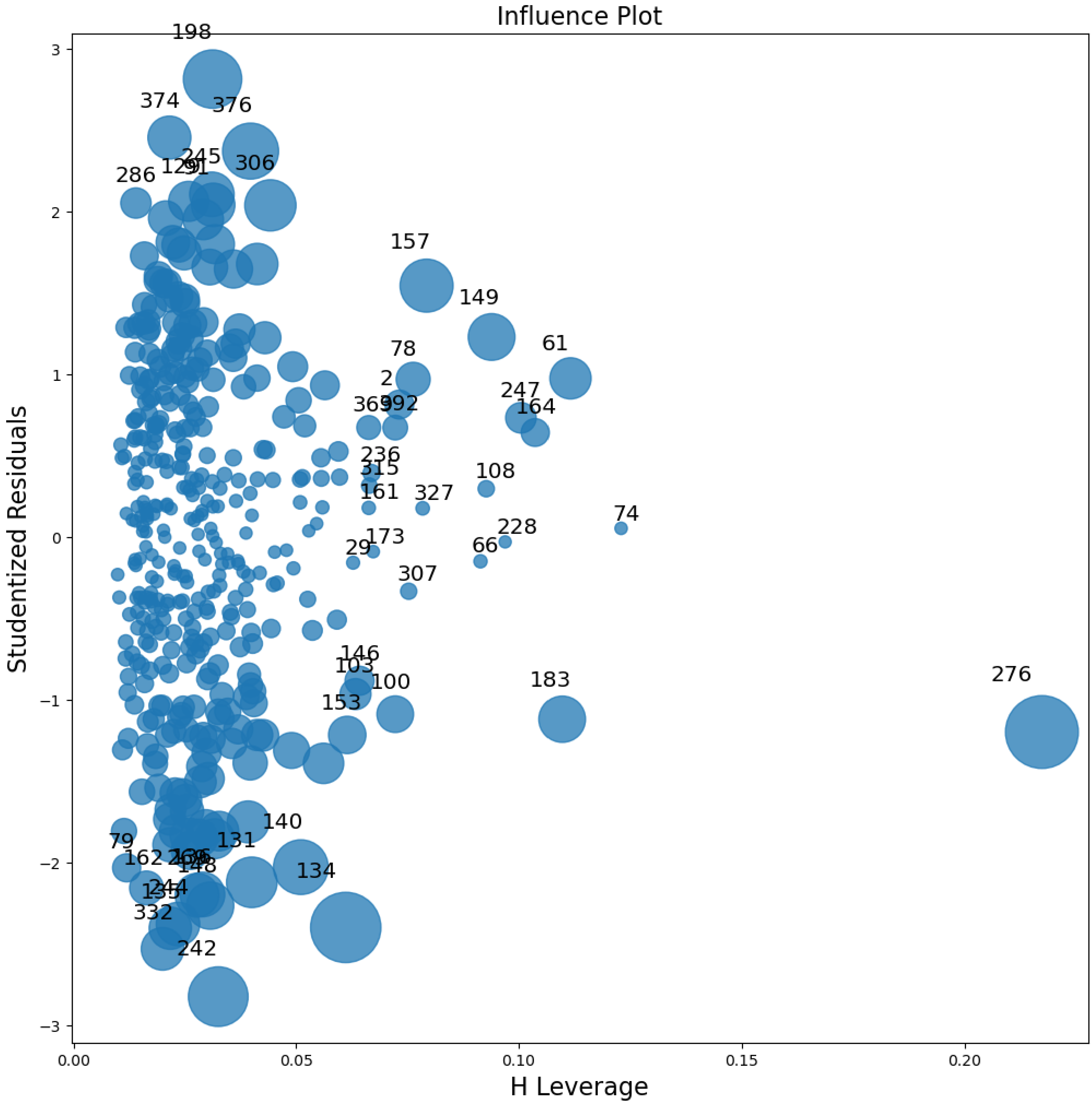
Metric	Value
LM Statistic	24.8202
LM-Test p-value	0.0097
F-Statistic	2.3345
F-Test p-value	0.0087

The Breusch-Pagan test orivudes twi sets if statistics: the Lagrange Multiplier (LM) test statistics and the F-Test. Both tests are essential but would give a slightly different assumptions:

- The P-vlaue for the LM-test is 0.0097, which is less than 0.05(commonly used alpha level). Therefore, we would reject the null hypothesis in favor if the alternative hypothesis using the LM test, which would suggest that the presence of heteroscedasticity.
-
- the p-value for the F-Test is 0.0087, which is also less than 0.05. Again, this would lead you to reject the null hypothesis in favor of the alternative hypothesis using the F test, suggesting the presence of heteroscedasticity.

Outlier and Influential points detection





The two diagnostic plots provide insights into the potential weaknesses of the regression model. The "Studentized Deleted Residuals" plot reveals that while most residuals cluster around the zero line, suggesting an adequate model fit, there are some notable outliers. The "Influence Plot" further categorizes these outliers based on their influence and leverage. Observations such as 198, 376, 306, and notably 276 stand out as potentially problematic, with 276 exhibiting high leverage, making it particularly influential on the model's predictions. Investigating the nature of these data points and considering model refinement might be essential for a more reliable and robust regression outcome.

model selection:

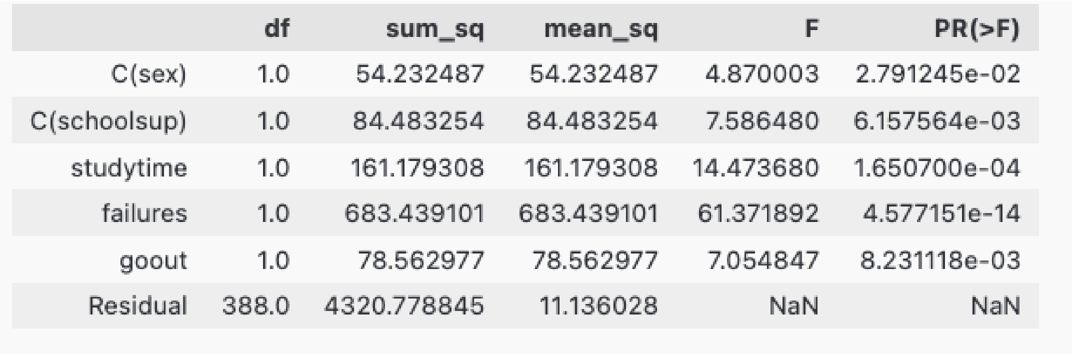
Following an exhaustive model diagnosis, we detected heteroscedasticity and influential data points. Our initial model integrated 11 predictors: sex, age, travel time, study time, school support, failures, romantic involvement, Dalc, Walc, absences, and going out.

To rectify the identified issues, we took several steps. First, we excluded the data at the 276th index. Next, we executed t-tests and ANOVA analyses. Based on significance levels, our final model retained five predictors: sex, school support, study time, failures, and going out. Our study yielded several insights:

- Male students tend to outperform female students on average.
- There is a positive correlation between study time and grades.
- Conversely, both failures and frequent outings correlate with decreased grades.

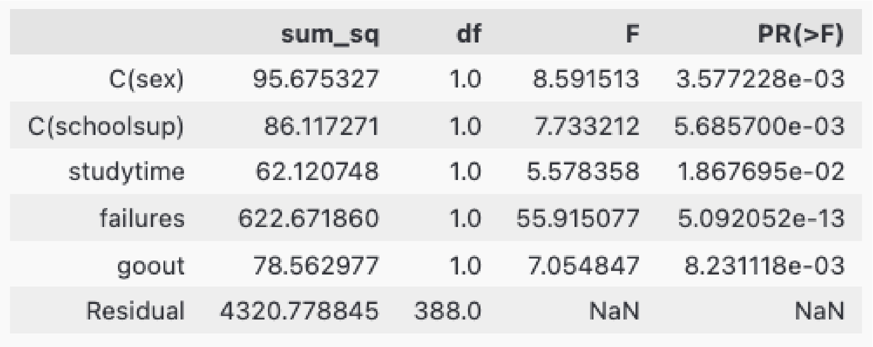
Interestingly, students availing school support showed diminished grades compared to those without. This raises a hypothesis: perhaps the aid diminishes motivation, adversely affecting academic prowess.

Anova type 1:



	df	sum_sq	mean_sq	F	PR(>F)
C(sex)	1.0	54.232487	54.232487	4.870003	2.791245e-02
C(schoolsup)	1.0	84.483254	84.483254	7.586480	6.157564e-03
studytime	1.0	161.179308	161.179308	14.473680	1.650700e-04
failures	1.0	683.439101	683.439101	61.371892	4.577151e-14
goout	1.0	78.562977	78.562977	7.054847	8.231118e-03
Residual	388.0	4320.778845	11.136028	NaN	NaN

Anova type 2:



	sum_sq	df	F	PR(>F)
C(sex)	95.675327	1.0	8.591513	3.577228e-03
C(schoolsup)	86.117271	1.0	7.733212	5.685700e-03
studytime	62.120748	1.0	5.578358	1.867695e-02
failures	622.671860	1.0	55.915077	5.092052e-13
goout	78.562977	1.0	7.054847	8.231118e-03
Residual	4320.778845	388.0	NaN	NaN

Subsequent to our analysis, we refined the regression model with 'score' as the dependent variable and 'C(sex)', 'studytime', 'failures', and 'C(school)' as the independent variables. The model showcased an R-squared value of 0.197, suggesting that our model accounts for approximately 19.7% of the 'score' variation.

	coef	std err	t	$P > \|t\|$
Intercept	11.1881	0.734	15.242	0.000
C(sex)[T.M]	1.0477	0.357	2.931	0.004
C(schoolsup) [T.yes]	-1.4072	0.506	-2.781	0.006
studytime	0.5045	0.214	2.362	0.019
failures	-1.7285	0.231	-7.478	0.000
goout	-0.4073	0.153	-2.656	0.008

$AdjustedR - squared$	$F - statistic$	$p - value$
0.187	19.07 on 5 and 387 DF	$5.67e - 17$

In conclusion, our optimized model sheds light on the nuances of academic achievement. The counterintuitive relationship between school support and grades emphasizes the intricate factors affecting student outcomes. As we look ahead, expanding our dataset or enhancing our feature engineering may provide even more robust insights.

Conclusion

The objective of this paper was to scrutinize the supposition that student performance in mathematics can be influenced by factors such as alcohol consumption and absences. A multifaceted linear regression model was employed, Key findings and insights include:

1. **Alcohol Consumption**: The study found no significant evidence to suggest that alcohol consumption influences students' performance. This is further supported by the regplot, indicating no strong linear association between daily alcohol consumption and scores.
2. **Significant Predictors**: Variables such as students' gender, travel time, study time, educational support, and outside activities had statistically significant effects on performance. Specifically:
   - Male students, on average, scored higher than female students.
   - Those receiving extra educational support tended to perform worse.
   - Longer study times were associated with better scores.
   - More failures corresponded with lower scores.
   - Going out more frequently led to a decrease in performance.
3. **Model Diagnostic**: Even though the residuals of the model deviated from normality (indicating potential issues with heteroscedasticity), there was no autocorrelation present. There was also no major issue of multicollinearity among the model predictors. However, a potential outlier was observed in the 'absences' variable, which could influence the model's accuracy and requires further investigation.
4. Limitations:
   - The dataset encompasses students from two schools, potentially limiting the extrapolation of these findings to a broader educational context or different cultural settings.
   - The methodology, reliant on self-reported data, might instigate biases. Students' inclination to underreport or overreport certain behaviors, such as alcohol consumption or absences, can impact the accuracy and reliability of the results. Furthermore, the study's focus on mathematical performance might not reflect the holistic academic capabilities of the student body, as performance in other subjects wasn't evaluated.

In summary, while alcohol consumption doesn't appear to be a primary factor influencing student grades based on this dataset, several other variables like study time, number of failures, and extra educational support have marked impacts on academic performance. Future work can enhance these findings by refining the analysis and exploring other related factors.