A group of people in a meeting

Description automatically generated

Statistic Model to Analyze Student’s Performance

Xuchuan Zheng, Sungki Park, Prashant Sharma

Contents

[Introduction 3](#_Toc162994268)

[Methodology 3](#_Toc162994269)

[Data Source 3](#_Toc162994270)

[Variable Explanations and Data Assumptions 3](#_Toc162994271)

[Approach and Workflow 5](#_Toc162994272)

[Workload Distribution 6](#_Toc162994273)

[Result 6](#_Toc162994274)

[Variable Selection Procedures 6](#_Toc162994275)

[Main Effects Individual T-tests: 6](#_Toc162994276)

[Hypothesis Statement for Individual T-tests 6](#_Toc162994277)

[Hypothesis Statement for Individual T-tests (Interaction Terms) 6](#_Toc162994278)

[Interaction Term T-tests: 6](#_Toc162994279)

[Hypothesis Statement for ANOVA Test: 6](#_Toc162994280)

[Multiple Regression Assumptions 6](#_Toc162994281)

[Conclusion 7](#_Toc162994282)

[Discussion 7](#_Toc162994283)

# Introduction

Academic success is important because it is strongly connected to the positive outcomes we value. Students who are academically successful and with high levels of education are more likely to get employed, have stable and better job, have more employment opportunities than those who with less education. Especially, academically successful adolescents have higher self-esteem, have lower level of depression and anxiety, and are less likely to abuse alcohol and engage in substance abuse.

In our final project for Data 603 - Statistical Modelling with Data, we have tried to develop a model to analyze the impact of various demographic and social factors on the performance of students. Academic performance, though it is not the only factor but is one of the crucial factors in shaping a student's future. To get into a good collage/university, student must score grades in school, a good college can lead a better future and economic stability. So, to secure good grades, getting into a great school is enough? Is there something more than a great school that can help a student to perform better? Do the social and demographic factors plays any role in student's performance? In our project we are trying to answer these questions.

Our project aims to study the internal and external factors that influence student performance using the given dataset based on the questions above. Also, we will identify and evaluate the factors that have a significant impact on student’s final grade. Finally, we will predict the student’s final grade based on the significant factors found by modeling process.

# Methodology

### Data Source

We found our datasets for our regression analysis from UC Irvine Machine Learning Repository which is a website is providing a collection of databases, domain theories, and data generators for the analysis of machine learning algorithms. Data attributes include student grades, demographic, social and school related feature. The two datasets we downloaded are provided regarding the performance in two distinct subjects: Mathematics and Portuguese language. Since we didn’t have to analyze our data by the subject, we combined those two datasets into one dataset and used it as a simple.

### Variable Explanations and Data Assumptions

The dataset we are working with is collected during 2005-2006 at 2 Portuguese schools for Mathematics and Portuguese subject. In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system and there are several courses that share core subjects as the Portuguese Language and Mathematics. A 20-point grading scales is used, where 0 is the lowest grade and 20 is the highest score. During the school year, students are evaluated in three periods and the last evaluation G3 corresponds to the final grade. There are closed questions related to several demographic (e.g. mother’s education, family income), social/emotional (e.g. alcohol consumption) and school related variables (e.g. number of past class failure) that were expected to affect student performance.

In our dataset, most attributes are ordinal variables (e.g. In mother’s education, numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education). We considered these variables as qualitative data.

There are 649 rows instances and 30 features in the dataset. The following table is a complete list of variables used in our modeling process.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Scale** | **Type** |
| school | student's school | binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira | Qualitative |
| sex | student's sex | binary: 'F' - female or 'M' - male | Qualitative |
| age | student's age | numeric: from 15 to 22 |  |
| address | student's home address type | binary: 'U' - urban or 'R' - rural | Qualitative |
| famsize | family size | binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3 | Qualitative |
| Pstatus | parent's cohabitation status | binary: 'T' - living together or 'A' - apart | Qualitative |
| Medu | mother's education | numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education | Qualitative |
| Fedu | father's education | numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education | Qualitative |
| Mjob | mother's job | nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other' | Qualitative |
| Fjob | father's job | nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other' | Qualitative |
| reason | reason to choose this school | nominal: close to 'home', school 'reputation', 'course' preference or 'other' | Qualitative |
| guardian | student's guardian | nominal: 'mother', 'father' or 'other' | Qualitative |
| 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 | Qualitative |
| studytime | weekly study time | (Numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours | Qualitative |
| failures | number of past class failures | numeric: n if 1<=n<3, else 4 | Qualitative |
| schoolsup | extra educational support | binary: yes or no | Qualitative |
| famsup | family educational support | binary: yes or no | Qualitative |
| paid | extra paid classes within the course subject (Math or Portuguese) | binary: yes or no | Qualitative |
| activities | extra-curricular activities | binary: yes or no | Qualitative |
| nursery | attended nursery school | binary: yes or no | Qualitative |
| higher | wants to take higher education | binary: yes or no | Qualitative |
| internet | Internet access at home | binary: yes or no | Qualitative |
| romantic | with a romantic relationship | binary: yes or no | Qualitative |
| famrel | quality of family relationships | numeric: from 1 - very bad to 5 - excellent | Qualitative |
| freetime | free time after school | numeric: from 1 - very low to 5 - very high | Qualitative |
| goout | going out with friends | numeric: from 1 - very low to 5 - very high | Qualitative |
| Dalc | workday alcohol consumption | numeric: from 1 - very low to 5 - very high | Qualitative |
| Walc | weekend alcohol consumption | numeric: from 1 - very low to 5 - very high | Qualitative |
| health | current health status | numeric: from 1 - very bad to 5 - very good | Qualitative |
| absences | number of school absences | numeric: from 0 to 93 | Quantitative |
| G1 | first period grade | numeric: from 0 to 20 | Quantitative |
| G2 | second period grade | numeric: from 0 to 20 | Quantitative |
| G3 | final grade | numeric: from 0 to 20, output target | Quantitative |

Table 1

There are three different scores for student performance in the dataset (see Table). We used the final grade G3 as the dependent variable, so we dropped G1 and G2, and then used the remaining variables as independent variables for our analysis. We assumed that a student’s gender, age, address, availability of internet, and family size would not affect a student’s final grade. However, we are expecting that there would be positive affect on parent’s education level, parent’s job, study hours, school support, family education support, and extra paid class. Also, we assumed that there is a negative impact on travel time, number of past class failures, romantic, free time after school, going out with friends, alcohol consumption.

### Approach and Workflow

For the project we are going to use the techniques we learn in Data-603 Statistical Modeling with Data. We will build a multi linear regression model with

We plan to approach this analysis using the methods we have learned in Data 603 – Statistical Modeling with Data. We will run a linear regression model to find the best model using all variables and test the variables for multicollinearity. Since we have many variables, we assume that we have high multicollinearity among variables.

Once we are satisfied with our main effects, we will use the individual t-test to check for significant higher-order terms and interactions. We intend to test this model with another F-test

to evaluate if the higher order terms and interactions are significant. Any significant higher-order

or interaction terms will be added to our main effects to produce our final model. Our model will

then test for the following 6 assumptions as shown below:

1. Multi-collinearity test
2. Main Effects Individual T-test
3. Hypothesis Statement for Individual T-tests
4. Hypothesis Statement for Individual T-tests (Interaction Terms)
5. Interaction Term T-tests:
6. Hypothesis Statement for ANOVA Test:
7. Multiple Regression Assumptions

### Workload Distribution

# Result

### Variable Selection Procedures: (Model building)

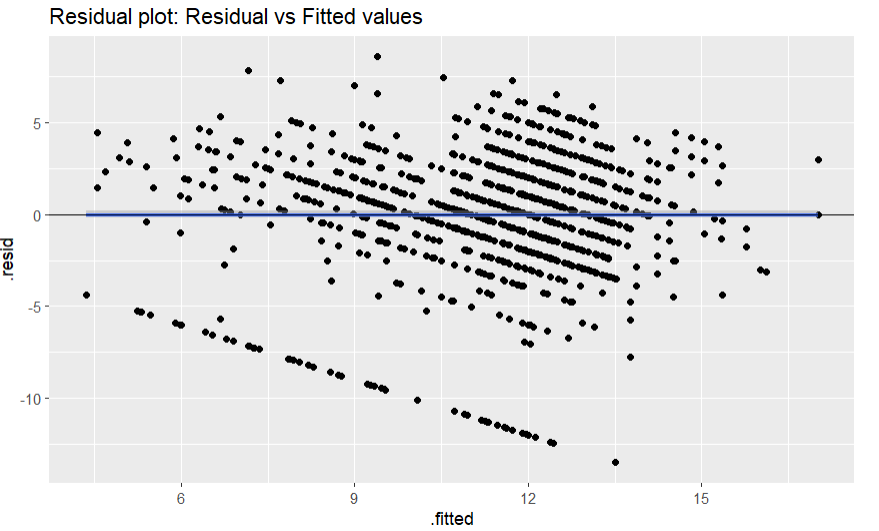
1. Additive model
2. Stepwise forward selection process
3. Drop variable using p-value perform F-test
4. Interaction model
5. Higher order model

### Main Effects Individual T-tests:

Assumption Verification

1. Linearity Assumption

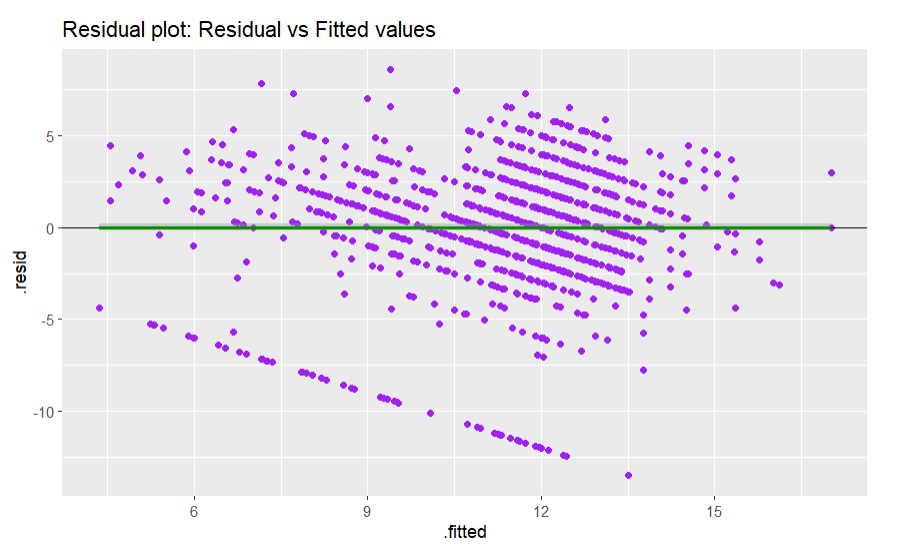
The linear regression model we build is based on the assumption that their is a linear relation between predictors and response variable. To confirm the linear relation we can use Residual plot.



The plot shows that the Residual evenly distributed on both sides so we conclude that the model is linear. There appears to be no pattern of the residuals at all.

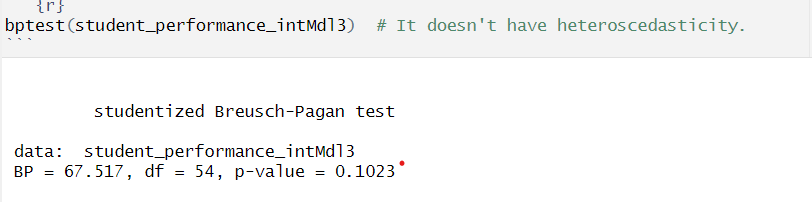
1. Equal Variance Assumption

Another important assumption for our liner regression model is that the error term has a constant variance. To verify the homoscedasticity assumption we can again use the residual vs fitted plot and check if there is any patter.



To confirm homoscedasticity we can perform the Breusch-Pagan Test (bptest) using below hypothesis:

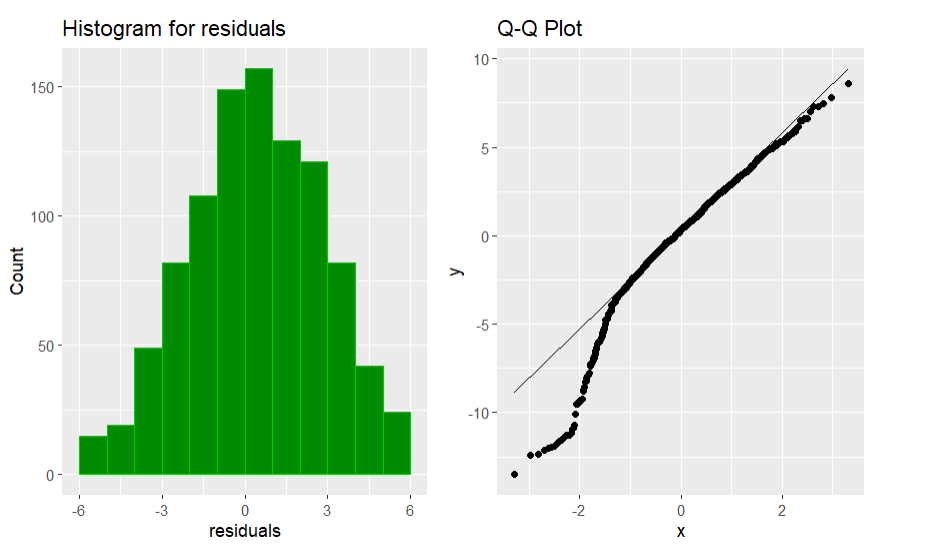
H0 : Hetroscedasticity is not present (error term has comman variance)

Ha : Hetroscedasticity is present (error term do not have comman variance)

Since the p-value is 0.3235 higher then our assumed α=0.05 we cannot reject the H0 and conclude that our model meets the assumption of common variance.

1. Normality Assumption

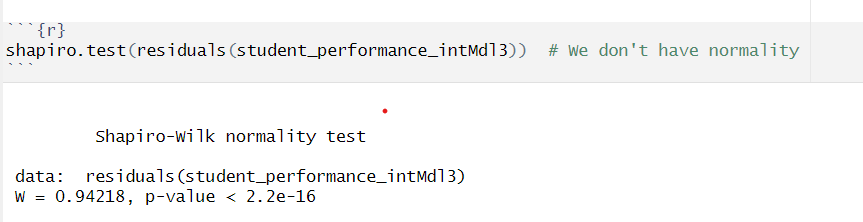
To confirm normality assumption, we plot the histogram for residuals and the Q-Q Plot.



The Q-Q plot shows that the model may not meet the normality assumption so we made Shapiro-Wilk test on the ANOVA residuals.

H0 : The residuals is normal

Ha : The residuals is not normal



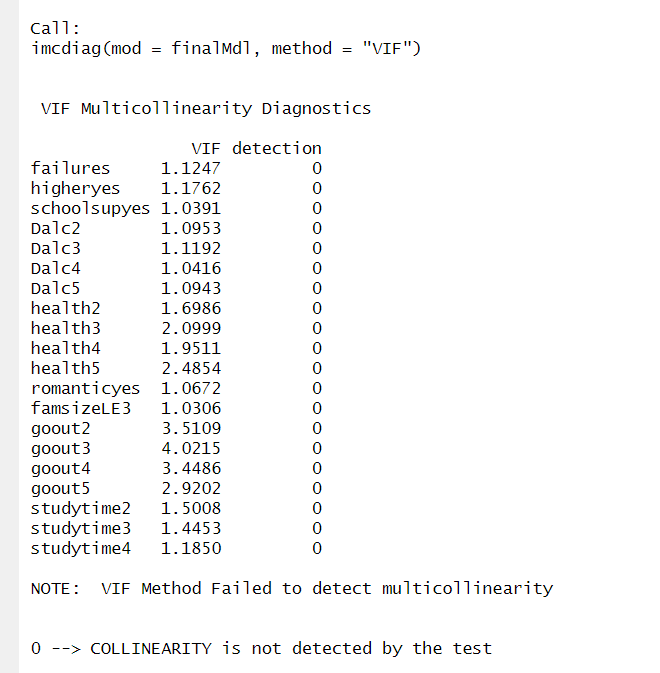
Since the p-value < 2.2e-16 lower then our assumed α=0.05 we can reject the H0 and conclude that our model doesn’t meets the normality assumption.

1. Independence Assumption

Science all the student are independent to each other, we meet the independence assumption.

1. Multicollinearity

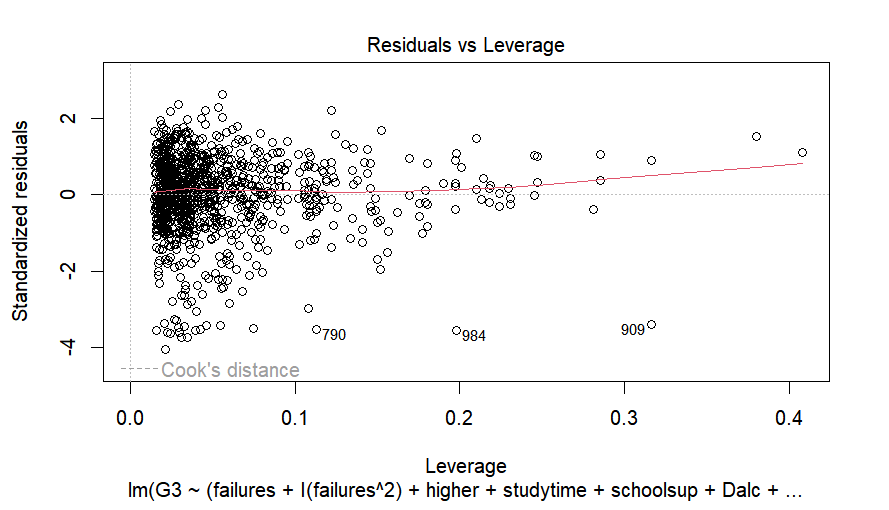
To detect multicollinearity, we use VIF to dentify correlations between variables.



All VIFs are less than 5 so we conclude that there isn’t multicollinearity in our model.

1. Outlier

To find outliers, we plot the Residuals vs Leverage Plot.



From the plot, we can see that there is no points outside the cook’s distance so we can conclude that there is no outliers in the model.

### Hypothesis Statement for Individual T-tests

### Hypothesis Statement for Individual T-tests (Interaction Terms)

### Interaction Term T-tests:

### Hypothesis Statement for ANOVA Test:

### Multiple Regression Assumptions

1. Linearity Assumption
2. Independence Assumption
3. Normality Assumption
4. Equal Variance Assumption
5. Multicolinearity Tests
6. Influential Points and Outliers
7. Interpreting Coefficients
8. Prediction

# Conclusion

# Discussion