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 (Student Performance, n.d.) 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: All Attributes**

There are three different scores for student performance in the dataset (see Table). We used the final grade G3 as a 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 final grade (G3) as the dependent variable, then we use variable selection techniques to select significant variables and perform hypothesis testing to confirm the significance of the selected variables. Once we have our best additive model, we will check for interaction terms and higher order terms. Once we have our final regression model, we will verify all the assumptions of multi-linear regression model.

Below are the workflow steps we are going to perform:

1. Build full additive model.
2. Use forward selection procedure to find significant variables.
3. Perform F-test to check the model usability.
4. Check for interaction between variables and higher order terms.
5. Check usability of final model
6. Provide final model for G3.
7. Verify assumptions for multi-linear regression model.

# Result

### Variable Selection Procedures: (Model building)

1. Build full additive model.

First, we want to test a relationship between the response and the set of independent variables. To address the overall question, we will test:

A math equation with black text

Description automatically generated

We can perform this test through ANOVA.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source of variation** | **DF** | **Sum of Square** | **Mean Square** | **F-statistic** | **P-value** |
| Regression | 66 | 4332.5 | 65.6439 | 5.7027 | 2.2e-16 |
| Residual | 977 | 11246 | 11.5107 |  |  |
| Total | 1043 | 15579 |  |  |  |

**Table 2:** ANOVA table Null Model vs Full Model

For overall modeling test, you can see the output of ANOVA table in Appendix I-2 that shows that *Fcals*=5.703 with *df*=66 and (p-value 2.2e-16 < α=0.05), indicating that we should clearly reject the null hypothesis. In other word, it suggests that at least one of the student performance variables must be related to the final grade.

1. Since we have many variables, we will use forward selection procedure to select significant variables from our full additive model. In forward selection process, we use R to perform individual t-test to test below hypothesis:

A number and equation on a white background

Description automatically generated

From the result of forward selection procedure in Appendix I-3, we could still see some non-significant variables, so we removed these ones and then generate the best additive model in Appendix I-4.

Below table contains significant variables with coefficients, t-value and p-value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Significant variables** | **Coefficient** | **t-value** | **p-value** |
| 1 | failures | -1.82887 | -10.595 | < 2e-16 \*\*\* |
| 2 | higheryes | 1.54720 | 3.732 | 0.000200 \*\*\* |
| 3 | studytime2 | 0.26960 | 1.030 | 0.303179 |
| 4 | studytime3 | 1.18603 | 3.346 | 0.000849 \*\*\* |
| 5 | studytime4 | 0.62526 | 1.272 | 0.203738 |
| 6 | schoolsupyes | -1.16127 | -3.392 | 0.000721 \*\*\* |
| 7 | Dalc2 | -0.73721 | -2.577 | 0.010107 \* |
| 8 | Dalc3 | -0.14301 | -0.315 | 0.753108 |
| 9 | Dalc4 | -1.65491 | -2.367 | 0.018107 \* |
| 10 | Dalc5 | 0.11089 | 0.155 | 0.877041 |
| 11 | health2 | -0.66440 | -1.540 | 0.123946 |
| 12 | health3 | -1.21726 | -0.38251 | 0.001505 \*\* |
| 13 | health4 | -0.74820 | -1.870 | 0.061751 . |
| 14 | health5 | -0.99991 | -2.882 | 0.004038 \*\* |
| 15 | romanticyes | -0.59306 | -2.574 | 0.010190 \* |
| 16 | famsizeLE3 | 0.48561 | 2.040 | 0.041625 \* |
| 17 | goout2 | 1.15597 | 2.460 | 0.014070 \* |
| 18 | Goout3 | 0.67978 | 1.482 | 0.138527 |
| 19 | Goout4 | 0.38442 | 0.800 | 0.423919 |
| 20 | Goout5 | 0.08698 | 0.173 | 0.862617 |

**Table 3:** lm(G3 ~ (failures+higher+studytime+schoolsup+Dalc+health+romantic+famsize+goout), data = studentDataset)

Using the above table, our final additive model is shown below:

A screenshot of a computer code

Description automatically generated

This result tells you that there is a positive impact of higheryes (students desire to go for higher study.), study time, family size, going out with friends have a positive, but there is a negative impact on the number of past course failure, school support, alcohol consumption, current health status, and a romantic relationship.

1. We can compare the residual squared error (RSE) and Adjusted R2 of Full Model and our Fina Additive Model.

|  |  |  |
| --- | --- | --- |
| Model | RSE | Adjusted R2 |
| Full Model | 3.393 | 0.2293 |
| Final Additive Model | 3.449 | 0.2036 |

The Adjusted R2 has decreased which mean that the final model has lost some prediction power. But it is better to have less Adjusted R2 value than over-fitting the model.

1. After deciding our final additive model, we can check for interaction term between variables. From the summary of our full interaction model (Appendix I-5), we can see that all the interaction terms are significant.
2. Since we have finalized our interaction model, we can check for higher order term for our quantitative variables. For this we can pairs plot to see if there is any visual clue to indicate the higher order term.

A diagram of a graph

Description automatically generated

1. Though there is no clear visual indication that we should add higher order term for our variable, but we can still try by adding the higher order term and check the significance. Referring to the summary of our second order model (Appendix I-6), we can see that the second order term for *failure* is significant. Comparing the RSE and Adjusted R2 for the interaction model and higher order model.

|  |  |  |
| --- | --- | --- |
| **Model** | **RSE** | **Adjusted R2** |
| Interaction Model | 3.398 | 0.22271 |
| Higher Order Model | 3.373 | 0.2384 |

We can clearly see that the Adjusted R2 is increased which is a good thing. Hence, we can select our higher order model as the final model.

1. We conducted a final ANOVA test to ensure that interaction model with higher order term is significant in the presence of interaction variables. Thus, we compared the interaction model with the interaction model with higher order term. Table 3 below shows the result of the partial F-test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source of variation** | **DF** | **Sum of Square** | **Mean Square** | **F-statistic** | **P-value** |
| Regression | 1 | 178.24 | 178.24 | 15.668 | 8.087e-05 |
| Residual | 989 | 11250 | 11.3751 |  |  |
| Total | 990 | 11429 |  |  |  |

**Table 4**: ANOVA table (interaction model vs interaction model with higher order terms)

As you see Table 4, p-value that is 8.087e-05<0.05 indicating that we should reject the null hypothesis. So, we chose the interaction model with higher order terms in the presence of interaction variables.

After then, we compared the best additive model with the interaction model with higher order term. Table 3 below shows the result of the partial F-test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source of variation** | **DF** | **Sum of Square** | **Mean Square** | **F-statistic** | **P-value** |
| Regression | 34 | 918.55 | 27.0162 | 2.375 | 2.008e-05 |
| Residual | 989 | 11250 | 11.3751 |  |  |
| Total | 1023 | 12169 |  |  |  |

**Table 5**: ANOVA table (best additive model vs interaction model with higher order terms)

As you see Table 5, p-value that is 2.008e-05<0.05 indicating that we should reject the null hypothesis. So, we also chose the interaction model with higher order terms in the presence of interaction variables.

1. Provide final model for G3.

From the interaction model with higher order terms, we can see coefficient, t-value, and p-value for each significant variable in below table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Significant variables** | **Coefficient** | **t-value** | **p-value** |
|  | failures | -3.8453 | -7.851 | 1.07e-14 \*\*\* |
|  | I(failures^2) | 0.7576 | 3.958 | 8.09e-05 \*\*\* |
|  | higheryes | 1.4940 | 3.639 | 0.000288 \*\*\* |
|  | studytime2 | 2.5596 | 2.285 | 0.022504 \* |
|  | studytime3 | 7.4010 | 3.514 | 0.000462 \*\*\* |
|  | studytime4 | -1.7682 | -0.803 | 0.422394 |
|  | schoolsupyes | 0.4547 | 0.306 | 0.760012 |
|  | Dalc2 | -0.7863 | -2.727 | 0.006511 \*\* |
|  | Dalc3 | 0.1068 | 0.234 | 0.814862 |
|  | Dalc4 | -1.3768 | -1.975 | 0.048525 \* |
|  | Dalc5 | 0.2097 | 0.283 | 0.777071 |
|  | health2 | 0.7798 | 0.916 | 0.360150 |
|  | health3 | -0.9041 | -1.203 | 0.229152 |
|  | health4 | 0.2467 | 0.321 | 0.747981 |
|  | health5 | -0.4384 | -0.675 | 0.499822 |
|  | romanticyes | -0.59306 | -2.574 | 0.010190 \* |
|  | famsizeLE3 | 0.48561 | 2.040 | 0.041625 \* |
|  | goout2 | 1.15597 | 2.460 | 0.014070 \* |
|  | Goout3 | 0.67978 | 1.482 | 0.138527 |
|  | Goout4 | 0.38442 | 0.800 | 0.423919 |
|  | Goout5 | 0.08698 | 0.173 | 0.862617 |
|  | failures:schoolsupyes | 1.4629 | 2.773 | 0.005665 \*\* |
|  | studytime2:health2 | -2.2138 | -2.182 | 0.029361 \* |
|  | studytime3:health2 | -1.4987 | -0.922 | 0.356618 |
|  | studytime4:health2 | 0.2277 | 0.111 | 0.911696 |
|  | studytime2:health3 | -0.5703 | -0.627 | 0.530916 |
|  | studytime3:health3 | -1.9426 | -1.360 | 0.174180 |
|  | studytime4:health3 | -1.1523 | -0.743 | 0.457822 |
|  | studytime2:health4 | -1.4005 | -1.492 | 0.135926 |
|  | studytime3:health4 | -2.7377 | -1.897 | 0.058132 |
|  | studytime4:health4 | -2.1322 | -1.039 | 0.299055 |
|  | studytime2:health5 | -0.6203 | -0.779 | 0.436179 |
|  | studytime3:health5 | -2.3509 | -1.742 | 0.081736 |
|  | studytime4:health5 | -0.2144 | -0.137 | 0.891296 |
|  | studytime2:goout2 | -1.1239 | -1.066 | 0.286582 |
|  | studytime3:goout2 | -4.2356 | -2.345 | 0.019244 \* |
|  | studytime4:goout2 | 4.9803 | 2.287 | 0.022411 \* |
|  | studytime2:goout3 | -1.3213 | -1.287 | 0.198229 |
|  | studytime3:goout3 | -3.9000 | -2.190 | 0.028790 \* |
|  | studytime4:goout3 | 3.3769 | 1.517 | 0.129630 |
|  | studytime2:goout4 | -1.8888 | -1.793 | 0.073252 . |
|  | studytime3:goout4 | -5.1949 | -2.719 | 0.006670 \*\* |
|  | studytime4:goout4 | 1.8847 | 0.696 | 0.486475 |
|  | studytime2:goout5 | -1.7584 | -1.618 | 0.106088 |
|  | studytime3:goout5 | -3.2727 | -1.636 | 0.102214 |
|  | studytime4:goout5 | 2.3971 | 1.094 | 0.274432 |
|  | schoolsupyes:health2 | 2.3211 | 1.642 | 0.100989 |
|  | schoolsupyes:health3 | 2.3842 | 1.892 | 0.058844 . |
|  | schoolsupyes:health4 | 2.9360 | 2.304 | 0.021438 \* |
|  | schoolsupyes:health5 | 2.9259 | 2.471 | 0.013625 \* |
|  | schoolsupyes:goout2 | -5.5980 | -4.165 | 3.38e-05 \*\*\* |
|  | schoolsupyes:goout3 | -4.5044 | -3.316 | 0.000948 \*\*\* |
|  | schoolsupyes:goout4 | -3.8287 | -2.814 | 0.004993 \*\* |
|  | schoolsupyes:goout5 | -5.5483 | -3.573 | 0.000370 \*\*\* |

**Table 6**

lm(G3 ~ (failures+I(failures^2)+higher+studytime+schoolsup+Dalc+health+romantic+famsize +goout+failures:schoolsup+studytime:health+studytime:goout +schoolsup:health+schoolsup:goout), data = studentDataset)

Since we chose the interaction model with higher order terms as the best fit model for our project, the final model is:

A screenshot of a computer

Description automatically generated

# Assumption Verification

All the liner regression models are bases on some assumptions and if these assumptions are not met then the model itself is considered invalid. So, it is very important to verify all the assumption for our final model.

1. **Linearity Assumption:** The linear regression model we build is based on the assumption that there is a linear relation between predictors and response variable. To confirm the linear relation, we can use Residual plot.

A graph with dots and lines

Description automatically generated

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.

A graph with purple dots

Description automatically generated

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

H0 : Heteroscedasticity is not present (error term has common variance)  
Ha : Heteroscedasticity is present (error term do not have common variance)

A screenshot of a computer

Description automatically generated

Since the **p-value is 0.3235** higher than 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.

A graph and a graph

Description automatically generated with medium confidence

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

A screenshot of a computer program

Description automatically generated

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

The data we are working with has many identical values (many factors are between 1-5) and Shapiro-Wilk test do not work well for such dataset. (Kolmogorov–Smirnov test, n.d.)

1. **Independence Assumption:** Since all the students are independent to each other, the data will also be independent, and we can safely assume that we meet the independence assumption.
2. **Multicollinearity:** To detect multicollinearity, we use VIF to identify correlations between variables.

A screenshot of a computer

Description automatically generated

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.

A graph of a number of black dots

Description automatically generated

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.

### 

# Conclusion

In our final model we can see that the coefficients of some of the parameters, like higheryes (student’s desire to go for higher study), schoolsupyes (Extra support from school), health (health of student), goout (measure of time student spent with friends) are positive, this indicate that these factors has good impact on the academic performance of the student. While some other variables, like romanticyes (student is involved in romantic relation), failure (number of past class failures), Dalc (frequency of alcohol consumption) has negative coefficient indicating that they has bad effect on the academic performance of student.

Since we are meeting all the assumptions of linear regression model we can conclude that our model is good, but the Adjusted R2 of our final model is just 23.84% which means that only our model can explain only 23.84% of variation in the dependent variable (G3 - final grade) which is not very good.

The model may not be good for prediction, but it helped us in concluding that demographic and social factors do play some role in the academic performance of a student.

# Discussion

The academic performance of a student is majorly dependent on student’s own capability and to identify the demographic and social variables/conditions which can help students, we will need much bigger dataset. Our dataset had only 649 rows and the data were collected for students of 2 public school only. To successfully conduct this experiment.

For further study we can try some other model building techniques to come up with a better model, we can collect more data from different locations so that our data is not biased.

# Bibliography

*Kolmogorov–Smirnov test*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Kolmogorov–Smirnov\_test#:~:text=Various%20studies%20have%20found%20that,samples%20with%20many%20identical%20values.

*Student Performance*. (n.d.). Retrieved from UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/dataset/320/student+performance

# Appendix I

1. Full model test

A screenshot of a document

Description automatically generated

A screenshot of a computer

Description automatically generated

1. ANOVA test between null model and full model

A screenshot of a computer

Description automatically generated

1. Forward Selection Procedure

A screenshot of a computer

Description automatically generated

1. Dropping non-significant variables from forward selection process.

A screenshot of a computer

Description automatically generated

1. Interaction model.

A close up of a text

Description automatically generated

1. Higher Order

A group of buildings with different colored dots

Description automatically generated with medium confidence

1. ANOVA test between interaction model and interaction model with higher order

A computer screen shot of text

Description automatically generated

1. ANOVA test between the best additive model and interaction model with higher order.

A computer screen shot of a code

Description automatically generated