Statistic Model to Analyze Student’s Performance

Xuchuan Zheng, Sungki Park, Prashant Sharma

# Introduction

In out 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 in order 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.

To answer these questions, we are working with a dataset that is collected at 2 Portuguese schools for Mathematics and Portuguese subject. This data is collected by using school reports and questionnaires. The data attribute includes student’s grades, family size information, education level of parents, free time of student, any many other factors. By working on this project, we are hoping to develop more understanding about the factors which can impact the performance of a student.

# Methodology

Data Source

Variable Explanations and Data Assumptions

This data is from [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/dataset/320/student+performance). There are 649 rows instances and 30 features in the dataset. Below are details of each feature

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) [Qualitative]

2. sex - student's sex (binary: 'F' - female or 'M' - male) [Qualitative]

3. age - student's age (numeric: from 15 to 22)

4. address - student's home address type (binary: 'U' - urban or 'R' - rural) [Qualitative]

5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) [Qualitative]

6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart) [Qualitative]

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) [Qualitative]

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) [Qualitative]

9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') [Qualitative]

10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other') [Qualitative]

11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') [Qualitative]

12. guardian - student's guardian (nominal: 'mother', 'father' or 'other') [Qualitative]

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) [Qualitative]

14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) [Qualitative]

15. failures - number of past class failures (numeric: n if 1<=n<3, else 4) [Qualitative]

16. schoolsup - extra educational support (binary: yes or no) [Qualitative]

17. famsup - family educational support (binary: yes or no) [Qualitative]

18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) [Qualitative]

19. activities - extra-curricular activities (binary: yes or no) [Qualitative]

20. nursery - attended nursery school (binary: yes or no) [Qualitative]

21. higher - wants to take higher education (binary: yes or no) [Qualitative]

22. internet - Internet access at home (binary: yes or no) [Qualitative]

23. romantic - with a romantic relationship (binary: yes or no) [Qualitative]

24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent) [Qualitative]

25. freetime - free time after school (numeric: from 1 - very low to 5 - very high) [Qualitative]

26. goout - going out with friends (numeric: from 1 - very low to 5 - very high) [Qualitative]

27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) [Qualitative]

28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) [Qualitative]

29. health - current health status (numeric: from 1 - very bad to 5 - very good) [Qualitative]

30. absences - number of school absences (numeric: from 0 to 93) [Quantitative]

31. G1 - first period grade (numeric: from 0 to 20) [Quantitative]

32. G2 - second period grade (numeric: from 0 to 20) [Quantitative]

33. G3 - final grade (numeric: from 0 to 20, output target) [Quantitative]

Modeling Plan

# Result

Variable Selection Procedures

Main Effects Individual T-tests:

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