Students' grades analysis and prediction

Andrej Čerňanský & Dávid Števaňák

Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
Jún 2023

Contents

0.1 Introduction

The goal of our project is mainly focused on analyzing data and building up the most important factors which have impact on the performance of the students, from a dataset we found on kaggle.com. We are both interested in social sciences and psychology and wanted to deal with real life data in our final project so we tried to come up with some general hypothesis and then used this dataset to either support or reject them. Since both of us had a similar view on the matter that good grades cannot simply be explained by genetics or paid classes and therefore such thorough survey where students gave information about the quality of their family relationships, their alcohol consumption, absences and so on was a clear choice for us. Unfortunately, the data were not from a longer term so we could see some trends in students' behaviour or just look what might have caused the change in the grades. All in all, this dataset provides a really detailed view into students' lives and around 400 entries might be considered a sufficient amount of data but the results we conclude from it should not be generalized.

Chapter 1

Presenting our dataset

This data approaches student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. by Paulo Cortez from the department of Information Systems, University of Minho, Portugal, the author of the book *Modern Optimization with R*. There are 395 entries from the students from the two high-schools: Here is a brief explanation of each of the variables in the dataset: Gabriel Pereira and Mousinho da Silveira.

- **school** - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- **sex** - student's sex (binary: 'F' - female or 'M' - male)
- **age** - student's age (numeric: from 15 to 22)
- **address** - student's home address type (binary: 'U' - urban or 'R' - rural)
- **famsize** - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- **Pstatus** - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- **Mjob** - mother's job (nominal: 'teacher', 'health' care related, civil 'services' , 'at_home' or 'other')
- **Fjob** - father's job (nominal: 'teacher', 'health' care related, civil 'services' , 'at_home' or 'other')
- **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- **guardian** - student's guardian (nominal: 'mother', 'father' or 'other')
- **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)
- **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

- **failures** - number of past class failures (numeric: n if 1≤n<3, else 4)
- **schoolsup** - extra educational support (binary: yes or no)
- **famsup** - family educational support (binary: yes or no)
- **paid** - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- **activities** - extra-curricular activities (binary: yes or no)
- **nursery** - attended nursery school (binary: yes or no)
- **higher** - wants to take higher education (binary: yes or no)
- **internet** - Internet access at home (binary: yes or no)
- **romantic** - with a romantic relationship (binary: yes or no)
- **famrel** - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- **freetime** - free time after school (numeric: from 1 - very low to 5 - very high)
- **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
- **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- **health** - current health status (numeric: from 1 - very bad to 5 - very good)
- **absences** - number of school absences (numeric: from 0 to 93)
- **G1** - first period grade (numeric: from 0 to 20)
- **G2** - second period grade (numeric: from 0 to 20)
- **G3** - final grade (numeric: from 0 to 20, output target)

1.1 Methods used and its justifications

For the data analysis and finding the corresponding correlations, we used R, not only for its nice plots (library ggplot2) but also for easy writing into SQL databases. We built a flask app to sum up and visualise those results and also to build an interactive way to help the user understand and explore the dataset more easily. For the charts we used Google Charts as well as some Python libraries such as plotly and pandas. The rest was done using Bootstrap and some basic JavaScript.

1.2 Some basic information about the dataset

By plotting a couple of pie charts we got a clearer view on the dataset and could amend our goals and expectations accordingly. Seeing that only 28% of the students are from the *Mousinho da Silveira* school, we realised that it would not be the best idea to compare results between the two schools since this could be simply biased by the size of each of the groups. We also noticed that the dataset does not contain a large number of recipients (only 395 students), so the results are rather restricted to this dataset and should not be interpreted in general. However, categories such as gender and age are pretty balanced when considering high school students only (there is a small portion of students who are 19+ years old but those were not relevant in our study).

Chapter 2

Analysis

2.1 T-test results

We performed t-tests by splitting the students into two categories. For binary variables it was rather trivial, for numeric/categorical variables we decided to split them into students above and below mean in the given variable. P-value was less than 0.05 in 13 of the observed variables when considering the mean from the G3 test. Namely in *sex* where the mean for females is 9.9663 and 10.9144 for males. *Address* of the students with mean of rural 9.5114 and urban of 10.6743. Other variables were of a FALSE/TRUE type or below/above mean and are summed up in the following table:

Category	**False/Below mean**	**True/Above mean**
schoolsup	10.5610	9.4314
paid	9.9860	10.9227
Medu	9.4121	11.1348
Fedu	9.8342	11.0051
studytime	10.1287	11.3587
failures	11.2532	7.2651
goout	10.9570	9.4173
traveltime	10.7821	9.7319
Dalc	10.7319	9.6807
G1	7.2487	13.4406
G2	6.9740	13.6700

Table 2.1: Means in G3 for considered variables

The most drastic differences in means were in Mother's education, Past failures and in G1, G2 columns.

2.2 Grades distribution

Judging by the graph 2.1, where the y-axis represents the count of students that received a specific grade, the distribution of the grades appears to be normal (the most common value from the range 0 to 20 is 10), but only if we exclude the students who had 0 points. There were 38 students who scored 0 points from the test but unfortunately, from the description of the dataset we were not able to tell what factor could cause this. Some might have had health problems, but others could probably scored the grade simply due to the lack of preparation.

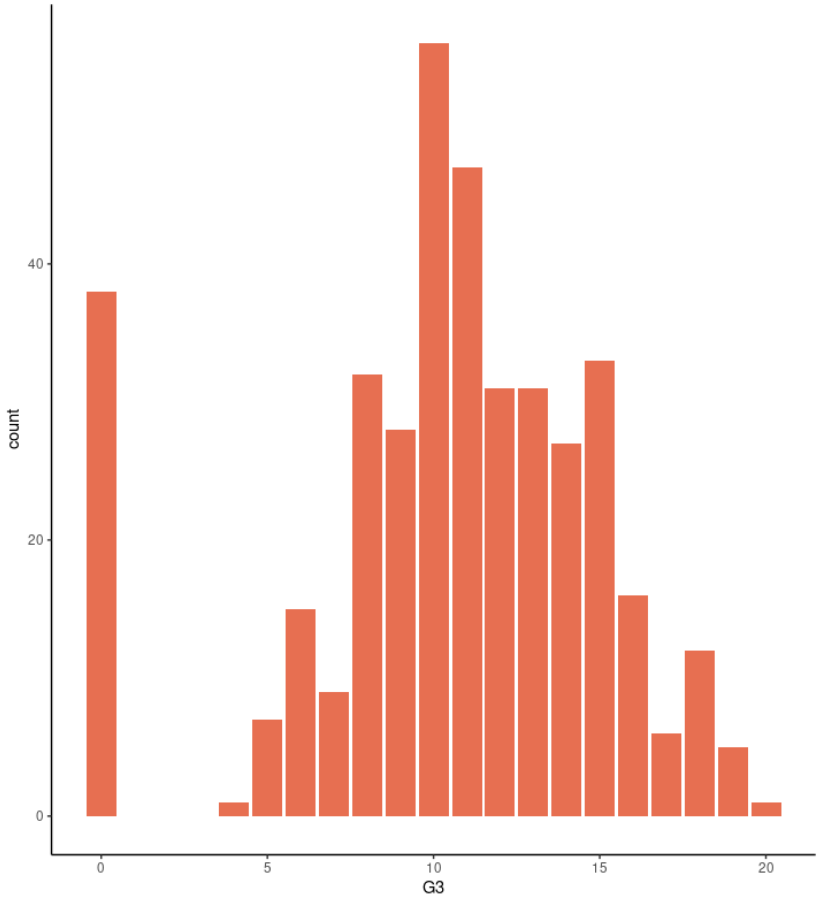

Figure 2.1: Grades distribution

2.3 Parents' education impact

We were curious whether the acquired education of parents plays a major role in the results of their children. In the graph 2.2 the mean scores of the groups from the G3 test is displayed - the color of the dot (the lighter the better) and the size of the group is represented by the size of the dots (so we know how significant it is). All in all, the means are rising diagonally (meaning the higher the education of both parents the better grades), so this supports our hypothesis.

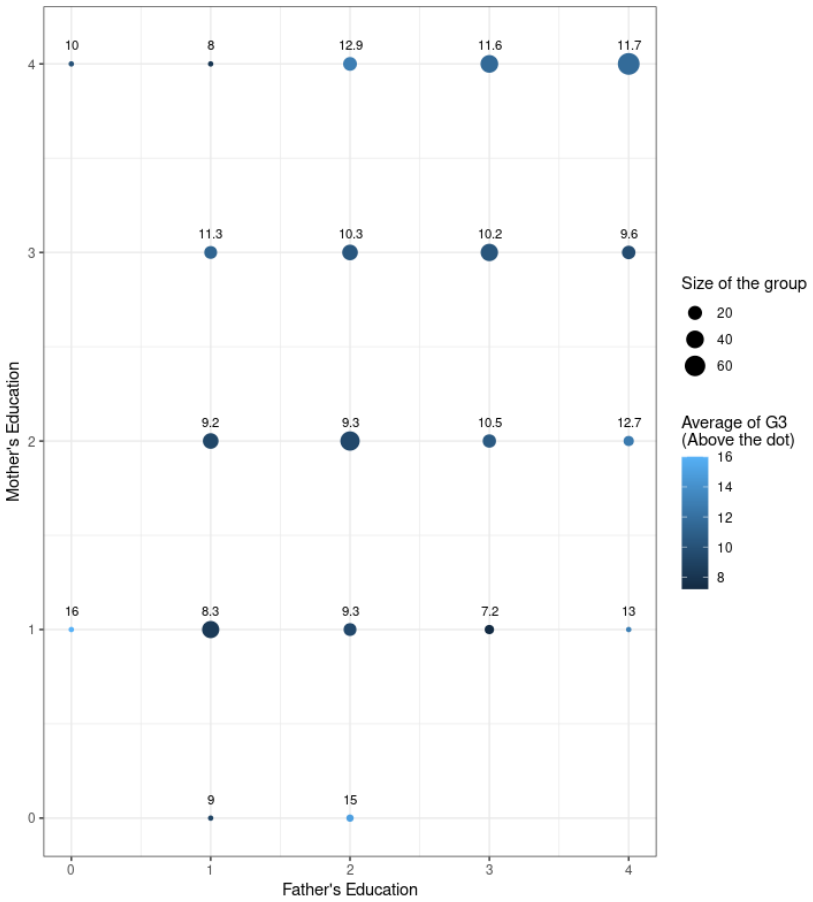

Figure 2.2: Parents' education impact

2.4 Family size impact

Another hypothesis was that having smaller family could mean more time for the parents to spend it helping their children with their studies. Intuitively, this makes sense but we wanted to use the data to support this. Results are displayed in the graph 2.3. On the right, there are students, which come from families of size 3 or less and on the left there are students with families of size bigger than 3. The dots are coloured by the fraction (number of students with that amount of points from one of the groups divided by the size of the group). Seeing that the colors do not differ noticeably in the groups, we might say that the size of the family does not significantly impact the grades.

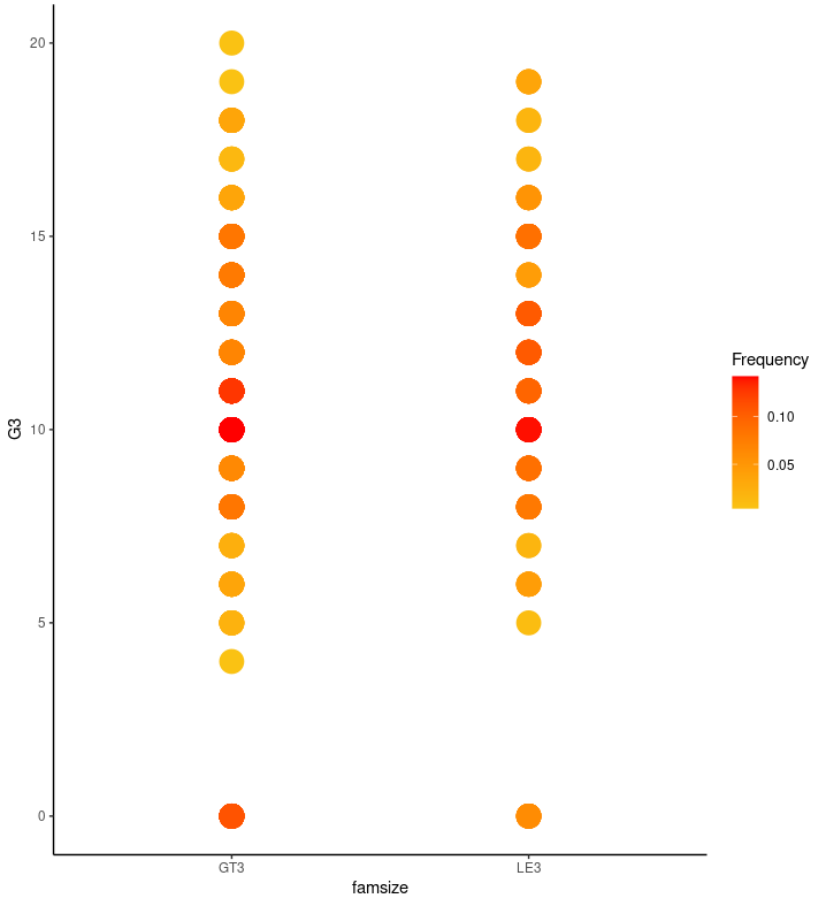

Figure 2.3: Family size impact

2.5 Comparing the worst and the best

Another way of determining what makes a difference in students' results is looking at the best/worst students and looking at biggest differences between them. In the following, we will look at the students divided into 4 groups - those with more than 17 points from the G3 test, all students, those with zero points and students with 4 to 6 points. For the simplicity, we will mention only the most significant differences in the means. The best students' mean of mother's education was higher by 0.7 compared to the overall mean (3.45 to 2.75). Unsurprisingly, students with zero points had a bigger amount of failures (0.92 compared to 0.334 overall). The majority of the best students are not in the relationship (mean of 0.055) compared to 0.33 overall mean (0.525 for the students with zero points). And of course, the best students were far better in G1 (17.389, 10.909, 7.526, 6.487) and G2 tests (17.95, 10.71, 4.6, 5.9) categories are in the above mentioned order.

Chapter 3

Linear and logistic regression

3.1 Correlation matrix analysis

Looking at the correlation matrix 3.1 we might say that the correlations between the variables are not strong and many of them are not much (linearly) dependent on each other. Mostly interconnected are mother's and father's education variables (0.62) along with weekend alcohol consumption which is also quite dependent on workday alcohol consumption (0.65). Nevertheless, those were not really helpful when trying to make the regression model. For G3 variable (the one that we look the most at) the negative correlation is with past failures (-0.36), and, as would be expected, the grades from previous tests - G1(0.8) and G2 (0.9).

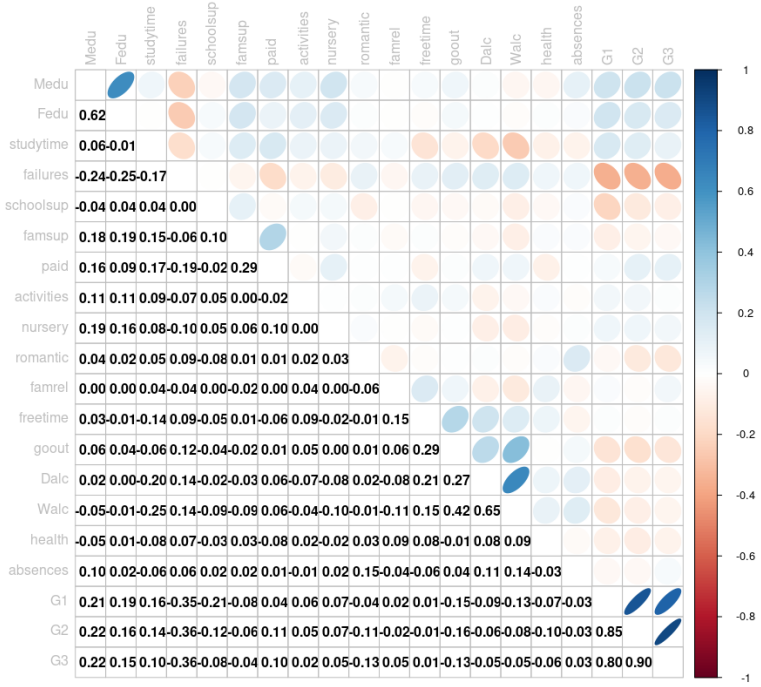

Figure 3.1: Correlation matrix

3.2 Linear model

We might assume that there is a linear relationships between the variables, so using LM could be appropriate, however, we know that G3 is a discrete variable ranging from 0 to 20, so linear regression might not be the best way of predicting results, but in the following we will use logistic regression as well. We are also right to think that the observations are independent of each other.

3.2.1 Results

We created two models - one predicting G3 results based only on G1 and G2 scores. This models' multiple R-squared score was 0.822. The variable G2 seemed to play a bigger role owing to the fact that its coefficient was higher. The second model consisted of all numerical variables, and got the multiple R-squared score of 0.832. The most important variables were G1, G2 and quality of family relationships.

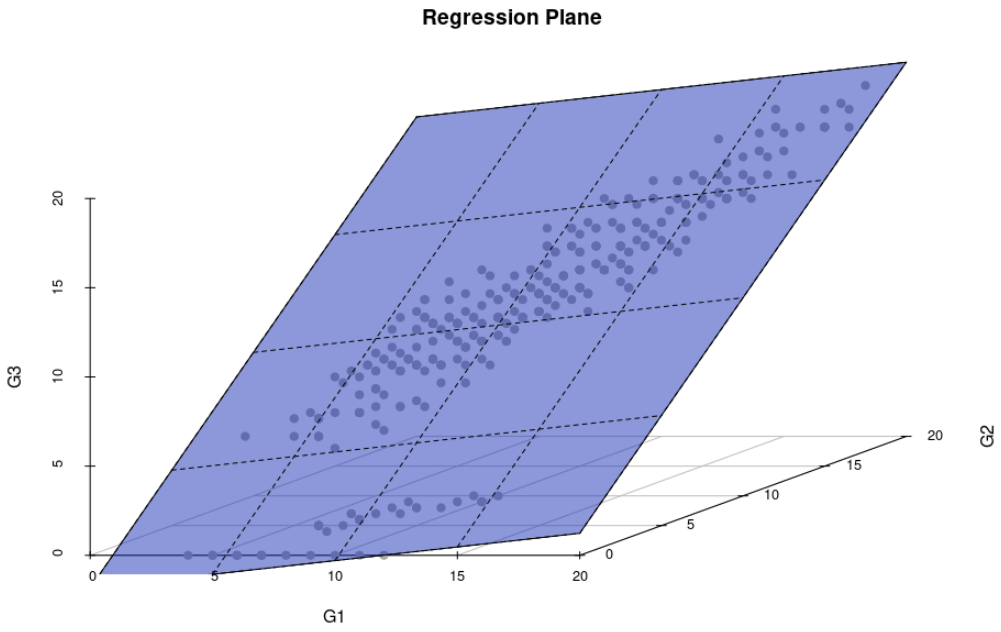

Figure 3.2: Linear model

3.3 Logistic regression

We can easily do a binary classification on the results of the G3 test. More than or equal to 10 points means passing the test, less than 10 points means fail and all of that is under the assumption that there are linear relationships between the variables. It is justified to use logistic regression to look for the importance of the variables.

3.3.1 Results

Coefficients:

Category	**Estimate**	**Std. Error**	**z value**	**Pr($>|z|$)**	**Significance**
(Intercept)	-26.23177	4.57218	-5.737	9.62e-09	***
sex	-1.38202	0.69475	-1.989	0.046675	*
Medu	0.33316	0.34572	0.964	0.335206	
Fedu	-0.97344	0.39981	-2.435	0.014903	*
studytime	-0.78216	0.39538	-1.978	0.047901	*
failures	-0.02564	0.34353	-0.075	0.940499	
schoolsup	0.05325	0.66280	0.080	0.935963	
famsup	-0.04166	0.56186	-0.074	0.940891	
paid	0.05974	0.59050	0.101	0.919421	
higher	-0.20051	1.27182	-0.158	0.874730	
romantic	-0.69137	0.57037	-1.212	0.225459	
famrel	1.15543	0.36508	3.165	0.001551	**
goout	-0.89136	0.29960	-2.975	0.002928	**
Dalc	-0.69721	0.38480	-1.812	0.070003	.
Walc	1.13231	0.34058	3.325	0.000885	***
absences	-0.05204	0.03029	-1.718	0.085846	.
G2	2.53692	0.43872	5.782	7.36e-09	***
G1	0.43326	0.21239	2.040	0.041361	*

Table 3.1: Results for considered variables

We can see that the most important factors which cause passing the test are the scores from G1 and G2, the quality of family relationships and what is odd, weekend alcohol consumption. Going out with friends does have a negative impact but oddly enough studytime somehow also belongs to this group.

Chapter 4

Our impressions

The most difficult part of the project was determining what could and what coludn't be done with the dataset since all variables are discrete, categorical or binary. Luckily there were no missing values within the data-frame and all columns seemed to be useful, so we didn't do much pre-processing of the data because it wasn't necessary. For more smooth calculation we needed to transform factor variables yes/no to 1/0 variables. Therefore to use binary classification and t-test was a good call. Working with dataframes in R was a smart choice and within few lines of code we were able to get out of the data what we wanted. Also the writing to the SQL3 database was trivial. Having also the code snippets for interactive java script graphs integrating them into our flask application was again quite easy. For the Flask app part there were some complications with adjusting the little details in the graphs such as background color or changing a color of a plane etc. - the solution itself was not that complicated but it took some time to dig it up in the documentation especially when it was the specific details. The same was for the web page but with that it was usually that repetitive work that need to be done until it finally works. Surprisingly enough, the Google charts were easy to implement and adjust as well as connecting the database to the page which went absolutely smoothly. I would probably focus more on plotting the graphs using Python rather than R even though R's libraries might be more intuitive to work with but making the graphs interactive on a webpage looks much better than just simple .png pictures.