ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS
ADVANCED DATA ANALYSIS WITH R

A-15 Student Performance Data Set

Panagiotis Symianakis

M.Sc. in Statistics

Athens,Greece
January 6, 2022

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

simianakisp@aueb.gr
f3612120

Contents

List of Figures

List of Tables

1 Introduction

According to UNESCO, *"education is the process of facilitating learning or the acquisition of knowledge, skills, values, beliefs and habits"*. Most people believe that all children are educated but there are really poor countries where education is not given. Schooling cultivates peace, reduces poverty and grows sustainable development. It cultivates critical thinking and unaffected thinking. Also, refines the inner world of man, sensitizes his soul and creates a moral character. The necessity and usefulness of education based on the above is considered self-evident. However, the effectiveness and performance of each student is influenced by various factors that we will consider below according to the given data set.
More specifically, the Portuguese education system consists of three levels. The basic which lasts 9 years, the second that lasts 3 years and the higher. The first two are mandatory and free. When students start the second level they have to choose the direction of the courses they will follow for their professional career. The categories are: a) science-humanities courses, b) vocational courses and c) other education and training provision. The evaluation is carried out in a formative and cumulative way. Moreover, the highest grade students can achieve is 20 and the lowest is 0. They are examined in two periods (variables G1,G2) and the last evaluation (G3) describes the final grade.
Above is a table with the names of given variables and their description.

Attribute	Description
school	student's school (binary: Gabriel Pereira or 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: urban or rural)
famsize	family size (binary: ≤ 3 or > 3)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4a)
Fedu	father's education (numeric: from 0 to 4a)
Mjob	mother's job (nominalb)
Fjob	father's job (nominalb)
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 $1\leq n<3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes (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)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday 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)
Attribute	Description
	the grades above are related with the course subject
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)

Table 1: Each variable's description of the given dataset.

The given data set (14-15 student.zip) comes from two Portuguese secondary schools. As the reader can understand from the above table, it contains student's grades at the "Portuguese language" subject, demographic, social and school related features which were collected through reports, based on paper sheets, and questionnaires with closed questions. Below will be an extensive analysis that will include:

- explanatory analysis for the dataset
- visual description for the most important findings
- pairwise associations
- regression models to asses the responses

2 Analysis for the dataset

Initially, the total number of students who participated in the research is 649, of which 383 are boys and the remaining 266 are girls. There are 423 students in the "Gabriel Pereira" school, which is 197 more than in the "Mousinho da Silveira" school. In particular, to facilitate our analysis, we made the distinction between the two schools.

School	Gabriel Pereira	Mousinho da Silveira
Female	237	146
Male	186	80
Urban	345	107
Rural	78	119
Family size ≤ 3	122	70
Family size >3	301	156
Parents live apart	55	25
Parents live together	368	201

Table 2: Differences between two schools

It's really important to notice that there are 151 students who do not have access to the internet. This factor expects someone to affect the effectiveness of the students as on the internet one can find a lot of information about the lessons, solved exercises and additional material. Furthermore, some other factors that affect a student's performance are the help he receives from his parents at home and whether he attends additional educational classes. Both offer dynamism and self-confidence to the student and he is more likely to succeed. According to the given data there only 68 out of 649 students who enjoy extra educational support. On the other hand, there seems to be uniformity in the numbers regarding family help. More specifically, 398 children receive help while 251 don't.

Finally, we will make an analysis about the children who passed or failed the course. After a separation of the students we found that 549 passed the course and the others 100 didn't make it. Below is a table with some characteristics of the students who passed the lesson or not.

		Pass	Fail
School	Gabriel Pereira	391	32
	Mousinho da Silveira	158	68
Gender	Female	333	50
	Male	216	50
extra educat. sup	Yes	60	8
	No	489	92
extra family sup	Yes	341	57
	No	208	43
Intenet access	Yes	430	68
	No	119	32

Table 3: Differences between students who passed and failled

In the above analysis, some factors first appeared that we believe play an important role in shaping a student's grade. Other factors that have been omitted but may affect a child's performance in relation to his or her image at school may later appear. Our data contains 33 variables and we hope several of them will be useful for us to be able to predict students' grades.

3 Visual methods to describe the dataset

In this section we aim to visualize the dataset so that the reader is able to draw some useful conclusions from the graphs that will follow. It is important to observe the differences that exist in the scale as well as what is presented in each plot.

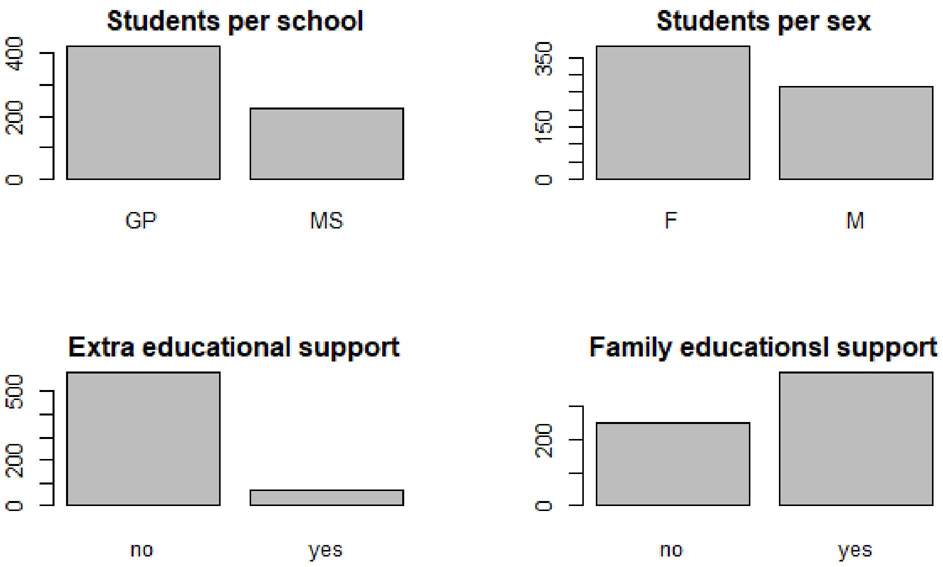

Figure 1: Students per school, sex and extra support

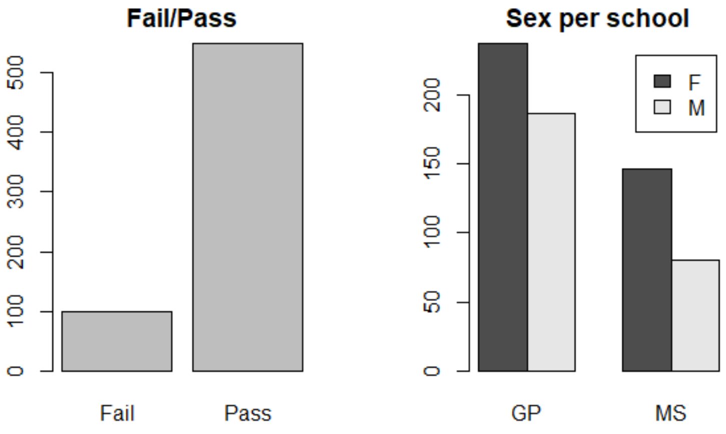

Figure 2: Fail/Pass & Sex per school

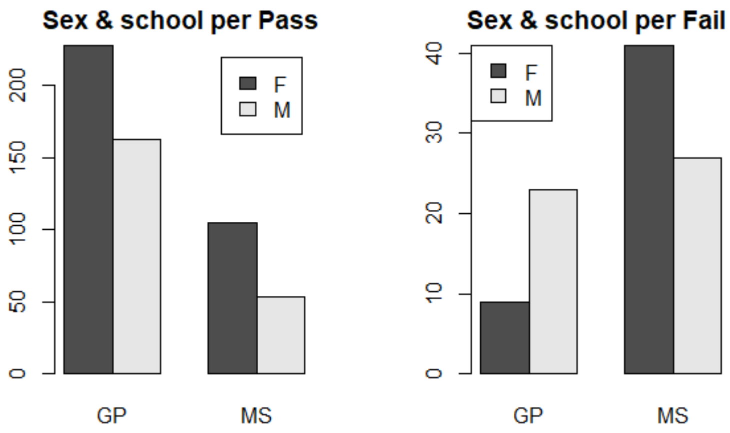

Figure 3: Sex and School per Fail/Pass

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. This definition does not strictly limit the analyst to what would be considered strange. On the contrary, it gives him the freedom to judge for himself what is normal and what is not. Boxplot is used to view outliers in the numerical data.

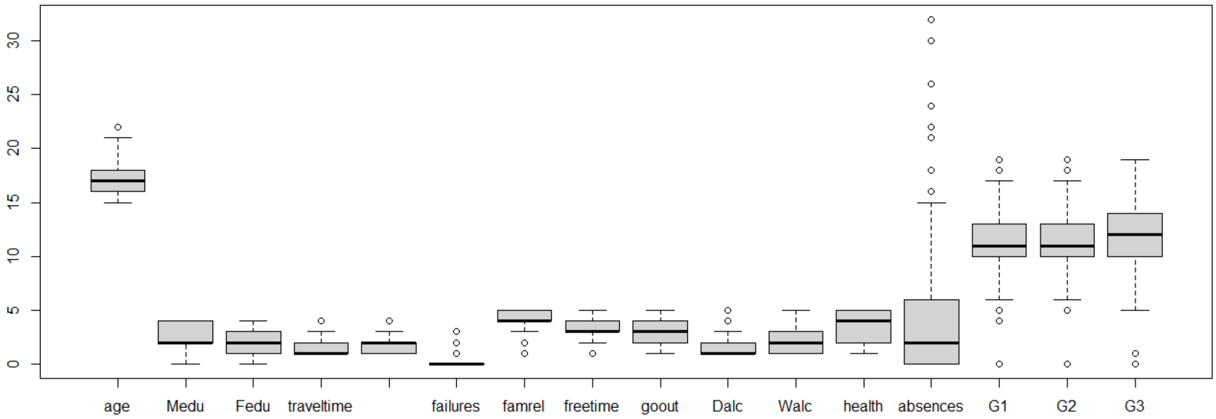

Figure 4: Outliers of nueric variables

Most outliers are observed in the variable that shows a student's absences **absences**. It is also expected as some children may were seriously ill and had to be absent for a long time. In general, the outliers do not seem to be many in other variables but the degrees that have outliers at both tails.
Now we will present two pie plots with the mother's and father's education level. It is significant to outline that amorphous level has been eliminated and knowledge and education are becoming even more important factors in life.

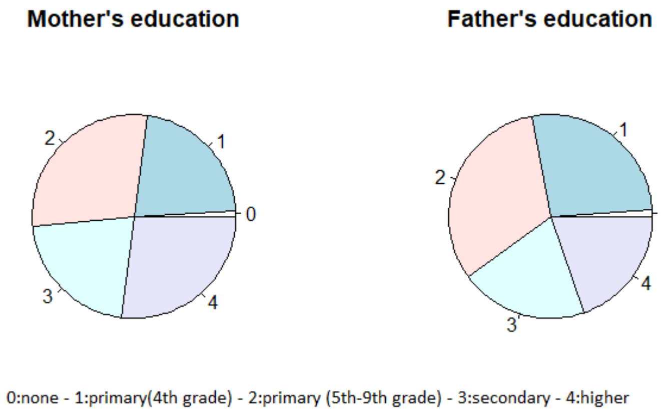

Figure 5: Mother's and Father's education level

4 Pairwise Associations

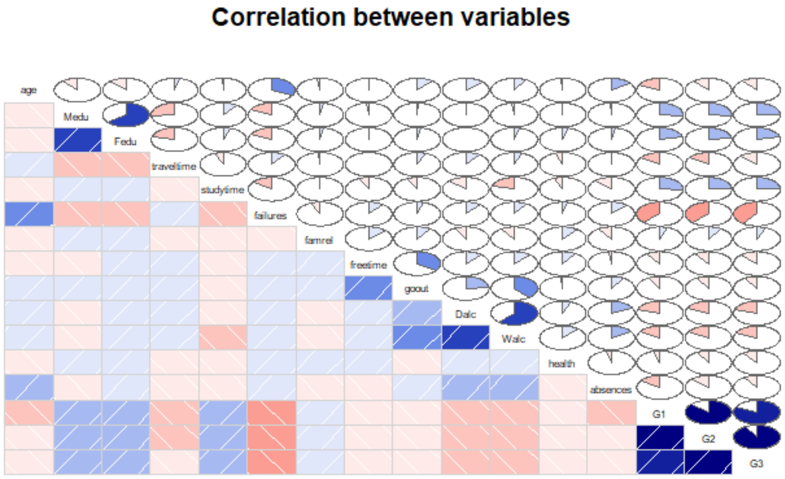

Figure 6: Pairwise Correlation

In this section we will deal with the relationships presented between the variables in pairs. Initially, as it is well known, one can observe the great correlation between the three grades (**G1**,**G2**,**G3**). Also, from the graph it seems that there is a great correlation between the education of the mother (**Medu**) and the father(**Fedu**), which look like they are correlated with the grades **G1**,**G2** and **G3**. Furthermore, the grades look like they get affected by the time a child spends to study.

Moreover,noteworthy is the small correlation that exists between daily alcohol consumption (**Dalc**) and that of the weekend (**Walc**). This is to be expected as both values are correlated with the time students spend outside the home (**goout**).

On the other hand, the number of failures has a minimal negative correlation with the time of studying and mother's and father's education level. Finaly, there seems to be a higher negative association of the failures with grades. The others variables do not seem to have significant influence each other.

5 Normality, Skewness and Kurtosis

Normality
The Shapiro–Wilk test tests the null hypothesis that a sample $x_1, x_2, ..., x_n$ came from a normally distributed population. The test statistic is:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

It is important not to confused the $x_{(i)}$ which the ith-smallest number in the sample with the x_i which is the ith observe of the sample. By the q-q plots which used and Saphiro-Wilk test we can not assume normality.
Skewness
Formula for skewness:

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}}$$

If the coefficient of skewness is greater than 0 , then the graph is said to be positively skewed with the majority of data values less than mean. Most of the values are concentrated on the left side of the graph.
If the coefficient of skewness is equal to 0 or approximately close to 0 , then the graph is said to be symmetric and data is normally distributed.
If the coefficient of skewness is less than 0 ,then the graph is said to be negatively skewed with the majority of data values greater than mean. Most of the values are concentrated on the right side of the graph.
It is observed that the values of the variables **failures**, **absences**, **Dalc** and **traveltime** show positive skewness. On the contrary, the variables **famrel**, **G3** and **health** are negatively skewed.
Kurtosis
Formula for Kurtosis

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}$$

There exist 3 types of Kurtosis values on the basis of which sharpness of the peak is measured:
Platykurtic. If the coefficient of kurtosis is less than 3 i.e. $\gamma_2 < 3$, then the data distribution is platykurtic. Being platykurtic doesn't mean that the graph is flat-topped.
Mesorkurtic. If the coefficient of kurtosis is equal to 3 or approximately close to 3 i.e. $\gamma_2 = 3$, then the data distribution is mesokurtic. For normal distribution, kurtosis value is approximately equal to 3.
Leptokurtic. If the coefficient of kurtosis is greater than 3 i.e. $\gamma_2 > 3$, then the data distribution is leptokurtic and shows a sharp peak on the graph.
According to "R"'s calculation we imply that:

Platykurtic	Mesokurtic	Leptokurtic
Medu Fedu freetime goout Walc health	age studytime G1	traveltime failures famrel Dalc absences G2,G3

Table 4: Kurtosis numeric variables

6 Regression Models

In this part, we are going to fit some regression models in order to understand better which variables are most important in extracting the final grade. It's known that the variables **G1**,**G2** and **G3** are highly correlated as **G1** and **G2** correspond to the first and second period grades. It is more difficult to predict **G3** without taking into account **G1** and **G2**, but it is what we will try.
First of all, LASSO (Least Absolute Shrinkage and Selection Operator) regression was held. LASSO regression is a type of linear regression that uses shrinkage. The lasso process encourages simple models with fewer parameters. It is based on the l_1 penalization.

$$minimize(\mathbf{y} - \mathbf{Z}\beta)^T(\mathbf{y} - \mathbf{Z}\beta)$$

$$s.t. \sum_{j=1}^{p} |\beta_j| \leq t$$

$$minimize\{(\mathbf{y} - \mathbf{Z}\beta)^T(\mathbf{y} - \mathbf{Z}\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

Steps for LASSO:

1. Run LASSO for a variety of values
2. Plot the regularization paths
3. Implement k-fold regularization
4. Estimate the coefficients using λ with minimum CV-MSE

It is important to depict the LASSO's result for each grade.

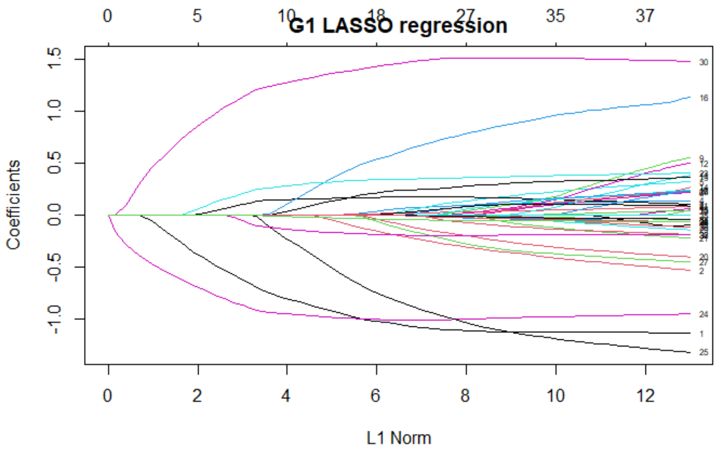

Figure 7: LASSO for G1

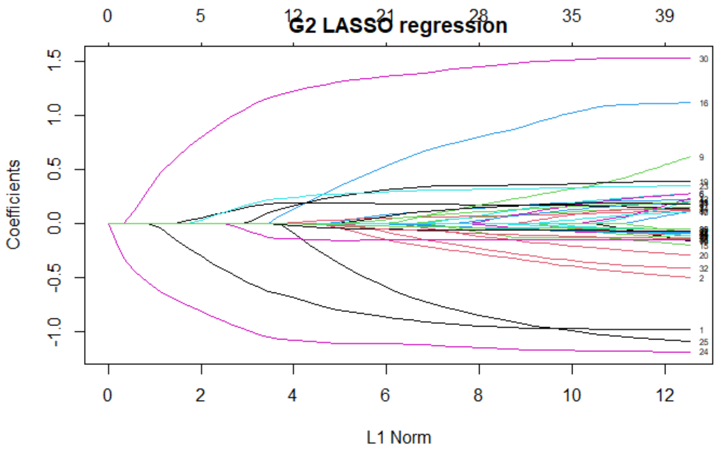

Figure 8: LASSO for G2

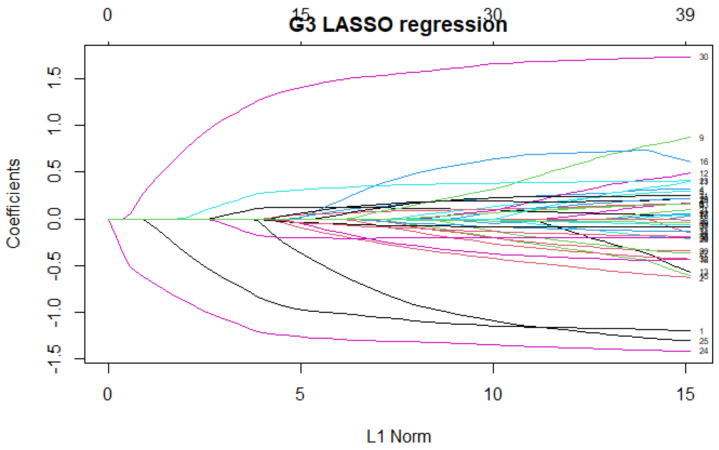

Figure 9: LASSO for G3

Through LASSO regression it is easy for the reder to understand which variables except the grades can affect the model we want to construct.

After that stepwise procedure and backword procedure were used in order to choose the best variables that should be contained in the model. The Bayes Information Criterion (BIC) was used as it is more suitable for prediction models. $\boxed{BIC : -2log\mathcal{L} + dlog(n)}$.

Stepwise procedure

It is step by step procedure of adding and removing variables. It started from the full model, excluding the other two grades, and in every step it was checked which covariate to include according the BIC. It takes place until no further improvment can be achieved.

Backward procedure

It is step by step removal of insignificant variables. It starts from the full model, excluding the other two grades, and in every step we check which covariate should be excluded according to the model which minimizes the BIC. In this process a variable which has been removed can not re-include in the model.

After several attempts and repetitions of the procedures we came up with the following models:

$$G1 = 10 - 1.3 * school(MS) + 0.2 * Medu + 0.5 * studytime - 1.1 * failures$$
$$-1.2 * schoolsup(YES) + 1.6 * higher(YES) - 0.3 * Dalc - 0.1 * absences + \epsilon$$

$$G2 = 10.6 - 1.2 * school(MS) + 0.3 * Medu + 0.4 * studytime - 1.2 * failures - 1.0 * schoolsup(YES)$$
$$+1.6 * higher(YES) - 0.3 * Dalc - 0.2 * health - 0.1 * absences + \epsilon$$

$$G3 = 11 - 1.5 * school(MS) + 0.2 * Fedu + 0.5 * studytime - 1.5 * failures$$
$$-1.3 * schoolsup(YES) + 1.8 * higher(YES) - 0.4 * Dalc - 0.2 * health + \epsilon$$

Some analysis for the three models.

G1

For G1 the constant is 10 which means that if all the other variables are zero then the grade for the first period will be 10. Moreover,if all the factors remain constant and someone chooses the MS school then there will be a reduction of 1.3 in the grade. Also, if all the covariates remain constant and the study time change by one it will have as result to increase the grade by 0.5. The coefficients of the other factors work in a similar way.

G2

In this model the constant is also about 10. It is important to notice that if somebody receives extra educational support (schoolsup) and the othres factors do not change there will be damage at his grade by 1. This fact looks strange as the reader waits such an action would help the student. Further, if mother's education level (Medu) increases by one ,and the other covariates remain constant, then the grade of the student for the second period will also increase by 0.3.

G3

The constant for the model which has as respnse variable the G3 is 11. Similar to the above, if all other variables are canceled out then the final score will be 11. In this model we have to notice the importance of wanting a child to study in higher education level. If the all covariates are same and somebody wants higher level studies then there is an increase at his final grade by 1.8. Finally, it is common known and approved by the models that health problems can affect negatively the effectiveness and as a result the grade of a student.

Unfortunately, we can not trust these models as much as the adjuster R-squared of the three models is about 0.3. Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. R-squared is the proportion of the variance in the response variable that can be explained by the predictor variables in a linear regression model. More simply, the higher the value of adjusted R-squared the better th response is described by the predictors through the model.

$$adjusted - R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

where:

- R2: The R2 of the model
- n: The number of observations
- k: The number of predictor variables

However, it is expected that there is no good description of the gradess from the models as they are highly correlated. For instance, if we can use **G1** and **G2** as predictors in a model with response the **G3** by a stepwise procedure, the adjusted R^2 equals to 0.85. Finally, from the graphs contained in appendix and below, it can be seen that it is not right to show much confidence in the models that were built. In contrast to the model contained in G1 and G2 there is greater uniformity of the residues. Both models contain outliers.

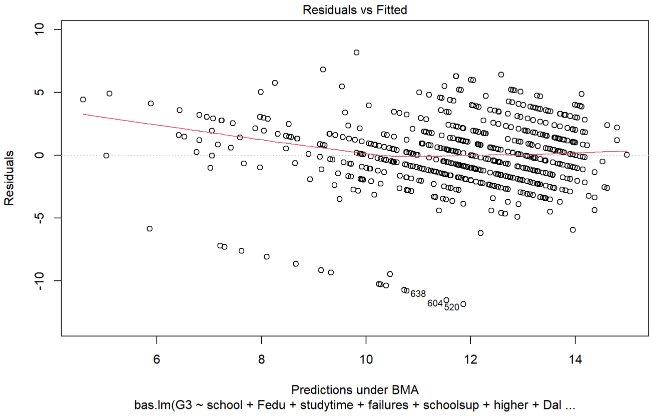

Figure 10: G3 without G1 and G2

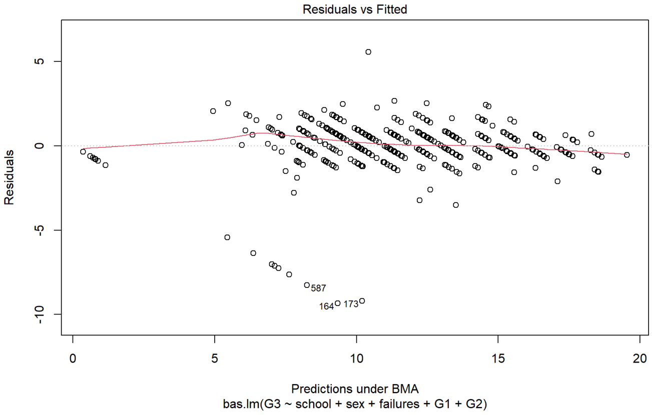

Figure 11: G3 with G1 and G2

7 Collaboration with Daskalaki Vasiliki

A14 – Student Performance Data Set
Vasiliki's dataset contains the same variables as mine but the grades **G1**, **G2** and **G3** are related with the "Mathematics" subject, unlike mine which concerns the "Portuguese language". Her dataset includes 395 observations. We had different approaches to model construction. I did not use any of the grades as predictors. On the other hand, Vasiliki used the grade she knew each time. For instance, when she had as response the **G1** she did not use the **G2** and **G3** as predictors, but when her response was the **G2**, she used **G1** as predictor because the grades of the first period were "known".
Vasiliki used generalized linear models with link function "logit". According to her analysis and mine there were different predictors for the same response. above is a table for each of the three scores.
For G1

Mathematics	Portuguese
Fedu	school
Mjob	Medu
failures	failures
schoolsup	schoolsup
goout	studytime
	higher
	Dalc
	absences

Table 5: Predictors for G1 for each subject

For G2

Mathematics	Portuguese
age	school
Medu	Medu
Mjob	failures
internet	schoolsup
goout	studytime
romantic	higher
freetime	health
G1	Dalc
absences	absences

Table 6: Predictors for G2 for each subject

For G3

Mathematics	Portuguese
age	school
Mjob	Fedu
traveltime	studytime
failures	failures
nursery	schoolsup
G1	higher
	health
	Dalc

Table 7: Predictors for G3 for each subject

Finaly, for the model about **G3**, Vasiliki did not use the **G2** as predictor because it completely separated the response and the model did not run. Although there are some common variables in each model. there are some differents which have great correlatio, like **Medu** and **Fedu** at "G1 model" or **goout** and **Dalc** at "G3 model". Last, Vasiliki's models look better for prediction as she used the previous grades to show the others.

Thank you!
Panagiotis Symianakis

References

Agresti, A. (2015). *Foundations of linear and generalized linear models.* Wiley.

Buhlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications.* Springer.

Cortez, P., & Silva, A. (2008). *Using data mining to predict secondary school student performance.* EUROSIS-ETI.

Croarkin, C., & Tobias, P. (2013). *Nist/sematech e-handbook of statistical methods.* Retrieved from http://www.itl.nist.gov/div898/handbook/

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r.* SAGE.

Portugal overview. (2020). European Commission and EACEA National Policies Platform and Eurydice. Retrieved 2021-12-20, from https://eacea.ec.europa.eu/national-policies/eurydice/content/portugal_en

[3] [6] [2] [4] [5] [1]