UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2020:28

A statistical analysis of the performance in mathematics of secondary students in Portugal

Camilla Molin

Examensarbete i matematik, 15 hp
Handledare: Silvelyn Zwanzig
Examinator: Martin Herschend
Juni 2020

Department of Mathematics
Uppsala University

Abstract

This thesis examines student performance in mathematics of secondary students in Portugal. The sample comprises of 395 observations and 33 variables and was collected during the 2005 – 2006 school year from two secondary schools in Portugal. A logistic regression models was used to predict whether or not a student received a pass or a fail grade in mathematics, in order to investigate if student background characteristics (and not grades) affected the performance in mathematics. Independence tests were used to pairwise examine association between the background variables. The final model was able to distinguish between pass and fail grades with a probability of 69% (just below the limit for an acceptable model). The explanatory variables of this model were: number of failures; whether or not a student had school support; whether or not a student was in a romantic relationship; and how much time the student spent with friends. There were some associations between the background variables like for example: between parent's job and education, and between number of failures and student alcohol consumption. Outliers, students with zero grades, also affected the performance of the model. A logistic regression model using only the previous term grade in mathematics as explanatory variable outperformed the final model with only background variables.

Contents

1 Introduction

In a study performed by Paulo Cortez and Alice Silva, data mining was used to predict student performance in mathematics and the Portuguese language in Portugal [2]. They concluded that it's possible to achieve a high predictive accuracy if previous grades are used in the model. Kotsiantis et. al. [18] make a similar conclusion that student achievement is highly affected by previous performances. But there are also other factors that affect student performance. Quoted from the PISA 2018 report for Portugal: "Socio-economic status was a strong predictor of performance in reading, mathematics and science in Portugal." [17].

Data for the study performed by Paulo Cortez and Alice Silva was collected during the 2005 – 2006 school year from two secondary schools in Portugal, Gabriel Pereira (GP) and Mousinho da Silveira (MS). Most of the observations came from Gabriel Pereira (349 compared to 46 from Mousinho da Silveira). Cortez and Silva used a questionnaire to collect data about characteristics and family background of students that might affect student performance. School reports were used for term grades and number of absences. The questionnaire was reviewed and first tested on a small set of students. [1]
.

1.1 Portuguese educational system

The first stage of the Portuguese compulsory school system is called *Basic education.* Basic education (students from 6 to 15 years of age) is divided into three cycles (1^{st} – 4^{th} grade, 5^{th} – 6^{th} grade, 7^{th} – 9^{th} grade). The second and final stage of the Portuguese compulsory school system is *Secondary education.* This cycle lasts for three years and corresponds to upper secondary school level.

The students in secondary school are 15 to 18 years old (but there are also some repeaters older than 18 years old). There are two paths in secondary school. One path is preparing students for higher education and one is preparing students for working life. The students are evaluated three times per school year. The third period grade is also the final grade. The grading scale is from 0 (lowest) to 20 (highest) i.e. 0, 1, 2, ..., 20. Pass grade is 10. [16]

1.2 Mathematics and PISA

How are Portuguese students' achievements in mathematics compared to other countries? To compare mathematics skills internationally one can for example use statistics from the PISA studies. PISA (Programme for International Student Assessment) is an international study of student performance in reading, mathematics and science. The last study is from 2018 and there's a new study every three years. A small fraction of all students of age 15 is randomly chosen to participate. The participants take tests but they also answer some survey questions about background characteristics and there attitudes towards the school and learning. Parents and principals also answer survey questions.

The test scores are standardized with a mean of 500 and a standard deviation of 100. The OECD average score in mathematics has slightly decreased from 2003 to 2018 but there has been an improvement in the Portuguese average scores during the same period of time (Figure 1). From performing lower than OECD average, the Portuguese students are now performing higher than OECD average.

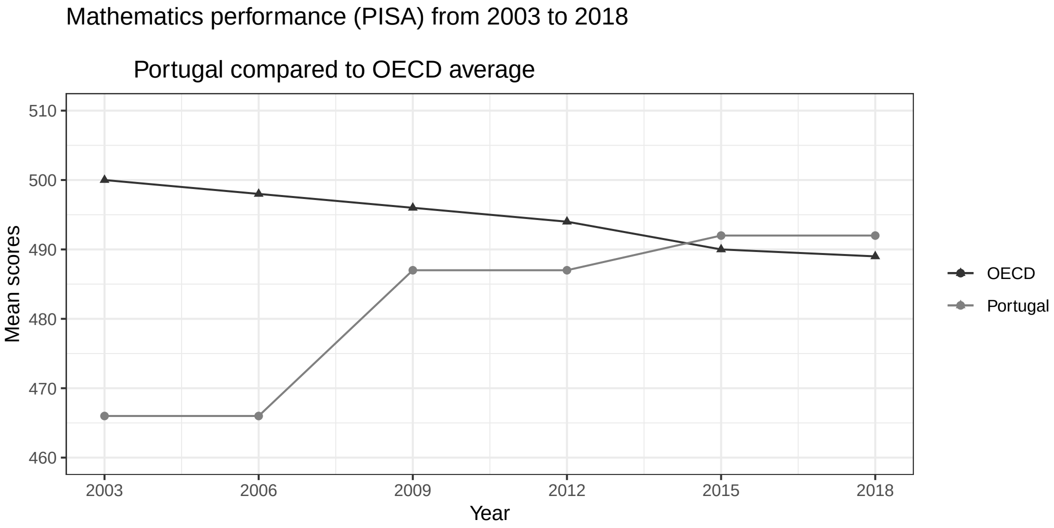


Figure 1: PISA 2003 to 2018.

The PISA study for Portugal based on the PISA scores in 2018 suggested that performance depends on many different factors [17].

- Socio-economic status explained 17% of the variation in mathematics performance for the Portuguese participants. This suggests that factors like parent's education and job, address, parental status, school, family support, going out with friends and the use of alcohol may be factors that affects student performances.
- Boys outperformed girls in mathematics which suggests that sex may be a factor affecting performance.
- On average the Portuguese students had a slightly higher absence average than the OECD average. Average may also be a factor affecting performance.

Not only previous performance but also family background seem to affect student performance. Pereira, M. conducted an analysis of Portuguese students' performance in the OECD programme for PISA using results from PISA 2003 to 2009. He concluded that "the variation in scores between PISA cycles has been substantially influenced by the changes in determinants, particularly with regard to the family background of children and, more importantly, the distribution of students by grades." [19].

1.3 Aim and research questions

This thesis analyses characteristics and family background of students and student performances in mathematics using data from the data set *Student Performance data set* [1].

The overall aim is to examine if other factors than previous grades affect secondary student's performance in mathematics in Portugal and if some of the background characteristics are associated.

The specific research questions are:

1. Are student background characteristics associated?
2. Are student background characteristics, other than previous grades, related to whether or not a student receives a pass grade in mathematics?

2 Data

The data set has 395 observations, 33 variables and no missing values. A list of all variables and their abbreviations can be found in Appendix. The data set consists mostly of categorical variables, both nominal and ordinal. Grades, age and absences are discrete quantitative variables.

2.1 Categorical variables

First a summary of the binary categorical variables.

- Sex (sex): 53%/47% (female/male)
- Address (address) : 22%/78% (rural/urban)
- School (school) : 88%/12% (Gabriel Pereira (GP)/Mousinho da Silveira (MS))
- Family size (famsize) : 29%/71% (3 or less/4 or more)
- Parental status (Pstatus) : 90%/10% (living together/living apart)
- Nursery school (nursery): 79%/21% (yes/no)
- Higher education (higher): 95%/5% (yes/no)
- School support (schoolsup): 13%/87% (yes/no)
- Family support (famsup): 61%/39% (yes/no)
- Extra paid classes within the course subject (paid) : 46%/54% (yes/no)
- Extra-curricular activities (activities) : 51%/49% (yes/no)
- Internet at home (internet) : 83%/17% (yes/no)
- Romantic relationship (romantic): 33%/67% (yes/no)

The categorical variables with more than two categories were plotted as bar charts. From the bar charts in Figure 2 we conclude that:

- The mother is usually the guardian and in the majority of families the family relations are good or excellent.
- The majority of the parents have an education level above 9^{th} grade. The proportion of higher education is larger among mothers. It's more common among mothers to work as a teacher, to stay at home or to work in the health category compared to the fathers. Other and services are the most popular working categories for both mothers and fathers.
- Approximately 50% of the students study between 2 and 5 hours per week. Approximately 25% of the student study less than 2 hours. The rest study more than 5 hours per week. Approximately 80% of the students don't have any failures.
- Most of the students have a low travel time to school. The most popular reason for choosing one of the schools is for the courses the school offers.
- A majority of the students consider their free time to be at a medium to high level. The mode for going out with friends is at a medium level.
- The weekend alcohol consumption is higher than the weekday alcohol consumption. During the weekdays a majority of the students have very low alcohol consumption but at the weekend the very low alcohol consumption is a minority.
- The majority of the students have medium to very good health.

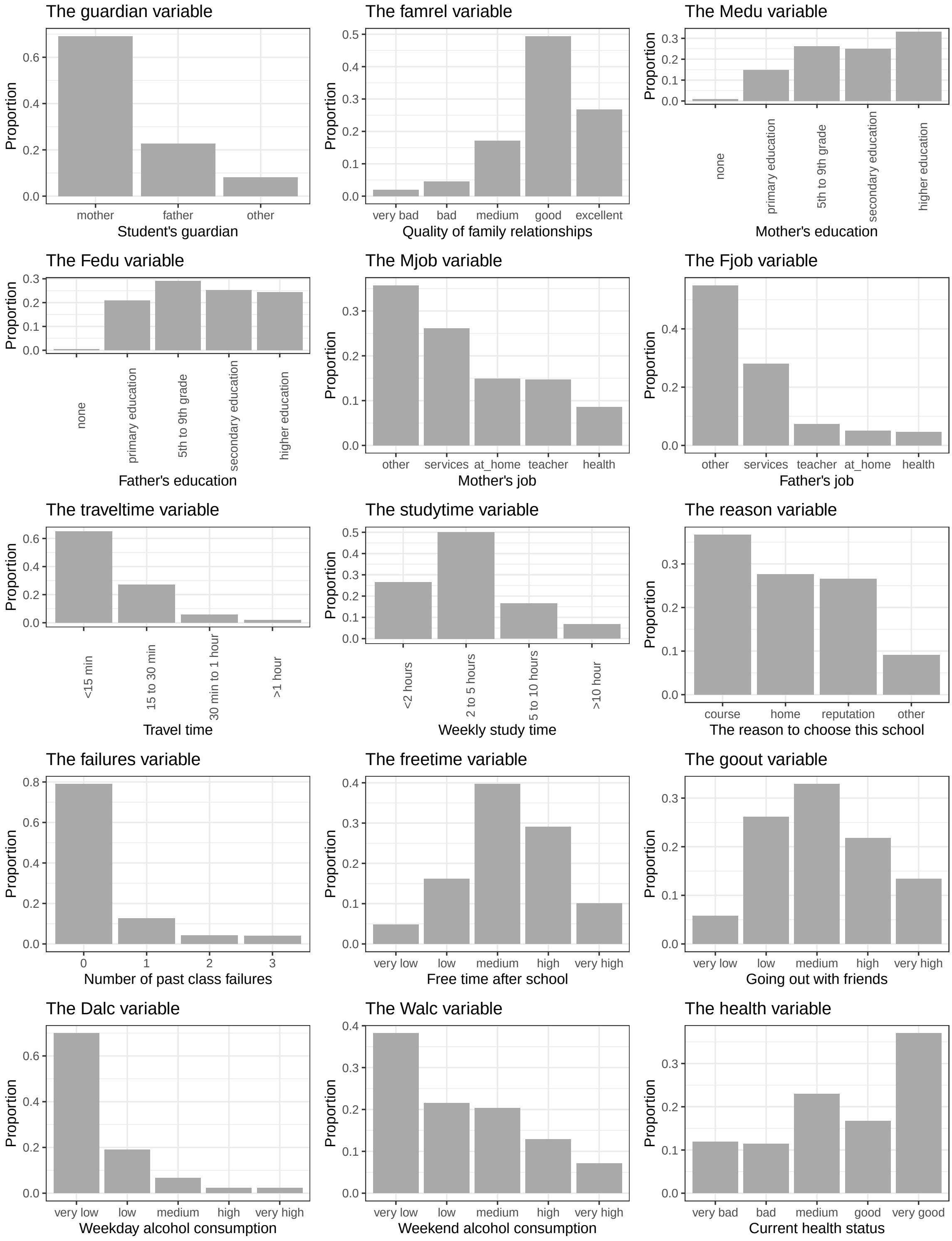


Figure 2: Categorical variables with more than two categories

2.2 Quantitative variables

The quantitative variables *absences* and *age* were plotted as histograms. The histograms are shown in the figure below (Figure 3).

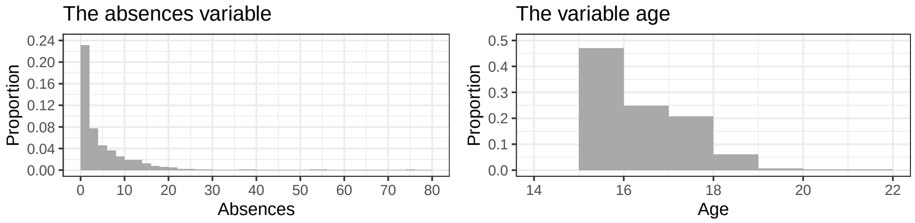


Figure 3: Student absences and age.

The numbers of absences are the number of lessons the student have missed. Most of the students have a low number of absences but there are also some outliers with a very high number of absences. The majority of the students were 15 to 18 years old (mainly secondary students but with some repeating students of 19 to 22 years of age).

The grade variables are plotted as boxplots in Figure 4. Grades for term 2 and 3 (final grades) include zero grades, while grades for term 1 do not. A possible explanation may be that the zeros represent drop-outs, students with high absences or students that didn't take an important exam . The median for the three term grades don't seem to differ much. The variation is larger for the final grade (term 3 grade).

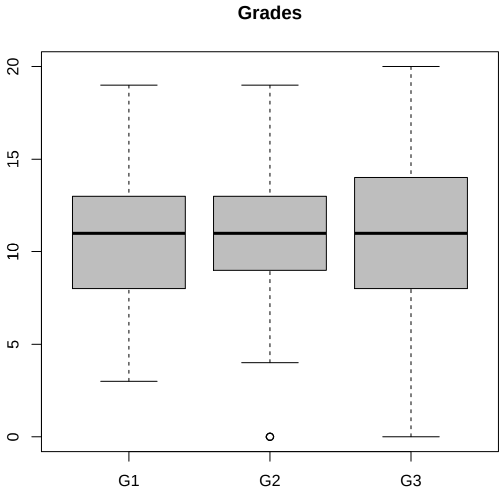


Figure 4: Boxplots of grades, G1 - first term grade, G2 - second term grade and G3 - third term grade / final grade.

2.3 Comparison between the schools

In this study data was collected from two schools, Gabriel Pereira (GP) and Mousinho da Silveira (MS). One need to bear in mind during the comparison that only 46 out of 395 students went to Mousinho da Silveira.

Address (address) and Mother's education (Medu) grouped by school are plotted in Figure 5. Failures (failures), school support (schoolsup) and study time (studytime) grouped by school are plotted in Figure 6. A summary of the analysis of the plots is shown in Table 1. It seems like Gabriel Pereira (GP) is a more academic school because the students study more and have less failures in this school. Mothers with higher education seems to choose this school for their children.

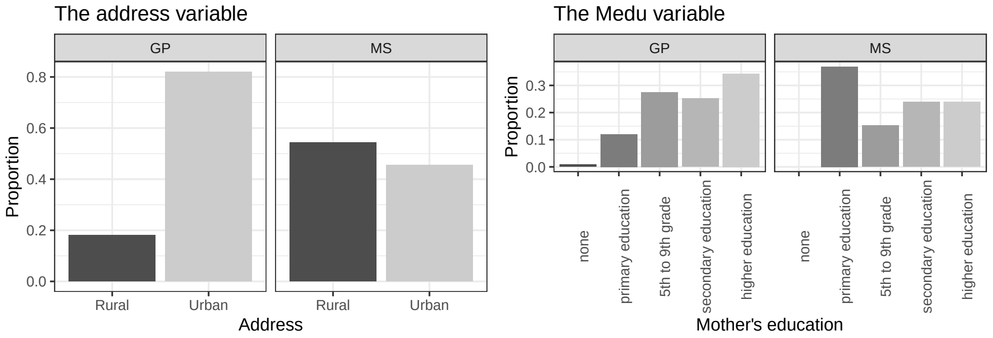


Figure 5: Comparison between the schools Gabriel Pereira (GP) and Mousinho da Silveira (MS) for the variables address and mother's education (Medu).

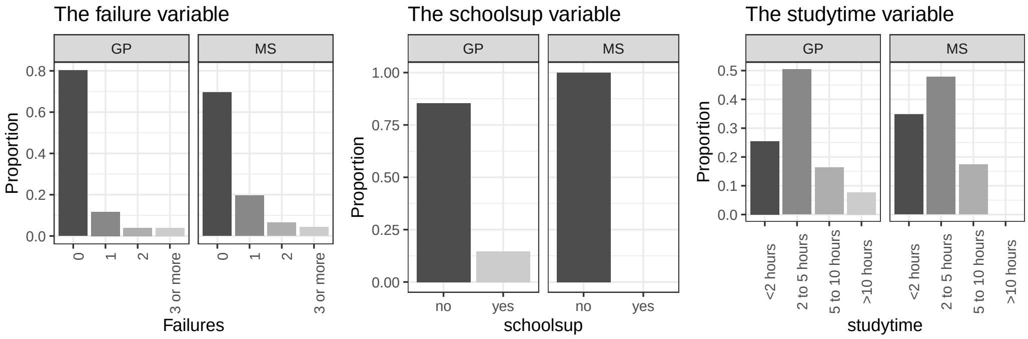


Figure 6: Comparison between the schools Gabriel Pereira (GP) and Mousinho da Silveira (MS) for the variables failures, school support (schoolsup) and study time.

Variable	GP	MS
Address (address)	mainly urban	urban $\approx$ rural
Mother's education (Medu), largest group has	higher education	primary education
Failures (failures)	lowest proportion	highest proportion
Offers school support (schoolsup)	yes	no
Study time (studytime)	highest	lowest

Table 1: Summary of the comparison between the two schools Gabriel Pereira (GP) and Mousinho da Silveira (MS).

3 Methods

This section describes the methods used in this thesis. R code used for the analyses can be found in the appendix.

3.1 Analysis of categorical data

The model used for pairwise analysis of categorical variables was a two-way table.

Two-way tables
Categorical variables describing student background characteristics were analyzed in pairs, A and B where

A had r categories: $A_1, ..., A_r$ and B had c categories: $B_1, ..., B_c$

The data consisted of n pairs of these variables:
$(A_1, B_1), (A_2, B_2), ..., (A_n, B_n)$

These pairs were summarized in two-way tables. The categories of A were the rows (with index $i = 1, 2, ..., r$) of the table and the categories of B (with index $j = 1, 2, ..., c$) were the columns (see Table 2 below). Each cell consisted of the number of observation with (A = i, B = j) denoted by n_{ij}.

	B			
A	1	...	c	**Total**
1	n_{11}	...	n_{1c}	$n_{1\bullet}$
...	...	n_{ij}	...	...
r	n_{r1}	...	n_{rc}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	...	$n_{\bullet c}$	$n_{\bullet\bullet} = n$

Table 2: Two-way table for A and B.

If one variable has more than two categories, say k distinct categories, and each category consists of n_i students with a constant probability for success (to fall into the category) π_i, then each $n_i \sim Bin(n, \pi_i)$ where $i = 1, 2, ..., k$. Then all categories for the variable is said to have a multinomial distribution with $\sum_{i=1}^{k} \pi_i = 1$ and $n_1 + n_2 + \cdots + n_k = n$. [8]. For multinomial distributed random variables with parameters n and $\pi_1, \pi_2, ..., \pi_k$, the joint probability function is given by

$$P(n_1, n_2, ..., n_k) = \frac{n!}{n_1!n_2!...n_k!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k}$$

The probability that a randomly chosen pair (A, B) from the population falls in the cell (i, j) of a two-way table is $\pi_{ij} = P(A = i, B = j)$. An item was a student and each student either fell into a category of not with probability for success, π_{ij}, i.e. each n_{ij} was binomial, $n_{ij} \sim Bin(n, \pi_{ij})$. The joint distribution of $\{n_{ij}\}$ is multinomial with parameters n and π_{ij}. [13].

The conditional probability of B for a given level of A in the population, $P(B = j|A = i) = \frac{P(A=i,B=j)}{P(A=i)}$, is denoted by $\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i\bullet}}$. A conditional probability distribution is a distribution that consists of $\{\pi_{j|i}\}$.

The analysis of categorical variables was performed in three steps. The aim was to find out if some pairs were associated (first specific research question).

1. **A two-way table or a stacked bar chart was made.**
   The purpose was to first get a visualization of the pair of variables to see if it looked like the variables were associated. For the variables plotted in stacked bar charts, the proportion of the levels of the second variable was plotted (vertical axis) for each level of the first variable (horizontal axis). If the proportions of the levels of the second variable i.e. the distribution of the second variable differed for the levels of the first variable, then there was an association between the variables (if the difference was statistically significant).

2. **Proper tests were performed and correlation coefficients were calculated.**
   The following tests were performed if possible:

   - Pearson's chi-squared test of independence (for all categorical variables).
   - The generalized Cochran-Mantel-Haenszel test of independence (if at least one ordinal, binary nominal needed if one nominal).

   The following correlations coefficients were calculated if possible:

   - Pearson C (for all categorical variables, no direction).
   - Kendall's tau-b (if both variables ordinal, direction).

3. **Assumptions were checked.**

3.1.1 Independence tests

Two variables are *dependent* if the value of one variable affects the value of the other variable. Two variables are *associated* if they are dependent. Since the variables were categorical and not normal distributed, non-parametric tests were used. The significance level for all independent tests was 5%.

Pearson's chi-squared test of independence
According to [3] two variables are said to be statistically independent if the true conditional distribution of B is identical at each level of A.

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i\bullet}} = \pi_{\bullet j} \implies \pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j} \ \forall i, j$$

The null hypothesis and the alternative hypothesis:

H_0 : A and B are independent i.e. $\pi_{ij} = \pi_{i\bullet} \cdot \pi_{\bullet j}$ for all pairs of (i, j)
H_1 : A and B are dependent i.e. $\pi_{ij} \neq \pi_{i\bullet} \cdot \pi_{\bullet j}$ for at least one pair of (i, j)

where $\pi_{i\bullet} = \sum_{j=1}^{c} P(A = i, B = j)$, $\pi_{\bullet j} = \sum_{i=1}^{r} P(A = i, B = j)$ and $\pi_{ij} = P(A = i, B = j)$

The test statistic for Pearson's chi-squared test is

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{[O_{ij} - E_{ij}]^2}{E_{ij}} \tag{3.1}$$

where
O_{ij} is the observed count in cell (i, j), in our case denoted by n_{ij}.
E_{ij} is the expected count in cell (i, j) under the null hypothesis.
$E_{ij} = \mu_{ij} = n\pi_{ij}$ for multinomial distributions.

The probabilities are estimated by [7]

$\hat{\pi}_{i\bullet} = \frac{n_{i\bullet}}{n}$ where $n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$, number of observations in row i

$\hat{\pi}_{\bullet j} = \frac{n_{\bullet j}}{n}$ where $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$, number of observations in column j

$\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ where $i = 1, ..., r$ and $j = 1, ..., c$

Under the null hypothesis, the expected counts E_{ij} can be estimated by

$$\hat{\mu}_{ij} = n \cdot \hat{\pi}_{i\bullet} \cdot \hat{\pi}_{\bullet j} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

The test statistic for the Pearson's chi-squared test for independence testing in two-way tables can then be rewritten as [3]

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{[n_{ij} - \hat{\mu}_{ij}]^2}{\hat{\mu}_{ij}} \tag{3.2}$$

When the total number of observations, n, is large, the test statistic X^2 (3.2) has approximately a χ^2-distribution with $(r - 1)(c - 1)$ degrees of freedom. If H_0 is true, the n_{ij}'s should not differ much from their expected values. A large value of X^2 gives evidence that H_0 should be rejected i.e. there's enough evidence to conclude that A and B are dependent (there's an association between variable A and B).

H_0 is rejected if $X^2 > \chi^2_{\alpha(r-1)(c-1)}$.

A p-value is the probability that χ^2 is at least as large as the observed value X^2 when H_0 is true i.e. $p - value = P(\chi^2_{\alpha(r-1)(c-1)} \geq X^2|H_0)$. If the p-value $< \alpha = 0.05$ (the significant level), H_0 is rejected.

The generalized Cochran-Mantel-Haenszel test of independence
If a chi-squared based test is used with ordinal data, the test will not take into account the ordering of the ordinal variables. According to [4] and [5] it's common to use a test based on correlation for ordinal data. The null and alternative hypothesis for independence test of ordinal variables using a correlation coefficient, ρ, is:

$H_0 : \rho = 0$
$H_1 : \rho \neq 0$

With the following test statistic:

$$M^2 = (n-1) \cdot r^2 \tag{3.3}$$

where r is the *sample* correlation coefficient and n the total number of observations.

This test is called the generalized Cochran-Mantel-Haenszel test. [13].

When the total number of observations n is large, M^2 has approximately a χ^2 - distribution with 1 degree of freedom. A large value of M^2 gives evidence that the H_0 should be rejected i.e. there's enough evidence to conclude that there is a correlation between A and B.

H_0 is rejected if $M^2 > \chi^2_{\alpha,1}$ or if the $p-value = P(\chi^2_{\alpha,1} \geq M^2|H_0) < \alpha = 0.05$.

To calculate r, scores having the same ordering as the levels, are assigned to rows and columns. The scores for the rows are denoted by u_i and $u_1 \leq u_2 \leq ... \leq u_R$ where R is the number of rows (number of categories for variable A). The scores for the columns are denoted by v_j and $v_1 \leq v_2 \leq ... \leq v_C$ where C is the number of columns (number of categories for variable B). The scores are usually integer scores i.e. $u_1 = 1$, $u_2 = 2$, ... and $v_1 = 1$, $v_2 = 2$, ... The *sample* correlation coefficient is calculated by

$$r = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}(u_i - \bar{u})(v_i - \bar{v})n_{_{ij}}}{\sqrt{[\sum_{i=1}^{R}\sum_{j=1}^{C}(u_i-\bar{u})^2 n_{ij}][\sum_{i=1}^{R}\sum_{j=1}^{C}(v_i-\bar{v})^2 n_{ij}]}} \tag{3.4}$$

where
$\bar{u} = \sum_{i=1}^{R}\sum_{j=1}^{C} u_i \frac{n_{ij}}{n}$ is the row mean and
$\bar{v} = \sum_{i=1}^{R}\sum_{j=1}^{C} v_j \frac{v_{ij}}{n}$ is the column mean

The generalized Cochran-Mantel-Haenszel test can also be used if one of the variables is a binary nominal variable. If the nominal variable is used as the row variable, then small p-values give evidence that there are a difference in row means i.e. there's a correlation between A and B.

Assumptions of independence tests based on the chi-squared distribution

- Independent random sampling.
- The categories of the variables must be mutually exclusive.
- The counts in each category should not be too small. A rule of thumb is that the estimated expected counts should exceed five.

3.1.2 Correlation

The strength of association can be measured by a correlation coefficient. If variables are independent, the population correlation coefficient is zero.

Nominal data
According to [14] Pearson's contingency coefficient (Pearson C) can be used as a correlation coefficient for categorical variables. This measurement uses χ^2 adjusted to the sample size n. The range of the coefficients is from 0 to 1 (no direction). A rule of thumb is that if the correlation coefficient is at least 0.1 there's an association between the two variables (weak), the correlation is moderate between 0.3 and 0.5 and the correlation is high above 0.5.

Pearson C:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} \tag{3.5}$$

where
χ^2 is the Pearson's chi-squared statistics.
n is the total number of observations i.e. the sample size.

If both variables are ordinal it's better to use some other kind of correlation coefficient since Pearson's contingency coefficient don't give a direction and don't take into consideration the ordering of variables.

Ordinal data
For ordinal data, correlation can be measured by methods using the ordering of the data. According to [13] the Kendall's tau-b correlation coefficient is a non parametric measure of association that can be applied to ordinal data. Kendall's tau-b is a modification of Kendall's tau when there are rank ties. Since the ordinal variables in this data set have few categories, there will be many ties and Kendall's tau-b will be used. The range of the coefficients is from -1 (negative correlation) to 1 (positive correlation). The observations are paired and the coefficient is based on the number of concordant and discordant pairs.

Two paired observations (A_i,B_i) and (A_j,B_j) are concordant if
$A_i < A_j$ and $B_i < B_j$ or $A_i > A_j$ and $B_i > B_j$

Two paired observations are discordant if
$A_i < A_j$ and $B_i > B_j$ or $A_i > A_j$ and $B_i < B_j$

Two paired observations are tied if $A_i = A_j$ and/or $B_i = B_j$.

The total number of pairs N for a sample size of n is:

$N{=}\binom{n}{2}{=}P + Q + A_0 + B_0 + (AB)_0$
where

P = number of concordant pairs
Q = number of discordant pairs
A_0 = number of pairs tied only on the A variable
B_0 = number of pairs tied only on the B variable
$(AB)_0$ = number of pairs tied on both A and B

Kendall tau-b:

$$t_b = \frac{P - Q}{\sqrt{(P + Q + A_0)(P + Q + B_0)}} \tag{3.6}$$

t_b is the sample estimate of the population Kendall tau-b, τ_b , where
$\tau_b = P(concordance) - P(discordance) = \dfrac{P - Q}{N}$.

3.2 Multiple logistic regression

The investigate if student background characteristics (and not grades) affect whether or not a student receives a pass grade in mathematics, a model with binary response was used. The binary response variable was chosen to take one of two values, 0 for a fail grade in mathematics and 1 for a pass grade. The grade for passing a course in Portuguese schools is 10. The final mark (G3) was assigned to two groups: pass ($grades \geq 10$) or fail ($grades < 10$). All the outliers (grade 0) were then assigned to the group with fail grades.

A possible function to use for modelling binary responses is the logistic mean response function. The logistic mean response function is a sigmoidal (S-shaped) response function where the probabilities 0 and 1 are reached asymptotically.

The multiple logistic regression model is:

$$Y_i = E(Y_i) + \epsilon_i$$

Y_i is the value of the binary response variable in the ith trial and $Y_i = \{0, 1\}$.
ϵ_i is the random error term in the ith trial.
$E(Y_i)$ is the logit mean response function.

Y_i is a Bernoulli random variable with the probability distribution shown in Table 3.

Y_i	**Probability**
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Table 3: Bernoulli probability distribution.

Since $Y_i \sim Be(\pi_i)$, the expected value of Y_i is:

$$E(Y_i) = 1 \cdot \pi_i + 0 \cdot \ (1 - \pi_i) = \pi_i \tag{3.7}$$

Note that the expected value of Y_i is a probability.

The logit mean response function is:

$$E(Y_i) = \pi_i = \frac{1}{1 + exp(-\boldsymbol{X_i^T \beta})} \tag{3.8}$$

where

$\boldsymbol{X_i^T \beta} = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i,p-1}$

$X_{i1}, X_{i2}, ..., X_{i,p-1}$ are the values of the explanatory variables in the ith trial.

$\beta_0, \beta_1, ..., \beta_{p-1}$ are parameters.

The logit response function:

$$logit(\pi) = \boldsymbol{X^T \beta} \tag{3.9}$$

where

$$logit(\pi) = log_e(\frac{\pi}{1 - \pi}) \tag{3.10}$$

Note that the logistic regression model has a linear form for the *logit* of the probability π. The right-hand side of (3.9) is called the *linear predictor*. The ratio $\frac{\pi}{1-\pi}$ is the odds ratio (OR). The right-hand side of (3.10) if often referred to as the *log odds* or shorter the *logit* of π_i.

Assumptions of binary logistic regression

- Independent random sampling.
- Binary response variable.
- Little or no multicollinearity among the explanatory variables.
- Linearity of explanatory variables and the logit response function.
- Large sample.

Since the distribution of the grades don't seem to differ much (Figure 4), one can use only either term 1 grades or term 2 grades to predict if a student will pass or fail. Since the aim of the logistic regression model building was to see if student background characteristics (and not grades) affect whether or not a student receives a pass grade in mathematics, the grades for term 1 and 2 were not used in the logistic regression models.

3.2.1 Model selection

The number of possible logistic regression models is large if there is a large number of explanatory variables. In a selection search procedure, one either start with a null model and then add variables one at a time (forward selection) or start with a full model and then eliminate variables one at a time (backward elimination). The aim with both methods is to only keep a smaller set of variables that account for as much of the total variance as possible. Since there was a large number of possible explanatory variables in this data set, the time consuming forward selection procedure was rejected in favour of the backward elimination procedure. The method of backward elimination will be further explained below. [9].

Backward elimination

If variables are eliminated one at a time the Wald statistic, $z^* = \frac{b_k}{s(b_k)}$, can be used for testing whether or not $\beta_k = 0$ (the null hypothesis) for variable X_k. b_k is the maximum likelihood estimates of β_k and $s(b_k)$ is the standard error of the estimate. The p-value, $p = 2P(z > |z^*|)$. z is a standard normal random variable. For $p < \alpha$, H_0 is rejected and the variable is considered statistically significant.[9]

The backward elimination search procedure starts with a model containing all possible explanatory variables i.e. a full model. The explanatory variable with the largest p-value is first dropped from the model. A new model with the remaining variables is fitted and the one with the largest p-value is dropped. The process will proceed in this manner until all variables have a p-value below the chosen limit. According to [10] choosing a p-value of 0.05 often exclude important variables from the model. It's recommended to choose a p-value in the range from 0.15 to 0.20. A p-value of 0.15 was chosen as the limit for the first models and a p-value of 0.05 for the last models (see section 3.2.2 below).

Training and test set
The data set was randomly split into a training and a test set. The training set was used for fitting the model. The test set was used for performance evaluation of the fitted model. It's common to use 70/30 or 80/20 splits where the highest number is the proportion of the training set (in percent). Since I found it important to get at reliable model, a large training set was chosen (the 80/20 split). The training set (80%) was used for fitting the model. The testing set (20%) was used for performance evaluation of the fitted model.

Transformation of variables
A dummy variable, X_i, is a nominal variable having only two categories. If a dummy variable is used in a logistic regression model, the parameter β_i indicates how much higher (or lower) the fitted logit response function is for "success" ($X_i = 1$) compared to the baseline ($X_i = 0$) "failure". If a nominal variable has more than two categories, additional dummy variables are needed in the logistic regression model. One needs $k - 1$ dummy variables for a variable with k categories. To reduced the number of possible explanatory variables, the nominal variables having more than two categories were transformed to dummy variables.

Four of the nominal variables had more than two categories. These variables were transformed to dummy variables. The new groups of the nominal variables with more than two categories were:

reasonD: reputation (1) / other (0)
guardianD: mother (1) / other (0)
FjobD: teacher (1) / other (0)
MjobD: teacher (1) / other (0)

Note that all variables transformed to dummy variables were denoted by an extra D in the end, for example *reason* was denoted by *reasonD* when used as a dummy variable.

The variable age was also modified. Secondary school students in Portugal are 15 to 18 years old. There are 29 observations of students from 19 to 22 years of age. These students were considered as repeaters and was put in the same category. Age was transformed to an ordinal variable with five categories as follows:

1 - 15 years old, 2 - 16 years old, 3 - 17 years old, 4 - 18 years old, 5 - 19 years old or older.

The transformed *age* variable was denoted by *ageT*.

The same association tests as above were performed for the new transformed variables. The same correlation coefficient as above were calculated.

Akaike information criterion
The Akaike information criterion, *AIC*, was used as a selection criteria to select the best model.

$$AIC_p = nln(SSE_p) - nln(n) + 2p \tag{3.11}$$

where
SSE_p is the error sum of squares.
p is the number of parameters in model.
n is the sample size.

The model with the smallest AIC_p was considered the best model. Note that, since $nln(n)$ is fixed, a model with small SSE_p will do well as long as $2p$ is small. AIC penalize a model with many explanatory variables (large p).

Multicollinearity
A model can have problems with multicollinearity which occurs when the explanatory variables are correlated. Signs of multicollinearity is for example

- Estimated parameters change drastically when explanatory variables are added or removed from the model.
- The standard error for the estimated parameter is large.
- A parameter that is expected to be an important predictor is not statistically significant.
- Estimated parameters get the "wrong" sign.

Each variable was added as a single variable in a simple logistic regression model with pass as the response variable The purpose was to check the sign of the estimated parameter. The sign of the estimated parameters were then be compared to the sign in the models to see if there was a change of sign.

3.2.2 Model building

The model building started either from the entire data set (excluding only the previous grade variables G1 and G2) or from a reduced data set. For preparing the reduced data set, the pair of variables where tests show association and all correlation coefficients exceed 0.20 were further examined. The variables in the pair were used in a simple logistic regression model with pass as the response and the variables (one at a time) as the explanatory variable. The variable with highest p-value was dropped from the full data set to create the reduced data set. The reason for using a reduced data set was to avoid multicollinearity problems.

- The first model was a model starting from the **entire data set**. Backward elimination was used where variables with the highest p-value was dropped one at a time until all individual variables had a **p-value below 0.15**.
- The next model was a model starting from the **reduced data set**. Backward elimination was used where variables with the highest p-value was dropped one at a time until all variables had a **p-value below 0.15**.
- The next model was a model starting from the **entire data set**. Backward elimination was used where variables with the highest p-value was dropped until all variables had a **p-value below 0.05**.
- The next model was a model starting from the **reduced data set**. Backward elimination was used where variables with the highest p-value was dropped one at a time until all variables had a **p-value below 0.05**.
- The last model was a model with **only G2** as explanatory variable. This model was only used for comparison.

3.2.3 Model evaluation

Hosmer-Lemeshow Goodness of fit test
For unreplicated data, the Hosmer-Lemeshow goodness of fit test was used to check the overall fit of a model i.e. how well the observed values in the sample correspond to the expected values under the model. In the Hosmer-Lemeshow goodness of fit test, data was first grouped into classes with similar fitted values $\hat{\pi}_i$ or similar fitted logit values $logit(\hat{\pi}_i)$. After grouping the data, the Pearson chi-squared test statistic was used. [9]

Pearson chi-squared test statistic:

$$X^2 = \sum_{j=1}^{c}\sum_{k=0}^{1}\frac{[O_{jk} - E_{jk}]^2}{E_{jk}} \tag{3.12}$$

The null and alternative hypothesis are:

H_0: $E(Y) = \frac{1}{1+exp(-\boldsymbol{X^T\beta})}$

H_1: $E(Y) \neq \frac{1}{1+exp(-\boldsymbol{X^T\beta})}$

When n is large, X^2 is approximately χ^2- distributed with $c-2$ degrees of freedom.

For an adequate fit, we don't want to reject H_0. If $X^2 \leq \chi^2_{\alpha(c-2)}$ or if the *p-value* $= P(\chi^2_{\alpha(c-2)} \geq X^2|H_0)$ is high enough, we fail to reject H_0. Then there is not enough evidence to reject that the overall fit of the model is appropriate.

Logistic regression residuals
Since the response variable is binary for logistic regression, the residuals are also binary. The ordinary residuals ($e_i = Y_i - \hat{\pi}_i$) will not be normally distributed so plots of ordinary residual against for example fitted values used for linear regression models are uninformative. There are special kind of residuals used for logistic regression like *Pearson residuals* and *Studentized Pearson residuals.* Pearson and Studentized Pearson residuals are plotted against estimated probabilities, $\hat{\pi}_i$, in residual plots. If the logistic regression model is correct, then $E(Y_i) = \pi_i$ and $E(Y_i - \pi_i) = 0$. When fitting a smooth curve to the residuals (the method of lowess smooth) the result should be a horizontal line with zero intercept for a good model.[9].

Pearson residuals:

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{3.13}$$

Pearson residuals are ordinary residuals divided by the estimated standard error of Y_i. Since $\sum_{i=1}^{n} r^2_{P_i} = X^2$, the square of each r_{P_i} measure the contribution of each response to the Pearson's chi-squared test statistic.

Studentized Pearson residuals:

$$r_{SP_i} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}} \tag{3.14}$$

To get residuals with unit variance, the ordinary residuals are divided by their estimated standard deviation approximated by $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$. h_{ii} is the ith diagonal element of the estimated hat matrix, H, for logistic regression:

$$H = \hat{W}^{\frac{1}{2}}X(X^T\hat{W}X)^{-1}X^T\hat{W}^{\frac{1}{2}} \tag{3.15}$$

where
$\hat{W}$ is the diagonal matrix with elements $\hat{\pi}_i(1-\hat{\pi}_i)$.
X a matrix containing the explanatory variables.

Cook's distance
Some observations may have to much influence on the fitted linear predictor. The fit could be different if these influential observations are deleted. Cook's distance can be used to identify these kind of influential observations. Cook's distance measures the standardized change in the linear predictor $\hat{\pi}_i$ when the ith case is deleted. [9]

Cook's distance:

$$D_i = \frac{r^2_{P_i}h_{ii}}{p(1-h_{ii})^2} \tag{3.16}$$

where p is the number of parameters in the logistic model.

To detect influential observations, Cook's distances were calculated and plotted for each observation. Observations were the distance exceeded 0.06 were considered as influential.

3.2.4 Further model evaluation

The models were also evaluated regarding to predictive power. The predictive power of a logistic regression model tells us how well the model classifies. The outcome of the logistic regression models was a probability whether or not a student would receive a pass grade in mathematics. A cutoff, p_{cut}, for getting a pass grade was chosen for the best model (using training data) and the students were classified as a student that passed (1) or failed (0). [6] [10]. The prediction rule was:

$$\hat{Y} = \begin{cases} 1 & if\, \hat{\pi} > p_{cut} \\ 0 & if\, \hat{\pi} \leq p_{cut} \end{cases}$$

The best model was then evaluated using test data. The predicted passing rate was calculated from a confusion matrix and compared with true passing rates. The actual and predicted classifications were summarized in a confusion matrix (TP = true positive, FP = false positive, FN = false negative,TN = true negative). See Table 4.

		Actual class, Y	
		0	1
Predicted class, $\hat{Y}$	0	TN	FN
	1	FP	TP

Table 4: Confusion matrix.

The *Overall Accuracy (ACC)* tells us the proportion of correct classifications:

$$ACC = \frac{P(\hat{Y} = 0|Y = 0) + P(\hat{Y} = 1|Y = 1)}{P(Y = 0) + P(Y = 1)} = \frac{TN + TP}{TP + FP + FN + TN}$$

In our case *Overall accuracy* (ACC) is the proportion of passing and fail grades that are correctly classified. $1 - ACC$ is the *prediction error rate* (the proportion of incorrect classifications). An ACC of 50% means that the predictions were no better than chance. A rule of thumb is an ACC of 95-100% is very good, 85-95% is good, 70-85% is satisfactory, 50-70% needs to be improved.

To display two types of errors for all possible cutoffs, it's common to use a *ROC curve* (receiver operating characteristic curve). The *ROC curve* plots the *sensitivity* as a function of (1 - *specificity*) where:

Sensitivity (true positive rate, TPR): $TPR = \frac{TP}{TP+FN}$
Specificity (true negative rate, TNR): $TNR = \frac{TN}{TN+FP}$

In our case *Sensitivity* is the proportion of pass grades that are correctly classified and *Specificity* is the proportion of fail grades that are correctly classified. A perfect classification (a model with an overall accuracy of 100%) will have a *sensitivity* of 100% i.e. no false negatives (FN) and *specificity* of 100% i.e. no false positives (FP). This will give a point in the upper left corner of the *ROC curve* (0,1). The area under the *ROC curve*, AUC, is a measure how close the *ROC curve* is to a perfect classification. For a perfect classification, the area will be 1 ($AUC = 1$) and for a model with predictions not better than random guessing the area will be 0.5 ($AUC = 0.5$). A rule of thumb is that an AUC of 0.9-1 is outstanding, 0.8-0.9 is excellent, 0.7-0.8 is acceptable, 0.6-0.7 is poor and 0.5-0.6 no discrimination.

ACC, *Sensitivity* and *Specificity* will differ for different cutoffs. The models were fitted on the same training data but with different explanatory variables. Since these models gave different optimal cutoffs, the AUC (the area under the ROC-curve) was used to choose the best model. AUC is not dependent on a cutoff and the model with the highest AUC was chosen as the best model. For the models where the plots of Cook's distances showed influential observations, adjusted models where the influential observations were removed, were also evaluated. The criteria for a good model was a model with high AUC and low *AIC*. The best model was also compared with the model with only term 2 grade (G2) as explanatory variable.

3.2.5 Interpretation of parameters

Since the logistic regression model is non-linear, the interpretation of estimated parameters is not as straightforward as for linear regression models. [10]. Using (3.9) and (3.10):

For X: $logit(\hat{\pi}_1) = b_0 + b_1 X$
For (X + 1): $logit(\hat{\pi}_2) = b_0 + b_1(X + 1)$

where b_0 and b_1 are the maximum likelihood estimates of β_0 and β_1.

$logit(\hat{\pi}_2) - logit(\hat{\pi}_1) = b_0 + b_1(X+1) - (b_0 + b_1X) = b_1$
$log_e(odds_2) - log_e(odds_1) = b_1$
$log_e(\frac{odds_2}{odds_1}) = b_1$
$\frac{odds_2}{odds_1} = exp(b_1)$

The estimated ratio of odds, the *odds ratio*, is denoted by $\widehat{OR}$:

$$\widehat{OR} = \frac{odds_2}{odds_1} = exp(b_1) \tag{3.17}$$

95% confidence interval for odds ratio
Confidence limits for β_k:

$$b_k \pm z(1-\alpha/2)s(b_k) \tag{3.18}$$

where b_k is the maximum likelihood estimates of β_k, $s(b_k)$ is the standard error of the estimate and $z(1-\alpha/2)$ is the $(1-\alpha/2)100$ percentile of the standard normal distribution. For a 95 % confidence level, $\alpha = 0.05$.

Confidence limits for the odds ratio $exp(\beta_k)$:

$$exp(b_k \pm z(1-\alpha/2)s(b_k)) \tag{3.19}$$

For estimated parameters with positive sign ($b_k > 0$) and if the variables is

- **an ordinal variable:**
  An increase in level of the variable (holding all other variables constant) increases the odds of passing.
- **a dummy variable:** The estimated value of the parameter is the exponent of e and $e^{b_k} > 1 \, \forall \, b_k > 0$. For positive parameters, the factor $e^{b_k} > 1$ and the estimated odds for passing is then lower for the baseline.

The situation is the opposite for estimated parameters with negative sign ($b_k < 0$).

4 Results

This section describes the findings from the analysis in R using the methods described in the Methods section.

4.1 Analysis of categorical data

4.1.1 Association and correlation for mother's education (Medu) and mother's job (Mjob)

Association between mother's education (ordinal variable) and mother's job (nominal variable) was first studied since it is reasonable to think that there is an association between these two variables. Two two-way tables, one with observed frequencies (Table 5) and one with observed row proportions (Table 6), were created.

	Mjob					
Medu	at home	health	other	services	teacher	**Total**
0	2 (0)	0 (0)	1 (1)	0 (1)	0 (0)	3
1	25 (9)	1 (5)	28 (21)	5 (15)	0 (9)	59
2	22 (15)	2 (9)	47 (37)	32 (27)	0 (15)	103
3	8 (15)	5 (9)	43 (35)	40 (26)	3 (15)	99
4	2 (20)	26 (11)	22 (47)	26 (34)	55 (19)	131
Total	59	34	141	103	58	395

Table 5: Two-way table showing observed counts and expected counts (in brackets) for Mother's education (Medu, rows) and Mother's job (Mjob, columns). 0 – none, 1 – primary education, 2 – 5th to 9th grade. 3 – secondary education, 4 – higher education.

In Table 5 there are five cells with expected counts below 5 (from category: Mother's education 'none'). This will be further discussed in section (5.1) below.

	Mjob				
Medu	at home	health	other	services	teacher
0	0.67	0.00	0.33	0.00	0.00
1	0.42	0.02	0.47	0.08	0.00
2	0.21	0.02	0.46	0.31	0.00
3	0.08	0.05	0.43	0.40	0.03
4	0.01	0.20	0.17	0.20	0.42

Table 6: Two-way table showing observed row proportions for Mother's education (rows) and Mother's job (columns).

In Table 6, the row proportions differed which indicates a possible association between mother's job and mother's education.

The stacked bar chart in Figure 7 shows the row proportions of mother's job (Mjob) as stacked bars for each level of Mother's education (Medu). Since I found it easier to see if the row proportions differed in a stacked bar chart, stacked bar charts were used instead of two-way tables (for visualization).

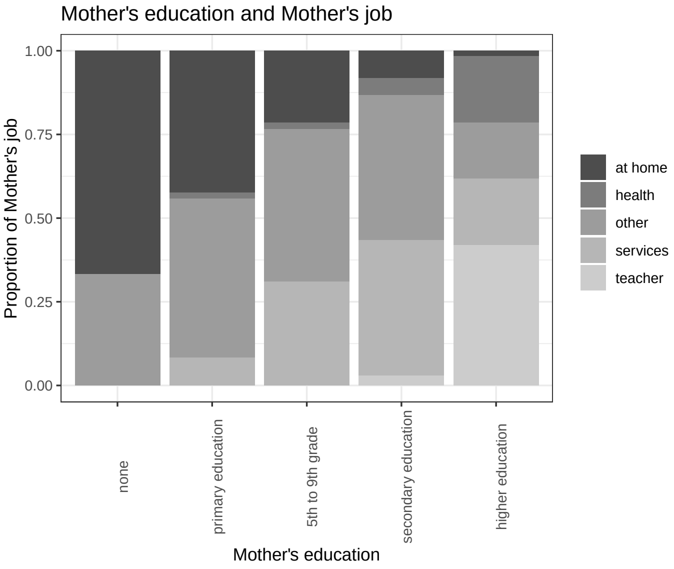


Figure 7: The proportion of the levels of mother's job are plotted for each level of mother's education.

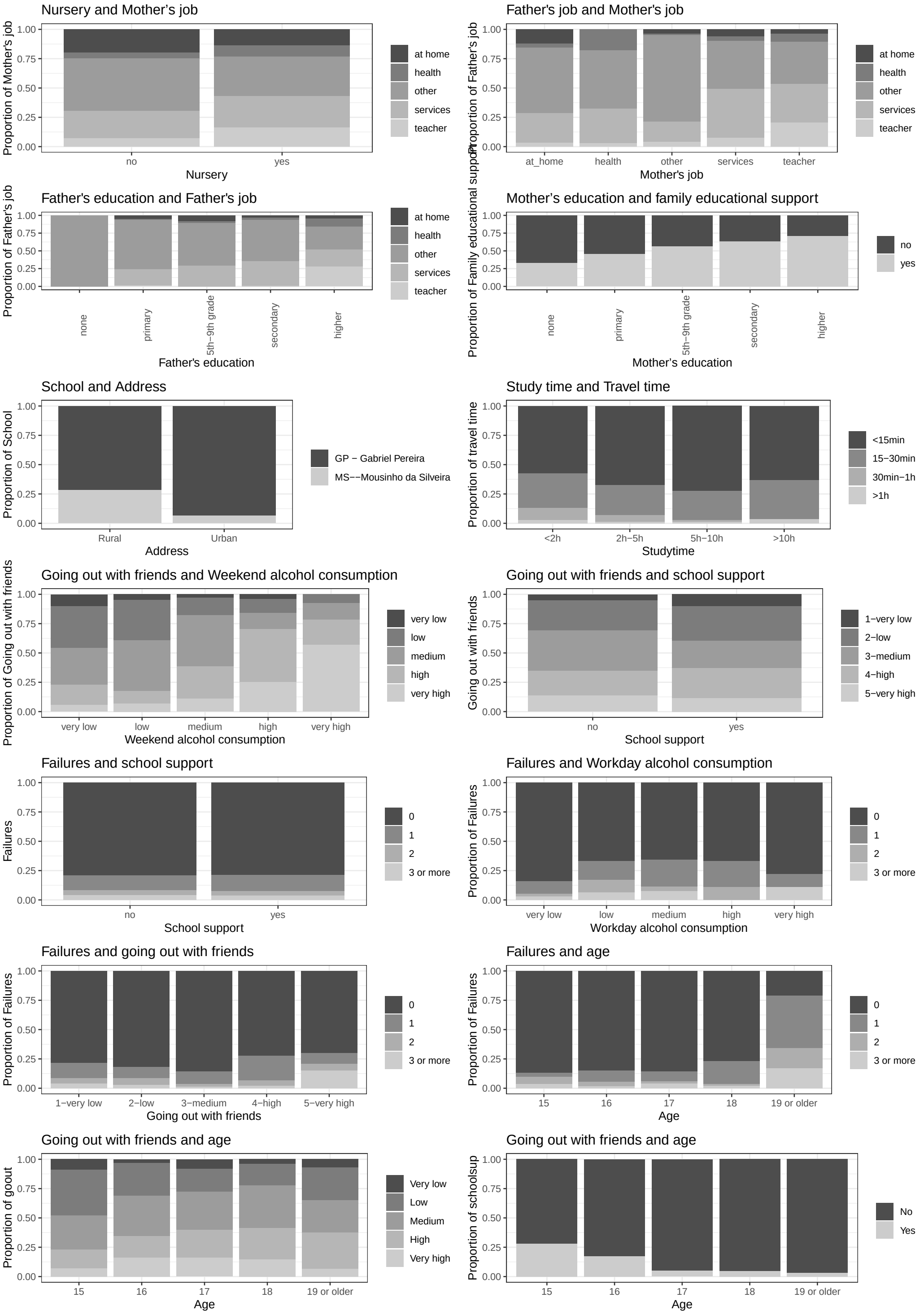
Since one of the variables was nominal (mothers education, Medu), Pearson chi-squared test of independence was performed and Pearson C was calculated. The results are shown in the table below (Table 7). The independence test gave significant results ($p < 0.001$) that gave enough evidence that there was an association between mother's education and mother's job. The correlation coefficient showed a high correlation.

Variable	X^2 (p-value)	Pearson C
Mother's edu. (O)	224.7 (<0.001)	0.602

Table 7: Association between mother's job (Mjob) and mother's education (Medu). X^2 is the statistics for Pearson's chi-squared test of independence.

4.1.2 More associations and correlations between categorical variables

14 stacked bar charts are shown in Figure 8. The proportion of the variable on the vertical axis didn't seem to differ much for each level of the variable on the horizontal axis for the following pairs of variables: Mother's job (Mjob) and nursery (nursery), going out with friends (goout) and school support (schoolsup) and failures (failures) and school support (schoolsup).


Nursery and Mother's job

Father's job and Mother's job

Father's education and Father's job

Mother's education and family educational support

School and Address

Study time and Travel time

Going out with friends and Weekend alcohol consumption

Going out with friends and school support

Failures and school support

Failures and Workday alcohol consumption

Failures and going out with friends

Failures and age

Going out with friends and age

Going out with friends and age

Figure 8: The proportion of the levels of the row variable are plotted for each level of the column variable.

The test results for the variables plotted in Figure 8 are shown in the table below (Table 8). For some ordinal variables (studytime/traveltime and age/goout), the test that didn't take into account the ordering of these variables (X^2) gave non-significant results while the tests taking into account the ordering (M^2) gave significant results.

	Variables	X^2	M^2
		(p-value)	(p-value)
1	Mjob (N)/ nursery (N)	9.3 (0.054)	- -
2	Mjob (N)/ Fjob (N)	73.4 (<0.001)	- -
3	Fedu (O)/ Fjob (N)	108.4 (<0.001)	- -
4	Medu (O)/ famsup (N)	13.5 (0.009)	13.3 (<0.001)
5	school (N)/ address (N)	30.9 (<0.001)	- -
6	studytime (O)/ traveltime (O)	11.2 (0.264)	4.0 (0.045)
7	goout (O)/ Walc (O)	116.6 (<0.001)	69.6 (<0.001)
8	goout (O)/ schoolsup (N)	3.90 (0.420)	0.560 (0.454)
9	failures (O)/ schoolsup (N)	<0.1 (0.994)	<0.1 (0.993)
10	failures (O)/ Dalc (O)	22.6 (0.031)	7.3 (0.007)
11	failures (O)/ goout (O)	77.5 (0.004)	17.2 (0.013)
12	age (O)/ failures (O)	29.0 (<0.001)	6.1 (<0.001)
13	age (O)/ goout (O)	23.0 (0.115)	5.8 (0.016)
14	age (O)/ schoolsup (N)	30.8 (<0.001)	25.9 (<0.001)

Table 8: Association tests. X^2 is the statistics for Pearson's chi-squared test of independence. M^2 is the statistics for the generalized Cochran-Mantel-Haenszel of independence.

The correlation coefficients are shown in Table 9. No correlations were high (above 0.5). The correlation was moderate (between 0.3 and 0.5) for these pairs of variable: mother's job (Mjob)/ father's job (Fjob), father's education (Fedu)/ father's job (Fjob), going out with friends (goout)/weekend alcohol consumtion (Walc) and age/failures. The correlation was weak for mother's job (Mjob)/nursery, mother's education (Medu)/family support (famsup), school/address, studytime/traveltime, failures/weekday alcohol consumption (Dalc), failures/going out with friends (goout), age/going out with friends (goout) and age/ school support (schoolsup). The following pair of variables showed no association: going out with friends (goout)/schoolsup and failures/school support (schoolsup). The following pair of variables showed a positive correlation: going out with friends (goout)/weekend alcohol consumtion (Walc), failures/weekday alcohol consumption (Dalc), failures/going out with friends (goout), age/failures and age/going out with friends (goout). Study time and travel time showed a negative correlation (note that Pearson C is an absolute value and doesn't give a direction so there's no conflict in direction for study time and travel time).

	Variables	Pearson C	Kendall's tau-b
1	Mjob (N)/ nursery (N)	0.152	-
2	Mjob (N)/ Fjob (N)	0.396	-
3	Fedu (O)/ Fjob (N)	0.464	-
4	Medu (O)/ famsup (N)	0.182	-
5	school (N)/ address (N)	0.269	-
6	studytime (O)/ traveltime (O)	0.166	-0.096
7	goout (O)/ Walc (O)	0.477	0.337
8	goout (O)/ schoolsup (N)	0.099	-
9	failures (O)/ schoolsup (N)	0.014	-
10	failures (O)/ Dalc (O)	0.233	0.175
11	failures (O)/ goout (O)	0.156	0.094
12	age (O)/ failures (O)	0.405	0.209
13	age (O)/ goout (O)	0.234	0.114
14	age (O)/ schoolsup (N)	0.269	-

Table 9: Correlation coefficients for nominal (Pearson C) and ordinal variables (Kendall's tau-b) variables.

A summary of associations can be found in Table 10.

	Variables	Association (N/O)	Type of correlation
1	Mother's job / Nursery	No(N)	-
2	Mother's job / Father's job	Yes(N)	-
3	Father's education / Father's job	Yes(N)	-
4	Mother's education / Family support	Yes(N)/Yes(O)	-
5	School / Address	Yes(N)	-
6	Study time / Travel time	No(N)/Yes(O)	Negative
7	Going out with friends / Weekend alcohol consumption	Yes(N)/Yes(O)	Positive
8	Going out with friends / School support	No(N)/No(O)	-
9	Failures / School support	No(N)/No(O)	-
10	Failures / Weekday alcohol consumption	Yes(N)/Yes(O)	Positive
11	Failures / Going out with friends	Yes(N)/Yes(O)	Positive
12	Age / Failures	Yes(N)/Yes(O)	Positive
13	Age / Going out with friends	No(N)/Yes(O)	Positive
14	Age / School support	Yes(N)/Yes(O)	-
15	Mother's education / Mother's job	Yes(N)	-

Table 10: Summary of associations. N - both variables are treated as nominal. O - at least one variable is treated as ordinal.

For some pair of variables there were cells with expected counts below 5. For father's education (Fedu) and father's job (Fjob) there were seven expected counts below 5 (see Table 11). For mother's education (Medu) and family support (famsup) there were two expected counts below 5 (see Table 12). This will be further discussed in section 5.1 below.

	Fjob				
Fedu	at_home	health	other	services	teacher
0	0 (0)	0 (0)	2 (1)	0 (1)	0 (0)
1	4 (4)	1 (4)	57 (45)	19 (23)	1 (6)
2	9 (6)	3 (5)	69 (63)	34 (32)	0 (8)
3	3 (5)	3 (5)	58 (55)	35 (28)	1 (7)
4	4 (5)	11 (4)	31 (53)	23 (27)	27 (7)

Table 11: Two-way table showing observed and expected counts (bracket) for father's education (Fedu, rows) and father's job (Fjob, columns). 0 – none, 1 – primary education, 2 – 5th to 9th grade. 3 – secondary education, 4 – higher education.

	famsup	
Medu	no	yes
0	2 (1)	1 (2)
1	32 (23)	27 (36)
2	45 (40)	58 (63)
3	36 (38)	63 (61)
4	38 (51)	93 (80)

Table 12: Two-way table showing observed and expected counts (bracket) for mothers's education (Medu, rows) and family support (famsup, columns). 0 – none, 1 – primary education, 2 – 5th to 9th grade. 3 – secondary education, 4 – higher education.

4.2 Logistic regression

The classification was pass (1) or fail (0) mathematics. The true classifications for the entire data set is shown in Table 13. The passing rate in mathematics for this sample was about 67%.

	True classification	Proportion
fail, 0	130	0.329
pass, 1	265	0.671

Table 13: True classification.

4.2.1 Model selection

The training data set was used for model building and the test data set was used for model evaluation. The passing rate in mathematics in the training set (316 observations) was about 66% and in the test set (79 observations) about 71% (see Table 14).

	n	Passing rate
Training	316	0.661
Test	79	0.709

Table 14: Passing rate for training and test set.

The passing rate in the test set was higher than the passing rate in the entire data set (and in the training set). These passing rates were used for comparison with the predicted passing rate in the test set (see section 5.2).

Association for transformed variables
The nominal variables *Mjob*, *Fjob*, *reason* and *guardian* were transformed to dummy variables. Table 15 shows the results from association tests and correlations where the transformed variables are used. Only the pair of variable where previous tests for the original variables gave evidence for association are included in the table (see Table 10). Mother's job (MjobD) and nursery now showed a significant results for association but the *p-value* was quite close to (but below) 0.05 (before it was just above 0.05). All other tests gave the same results as before regarding to whether the variables are associated or not. Some of the correlation coefficient were slightly lower than before.

Variables	X^2 (p-value)	M^2 (p-value)	Pearson C	Kendall's tau-b
MjobD (N)/ nursery (N)	4.3 (0.038)	4.3 (0.038)	0.104	
Medu (O)/ MjobD (N)	117.1 (<0.001)	81.5 (<0.001)	0.478	
Fedu (O)/ FjobD (N)	80.7 (<0.001)	47.5 (<0.001)	0.412	
MjobD (N)/ FjobD (N)	17.8 (<0.001)	17.8 (<0.001)	0.208	
ageT (O)/ failures (O)	77.5 (<0.001)	17.2 (<0.001)	0.405	0.209
ageT (O)/ goout (O)	23.0 (0.115)	5.8 (0.016)	0.234	0.114
ageT (O)/ schoolsup (N)	30.8 (<0.001)	25.9 (<0.001)	0.269	

Table 15: Association tests for transformed variables. X^2 is the statistics for Pearson's chi-squared test of independence. M^2 is the statistics for the generalized Cochran-Mantel-Haenszel test of independence. The D in the end of a variable indicates that the variable is transformed to a dummy variable. The T in the end of the age variables indicates that this variable is transformed to an ordinal variable with five categories. See section 3.2.1.

The reduced data set

To form the reduced data set, associated variables were used one at a time in single logistic regression models (see section 3.2.2). The results are shown in Table 16. The variable with the highest *p-value* was dropped. The variables that were dropped from the full data set to create the reduced data set were mother's job (MjobD), father's job (FjobD), weekend alcohol consumption (Walc), school (school) and weekday alcohol consumption (Dalc) (see Table 16). An exception was made for the associations with age. Age (ageT) was associated with failures (failures), going out with friends (goout) and school support (schoolsup). All these variables were considered important for the model and all *p-values* were low. All these variables were kept.

Pair	p-value	Decision
Medu/MjobD	0.025/0.846	Drop Mjob
Fedu/FjobD	0.067/0.209	Drop Fjob
MjobD/FjobD	0.846/0.209	Both Mjob and Fjob dropped
goout/Walc	<0.001/0.937	Drop Walc
school/address	0.774/0.520	Drop school
Dalc/failures	0.660/<0.001	Drop Dalc
ageT/failures	0.003/<0.001	Keep both
ageT/goout	0.003/<0.001	Keep both
ageT/schoolsup	0.003 / 0.062	Keep both

Table 16: Drop variables to create reduced model.

Another reason to keep the age variable was that there was an age group of students above 18 years of age which probably was repeaters. We know from Table 10 that age was associated with failures and school support. Table 17 shows that the largest group with at least one failure was the group with repeaters.

	Number of failures	
Age	0	at least one
15	71	11
16	88	16
17	84	14
18	63	19
19 and older	6	23

Table 17: The distribution of failures for each age group.

One could think that this group would get support but Table 18 shows that the number of students getting school support seems to decrease with age and in this data set there was only one student above 18 years of age with school support.

	School support	
Age	No	Yes
15	59	23
16	86	18
17	93	5
18	78	4
19 and older	28	1

Table 18: The distribution of school support for each age group.

4.2.2 Model building

The backward elimination procedure was used for fitting the models.

Model 1: *starting from the entire data set, significant level of 15%*
The result for **model 1** is shown in Table 19.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3553	0.8886	2.65	0.0080
PstatusT	-0.6916	0.4685	-1.48	0.1398
Medu	0.2382	0.1548	1.54	0.1238
MjobDteacher	-0.8407	0.4458	-1.89	0.0593
reasonDreputation	0.4911	0.3271	1.50	0.1333
failures	-0.8705	0.2087	-4.17	0.0000
schoolsupyes	-0.9530	0.3989	-2.39	0.0169
famsupyes	-0.5478	0.2947	-1.86	0.0631
internetyes	0.6867	0.3640	1.89	0.0593
romanticyes	-0.6185	0.2970	-2.08	0.0373
freetime	0.2089	0.1428	1.46	0.1436
goout	-0.6087	0.1521	-4.00	0.0001
Walc	0.2562	0.1200	2.13	0.0328
ageT	-0.2112	0.1219	-1.73	0.0833

Table 19: Logistic regression output for model 1.

Model 1 was a model with the following explanatory variables (the brackets shows the baseline of the nominal variables and the star* indicates that the variable has a p-values below 0.05): Pstatus (living apart), Medu, MjobD (other than teacher), reasonD (other than reputation), failures*, schoolsup* (no), famsup (no), internet (no), romantic* (no), freetime, goout*, Walc*, ageT

Some variables showed a negative relationship (Table 19).

- The odds of passing is lower for students with parents living together (Pstatus), with a mother working as a teacher (MjobD), with school (schoolsup) or family support (famsup) or if the student is in a romantic relationship (romantic).
- An increase in level of failures (failures), going out with friends (goout) or age (age) decreases the odds of passing.

The rest of the variables showed a positive relationship (Table 19).

- The odds of passing is higher if the reason for choosing a school is for reputation (reasonD) or if the home has internet (internet).
- An increase in level of Mother's education (Medu), freetime (freetime) or weekend alcohol consumption (Walc) increases the odds of passing.

The failing rate in the training data set for model 1 was about 34% (Table 14). Failing rates grouped by family support, school support and students with mother's working as a teacher are shown i Table 20. The results from Table 20 will be further discussed in the Discussion section.

Variable	Failing rate
Family support	36%
School support	49%
Mother working as a teacher	33%

Table 20: Failing rates for different kind of support and for students with a mother working as a teacher.

Model 2: *starting from the reduced data set, significant level of 15%*
The result for **model 2** is shown in Table 21.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.9974	1.1160	1.79	0.0735
sexM	0.4733	0.2871	1.65	0.0992
PstatusT	-0.7351	0.4591	-1.60	0.1093
reasonDreputation	0.4613	0.3183	1.45	0.1473
failures	-0.7959	0.2058	-3.87	0.0001
schoolsupyes	-0.8742	0.4002	-2.18	0.0289
famsupyes	-0.4550	0.2889	-1.58	0.1152
higheryes	1.1716	0.7905	1.48	0.1383
internetyes	0.6780	0.3521	1.93	0.0542
romanticyes	-0.4575	0.2924	-1.56	0.1176
goout	-0.4140	0.1258	-3.29	0.0010
ageT	-0.1871	0.1220	-1.53	0.1252

Table 21: Logistic regression output for model 2.

Model 2 was a model with the following explanatory variables (the brackets shows the baseline of the nominal variables and the star* indicates that the variable has a p-values below 0.05): sex (female), Pstatus (living apart), reasonD (other than reputation), failures*, schoolsup* (no), famsup (no), higher (no), internet (no), romantic (no), goout*, ageT

None of the variables common to model 1 (Pstatus, reasonD, failures, schoolsup, famsup, internet, romantic, goout, ageT) changed signs (Table 21). Failures, schoolsup and goout still had p-values below 0.05. Romantic was still in the model but now with a p-value above 0.10. The variables sex and higher were now in the model. These variables both showed a positive relationship i.e. the odds of passing is higher for a male student (sex) or for a student that wants to take higher education (higher).

Model 3: *starting from the entire data set, significant level of 5%*
The result for **model 3** is shown in Table 22. The same result was received when starting from the reduced data set (significant level of 5%).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5211	0.4400	5.73	0.0000
failures	-0.8894	0.1956	-4.55	0.0000
schoolsupyes	-0.7788	0.3617	-2.15	0.0313
romanticyes	-0.5437	0.2741	-1.98	0.0473
goout	-0.3909	0.1200	-3.26	0.0011

Table 22: Logistic regression output for model 3.

Model 3 was a model with the following explanatory variables (the brackets shows the baseline of the nominal variables): failures, schoolsup (no), romantic (no) and goout

None of the variables changed sign and there was not a large change in the estimated parameters (Table 22). Influential observations were removed from model 3 for the final model.

Model 4: *G2 as the only explanatory variable, a model for comparison*
Model 4 was a simple logistic regression model with term 2 grades (G2) as the single explanatory variable. This model was used for reference only (Table 23). The variable G2 was statistically significant.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.5584	2.4130	-6.86	0.0000
G2	1.7982	0.2596	6.93	0.0000

Table 23: Logistic regression output for model 4.

AIC was calculated for all models (Table 24). Model 4 had the lowest AIC but this model was only used for comparison. Model 1 had the lowest AIC of the models excluding previous term grades and this was considered as the best model regarding to the Akaike information criterion (AIC).

Model	Limit	AIC
1	0.15	357.85
2	0.15	359.39
3	0.05	364.84
4	0.05	130.27

Table 24: AIC for model 1-4.

Each variable was added as a single variable in a simple logistic regression model to check the sign of the parameter. The results are shown in Table 25 (and will be further discussed in section 5.2.).

Variable	estimate	standard error	p-value
school	-0.103	0.357	0.774
sex	0.388	0.240	0.107
address	0.179	0.278	0.520
famsize	0.056	0.264	0.831
Pstatus	-0.649	0.420	0.122
Medu	0.244	0.109	0.025
Fedu	0.198	0.108	0.067
MjobD	0.066	0.340	0.846
FjobD	0.654	0.520	0.209
reasonD	0.442	1.557	0.119
guardianD	-0.079	0.261	0.762
traveltime	-0.054	0.168	0.746
studytime	0.123	0.143	0.392
failures	-0.954	0.192	<0.001
schoolsup	-0.620	0.332	0.062
famsup	-0.257	0.245	0.294
paid	0.430	0.243	0.077
activities	-0.063	0.239	0.793
nursery	-0.123	0.307	0.690
higher	1.957	0.670	0.003
internet	0.491	0.307	0.110
romantic	-0.622	0.250	0.013
famrel	0.004	0.135	0.979
freetime	0.001	0.117	0.996
goout	-0.399	0.112	<0.001
Dalc	-0.059	-0.441	0.659
Walc	-0.007	0.091	0.937
health	-0.090	0.089	0.310
absences	-0.016	0.017	0.247
failures	-1.520	0.292	<0.001
ageT	-0.291	0.098	0.003

Table 25: Simple logistic regression models for each of the variables.

4.2.3 Model evaluation

Hosmer-Lemeshow Goodness of fit test

Hosmer-Lemeshow Goodness of Fit Tests was performed for all four models. The hypotheses for all models:

$$H_0\colon\ E(Y) = \frac{1}{1+e^{-X'\beta}}$$
$$H_1\colon\ E(Y) \neq \frac{1}{1+e^{-X'\beta}}$$

The test statistic for model 1 (16 groups): $X^2 \approx 15.8 < \chi^2(1-\alpha, 14) \approx 23.7$ and the $p-value \approx 0.324$ was considered high enough to fail to reject H_0. There was not enough evidence to reject the null hypothesis that the overall fit of the model was appropriate.

Hosmer-Lemeshow Goodness of Fit Tests for all models are shown in table 26 below. All models gave the same result. There was not enough evidence to reject the null hypothesis that the overall fit was appropriate.

Model	X^2	df	p
1	15.8	14	0.324
2	14.5	12	0.268
3	5.26	5	0.385
4	0.39	2	0.823

Table 26: Hosmer-Lemeshow Goodness of Fit Tests for model 1-4.

Logistic regression diagnostics

Plots of Cook's distances are shown in Figure 9.

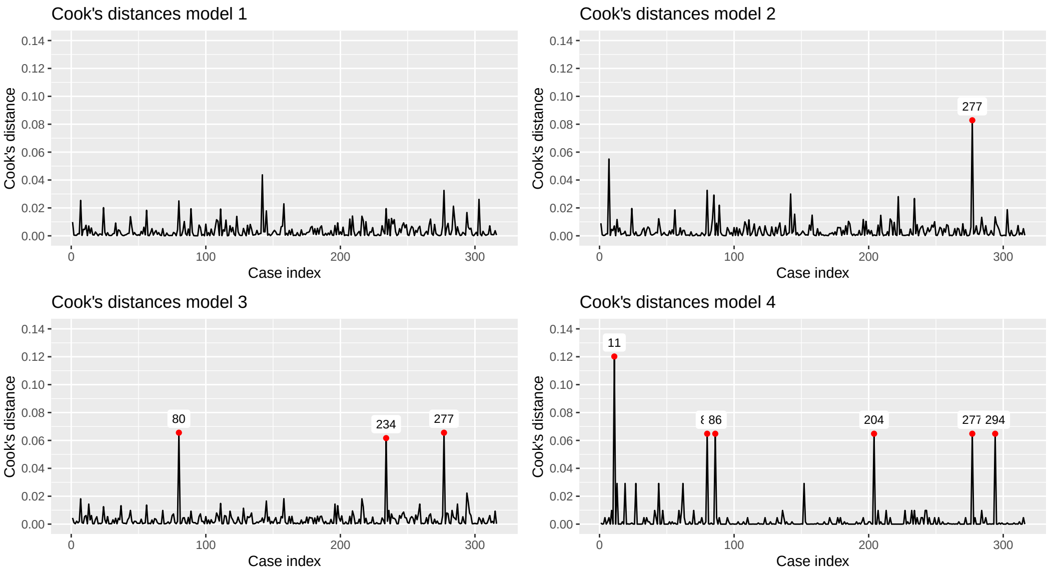


Figure 9: Cook's distances for model 1-4. Distances above 0.06 are marked with it's corresponding index.

Model 2-4 have some influential observations (the limit was set to 0.06). For model 2 and 3 two different models were evaluated: one model with and one model without influential observation (called *Model 2 without* and *Model 3 without*).

Pearson and Studentized Pearson residuals are plotted for all models (Figure 10). Notice that the plots for Pearson and Studentized Pearson residuals are similiar. The explanation is that the hatvalues used for standardisation are low which give a conversion factor close to 1 (see Equation (3.14)). No models showed problems.

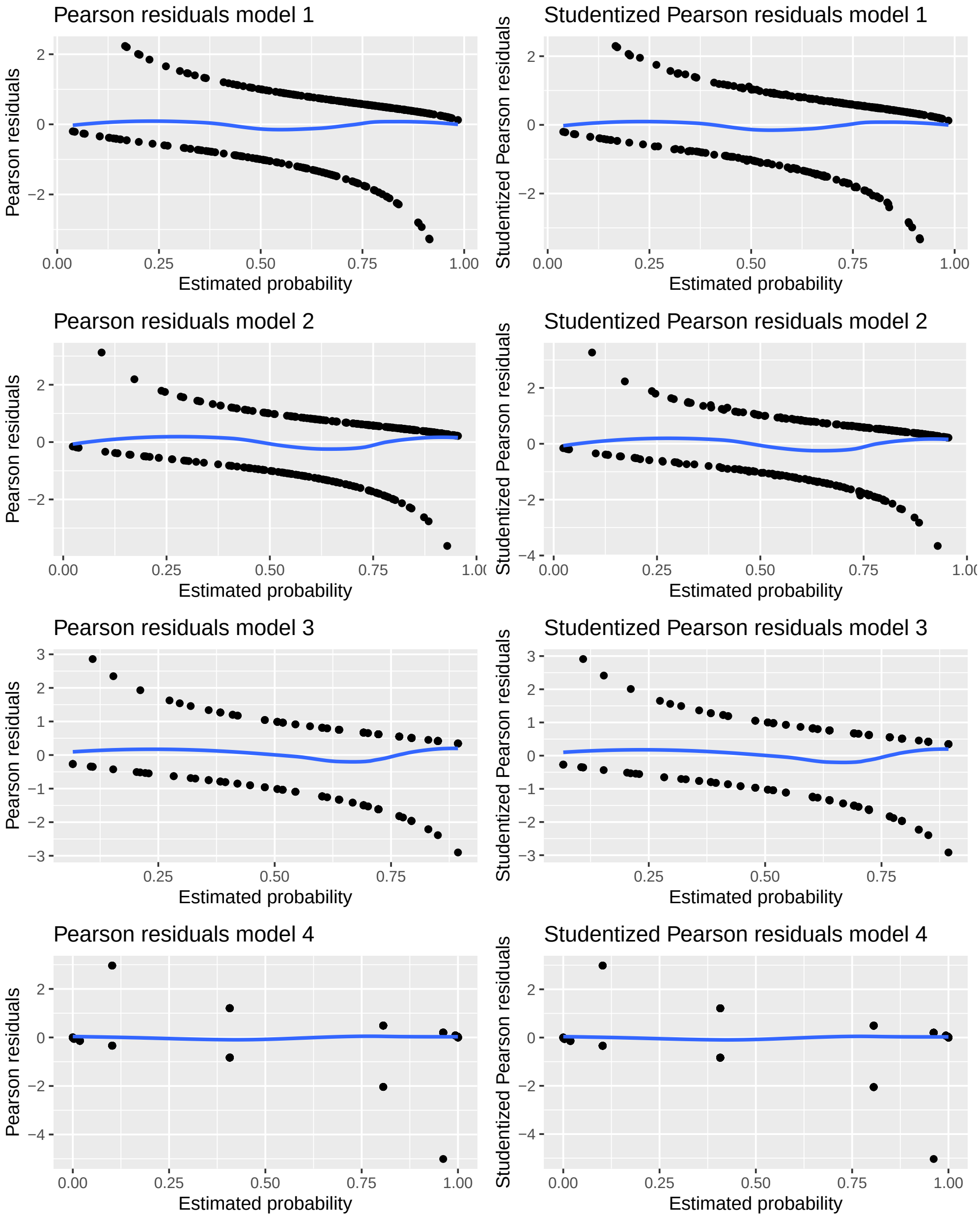


Figure 10: Pearson and Studentized Pearson residuals for model 1-4.

4.2.4 Further model evaluation

AIC and *AUC* was calculated for all models. The results are shown in Table 27.

Model	Limit	AUC	AIC
1	0.15	0.679	357.85
2	0.15	0.682	359.39
2 without	0.15	0.681	353.67
3	0.05	0.684	364.84
3 without	0.05	0.689	350.79
4	0.05	0.985	130.27

Table 27: *AUC* and *AIC* for model 1-4 (including model 2 and 3 without influential observation).

One model, model 3 without, had both highest AUC and lowest AIC (remember that model 4 is only for comparison). $AUC \approx 0.689$ for this model and according to the rule of thumb this means that the model is poor (but close to acceptable). This *AUC* can be compared with model 4 (the model with only term 2 grades as explanatory variable) with an *AUC* of 0.985 which is an outstanding model. Model 3 without influential observations was selected as the best model and was further evaluated.

Evaluation of model 3 without influential observations
Overall accuracy (ACC), *sensitivity* and *specificity* for different cutoffs are plotted in Figure 11 for model 3 without influential observations (using the training set).

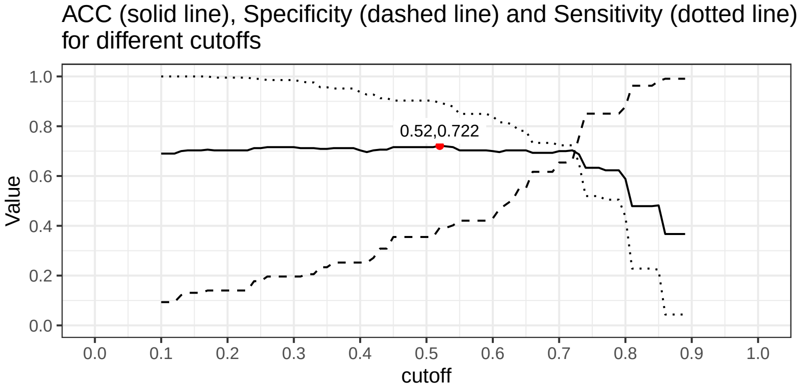


Figure 11: Error plot for model 3 without influential observations using test set. Marked point is the point for maximum ACC.

ACC didn't change much for cutoffs below 0.75. *Sensitivity* decreased and *specificity* increased with increasing cutoff. The cutoff was chosen for maximum ACC and where the *Sensitivity* exceeded 90%. The purpose was to get a high proportion of correctly classified pass and fail grades and a high proportion of actual pass grades that were correctly classified. The point for maximum ACC ($ACC \approx 0.722$) is marked in Figure 11. A cutoff of 0.52 was chosen which means that for estimated probabilities exceeding 0.52 the model will predict a pass grade (1) in mathematics.

Confusion matrix model 3 without influential observations
The confusion matrix for model 3 without influential observations (using the test set and a cutoff of 0.52) are shown in Table 28.

	Actual class, Y	
Predicted class, $\hat{Y}$	0	1
0	11	5
1	12	51

Table 28: Confusion matrix, cutoff = 0.52.

The predicted passing rate in the test set is about 80% (Table 28) which is higher than the true passing rate in the test set (71%, Table 14) and the true passing rate in the entire data set (67%, Table 13). *ACC*, *error rate*, *sensitivity* and *specificity* for model 3 without influential observation and for model 4 (for comparison) are shown in Table 29.

Model	cutoff	ACC	Error rate	Sensitivity	Specificity
3 without	0.52	0.785	0.215	0.911	0.478
4	0.52	0.937	0.063	0.929	0.957

Table 29: *ACC*, *error rate*, *sensitivity* and *specificity* for model 3 without influential observations and for model 4.

Further results from Table 29 will be discussed in the Discussion section.

ROC-curve for model 3 without influential observations
The *ROC*-curve for model 3 without influential observations is plotted in Figure 12.

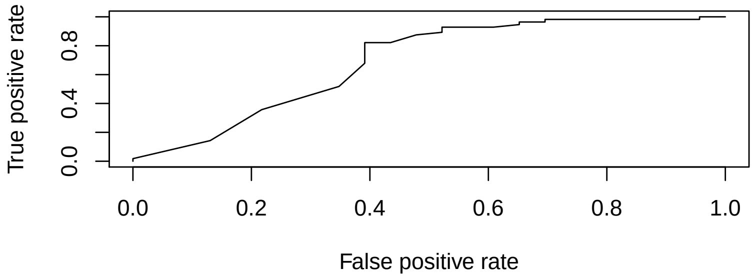


Figure 12: ROC-curve for model 3 without. True positive rate = *sensitivity* and false positive rate = $1 - specificity$.

The AUC value of the *ROC-curve* in Figure 12 is shown in Table 27 and discussed above.

4.2.5 Fitted logistic response function for model 3 without influential observations

Model 3 without influential observations was chosen as the final model. This model was fitted on the entire data set. The result is shown in Table 30.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3534	0.3964	5.94	0.0000
failures	-1.2150	0.2029	-5.99	0.0000
schoolsupyes	-0.9297	0.3328	-2.79	0.0052
romanticyes	-0.3291	0.2505	-1.31	0.1890
goout	-0.3193	0.1090	-2.93	0.0034

Table 30: Model 3 fitted on entire data set without influential observations.

Notice from Table 30 that the variable romantic was not significant at a 5% level when the influential observations were removed (this variable was significant at a 5% level in model 3 from Table 22). The standard error for this variable was large. This model had an AIC of around 436 which was much higher than the AIC (351) from model 3 without influential observations fitted on the training set (Table 31).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5347	0.4511	5.62	0.0000
failures	-1.2462	0.2433	-5.12	0.0000
schoolsupyes	-0.9655	0.3691	-2.62	0.0089
romanticyes	-0.4427	0.2828	-1.57	0.1175
goout	-0.3802	0.1234	-3.08	0.0021

Table 31: Model 3 fitted on training set without influential observations.

The fitted logit response function for the final model:

$$logit(\hat{\pi}) = 2.353 - 1.215 failures - 0.930 schoolsup - 0.329 romantic - 0.319 goout \tag{4.1}$$

The fitted logistic response function:

$$\hat{\pi} = \frac{1}{1 + exp(-2.353 + 1.215 failures + 0.930 schoolsup + 0.329 romantic + 0.319 goout)} \tag{4.2}$$

95% confidence interval for odds ratio
Using (3.18) and (3.19) with $z(1-\alpha/2) = z(1-0.05/2) \approx 1.960$.
The confidence limits for the odds ratio $exp(\beta_k)$ is $exp(b_k \pm 1.960s(b_k))$

Confidence intervals for:

- *failures (failures)*
  Confidence limits for the odds ratio $exp(\beta_1)$:
  $exp(b_1 \pm 1.960s(b_1)) = exp(-1.2150 \pm 1.960 \cdot 0.2029)$
  95% confidence interval for odds ratio $exp(\beta_1)$: $0.199 \leq exp(\beta_1) \leq 0.442$

- *school support (schoolsup)*
  Confidence limits for the odds ratio $exp(\beta_2)$ (schoolsup):
  $exp(b_2 \pm 1.960s(b_2)) = exp(-0.9297 \pm 1.960 \cdot 0.3328)$
  95% confidence interval for odds ratio $exp(\beta_2)$: $0.206 \leq exp(\beta_2) \leq 0.758$

- *in a romantic relationship (romantic)*
  Confidence limits for the odds ratio $exp(\beta_3)$:
  $exp(b_3 \pm 1.960s(b_3)) = exp(-0.3291 \pm 1.960 \cdot 0.2505)$
  95% confidence interval for odds ratio $exp(\beta_3)$: $0.440 \leq exp(\beta_3) \leq 1.176$

- *going out with friends (goout)*
  Confidence limits for the odds ratio $exp(\beta_4)$:
  $exp(b_4 \pm 1.960s(b_4)) = exp(-0.3193 \pm 1.960 \cdot 0.1090)$
  95% confidence interval for odds ratio $exp(\beta_4)$: $0.587 \leq exp(\beta_4) \leq 0.900$

4.2.6 Interpretation of parameters for model 3 without influential observations

- *failures*: $exp(-1.215) \approx 0.30$, the estimated odds of passing decreases by about 70% with each additional failure if all other variables are held constant in the sample. The 95% confidence interval for the odds ratio for *failures* is $0.199 \leq exp(\beta_1) \leq 0.442$. We are 95% confident that the estimated odds of passing decreases by between 56% and 80% with each additional failure if all other variables are held constant in the population.

- *schoolsup*: schoolsup is a dummy variable (the baseline is no school support). $exp(-0.930) \approx 0.39$, the estimated odds of student with school support passing is about 0.39 times those without school support if all other variables are held constant in the sample. The 95% confidence interval for the odds ratio for *schoolsup* is $0.206 \leq exp(\beta_2) \leq 0.758$. We are 95% confident that the estimated odds of students with school support passing is between 0.21 and 0.76 times those without school support if all other variables are held constant in the population.

- *romantic*: romantic is a dummy variable (the baseline is that the student is not in a romantic relationship). $exp(-0.329) \approx 0.72$, the estimated odds of students in a romantic relationship passing is about 0.72 times those not in a romantic relationship if all other variables are held constant in the sample. The 95% confidence interval for the odds ratio for *romantic* is $0.440 \leq exp(\beta_3) \leq 1.176$. We are 95% confident that the estimated odds of students in a romantic relationship passing is between 0.44 and 1.18 times those not in a romantic relationship if all other variables are held constant in the population. This means that we don't know whether or not there's a lower odds of passing for students in a romantic relationship in the population.

- *goout*: $exp(-0.319) \approx 0.73$, the estimated odds of passing decreases by about 27% with each additional increase in level of going out with friends if all other variables are held constant in the sample. The 95% confidence interval for the odds ratio for *goout* is $0.587 \leq exp(\beta_4) \leq 0.900$. We are 95% confident that the estimated odds of passing decreases by between 10% and 41% with each additional increase in level of going out with friends if all other variables are held constant in the population.

5 Discussion

The overall aim of this thesis was to examine if other factors than previous grades affected secondary student performance in mathematics in Portugal and if some of the background characteristics were associated.

5.1 Associations

The analysis started with some variables that were supposed to be associated like education and job of the parents to see how the tests worked. All tests showed a significant association between the education of parents (Fedu/Medu) and their jobs (FjobD/MjobD). There was also a significant association between school and address which was also something to expect from the comparison between the schools (Table 1). The address was divided in *rural* and *urban* and the comparison between the schools showed that one of the schools (Gabriel Pereira) had mainly students from an urban address.

The tests also showed that there was an significant association between mother's and father's job. Mother's education and family support (famsup) were associated which suggested that the mother was in charge of the family support. Going out with friends (goout) and weekend alcohol consumption (Walc) were also associated. Since the student probably can't consume alcohol at home, he or she needs to go out to do that. This association was not surprising. Failures (failures) was associated with weekday alcohol consumption (Dalc), going out with friends (goout) and with age (age). Age (age) was also associated with going out with friends (goout) and school support (schoolsup). Spending more time outside of the house also indicated that the student didn't spent as much time at home for studies as needed.

A surprising result was that failures and school support was not associated. From the comparison between the schools one found that one of the schools Mousinho da Silveira (MS) didn't offer support. One could have expected that past failures was a ground for getting school support. But maybe the failures was in a subject that didn't offer support. To understand why failures and school support wasn't associated, a further investigation of the school support offered at Gabriel Pereira (GP) is needed.

Assumptions for Pearson's chi-squared test of independence

- **Independent random sampling.**
  We can't say that the sampling was neither independent nor random. There's not enough information to say if the design of the study was completely randomised. It's not likely that it was. It's known that Cortez and Silva used a questionnaire to collect most of the data (some data was collected from school reports). But how was the student selected? Did they select entire classes or just a few students from all classes? Why were there so few observations (students) from the school Mousinho da Silveira (MS)? Did the student answer the questionnaire sitting next to each other? This might have affected their answers. Some students probably came from the same family. Students' behavior is unlikely to be independent.
- **The categories of the variables must be mutually exclusive.**
  The categories was mutually exclusive.
- **The counts in each category should not be too small. A rule of thumb is that the estimated expected counts should exceed five.**
  There are cells with estimated expected counts below five for the variables $Medu \sim Mjob$, $Fedu \sim Fjob$ and $Medu \sim famsup$ (see Table 5, Table 11 and Table 12). The chi-squared based tests can't be trusted for these pair of variables.

5.2 Logistic regression models

Model 4 with only term 2 grades (G2) as explanatory variable outperformed all models. The second specific research question was "Are student background characteristics, other than previous grades, related to whether or not a student receives a pass grade in mathematics?". Since model 4 used previous grades it was only a comparison model. Model 3 without influential observations (starting from the entire data set, significant level of 5%) was the best model (highest AUC and lowest AIC) using only background variables (excluding previous term grades).

In the models, some of the variables seemed to get the "wrong" sign and the standard errors of some of the parameters were high indicating that there could be a problem with the model. Maybe the problem was a multicollinearity problem. Table 25 showed that the variable *MjobD*, *freetime* and *Walc* had changed sign from when used as single variable in a simple logistic regression model (still with pass as the dependent variable). It's reasonable to think that the odds of passing increases if the student's mother is working as a teacher. Table 20 shows that the failing rate for these student is about the same as for all students in the entire data set (Table 13) but nothing tells us that the odds of passing would decrease if the student's mother is working as a teacher. It's also reasonable to think that the odds of passing decreases if the level of freetime or weekend alcohol consumption increases.

A surprising result was that the odds of passing decreased for a student getting school or family support. But one need to bear in mind that students usually get support because they are at risk of failing the subject. Table 20 shows that the failing rate is much higher in the group getting school support (49% failing rate compared to 34% for the entire data set) and slightly higher in the group getting family support (36% failing rate compared to 34%). So the signs of the parameters for support were reasonable.

Model evaluation

The results from the Hosmer-Lemeshow Goodness of fit tests for all four models showed that there was not enough evidence to reject the null hypothesis that the overall fit was appropriate for any of the models.

The plots of Cook's distances (Figure 9) showed some influential observations for model 2-4. The Pearson and Studentized Pearson residual plots (Figure 10) showed an adequate fit for all models.

Further model evaluation

For model evaluation different measurement can be used. *Overall accuracy*, *sensitivity* and *specificity* depends on chosen cutoff. A measurement that doesn't depend on the cutoff is the area under the ROC-curve, AUC. A model with high AUC is a good model. The model with highest AUC was model 3 without influential observations (see Table 27).

The predicted passing rate in the test set (80%, calculated from the confusion matrix in Table 28) was higher than the true passing rate (71% for test set, 66% for training set and 67% for the entire data set, from Table 14 and Table 13). The *Sensitivity* was about 91% i.e. 91% of actual pass grades were correctly classified (compare with model 4 with $Sensitity \approx 93\%$). The *overall accuracy* (ACC) was about 78% i.e. 78% of all grades were correctly classified (compare with model 4 with $ACC \approx 94\%$). The *error rate* was 0.215 (compare with model 4 with *error rate* ≈ 0.063). The confusion matrix in Table 28 also shows that the model predicted pass grades when the true grade was a fail grade too often. This is because of the low *specificity. Specificity* in this case was the proportion of actual fail grades that was correctly classified. Since the *Specificity* was about 48% (compare with model 4 with $Specificity \approx 96\%$), then 52% of the fail grades were classified as pass grades (Table 29). The model was much better at predicting pass grades when the actual grade was a pass grade (sensitivity in Table 29 is high). This was not surprising since the line between passing and failing is a fine line and it depends on many factors. A student can for example get a fail grade for missing an important exam without showing any other signs of failing.

The final logistic regression model

The model chosen as the final model was the model with lowest AIC (Table 27). This was model 3 without influential observations ($AIC \approx 351$, compare with model 4 with $AIC \approx 130$). The same explanatory variable was used in the final model (failures, schoolsup, romantic and goout) but the model was build from the entire data set but without the influential observations from the training data set. Notice that the variable *romantic* was non-significant at a 5% level when the entire data set was used (Table 30 shows that the p-value for romantic is about 0.19).

Assumptions of binary logistic regression

- **Independent random sampling.**
  See "Assumptions for Pearson's chi-squared test of independence" above.
- **Binary response variable.**
  The response variable was binary (fail or pass).
- **Little or no multicollinearity among the explanatory variables.**

There were some correlations among the explanatory variables and there were signs of multicollinearity (for example some of the parameters seemed to get the "wrong" sign.)

- **Linearity of explanatory variables and the logit response function.**
  This assumptions was not checked.

- **Large sample.**
  395 observation is usually considered a large sample.

5.3 Limitations

Data set
The data set has 395 observation from two schools but only 46 of these came from the school Mousinho da Silveira. The fact that the schools differed may have affected the results. For example Mousinho da Silveira did not offer school support while the school Gabriel Pereira did. It would have been better to either only focus on Gabriel Pereira or to collect more observations from Mousinho da Silveira to get a better balanced data set. The data set included some repeaters (older than 18 years old) that may have affected for example the association founded between failures (failures) and age (age). The repeaters probably had to repeat because of failures. Table 17 shows that in this age group many students had one or more failures.

One can ask if it was necessary with variables for example both parent's education and job. Usually there's a correlation between education and job. Maybe some of the variables describing how students spent their freetime could be removed. Weekend alcohol consumption and weekday alcohol consumption could be merged to one variable describing alcohol consumption. Alcohol consumption and going out with friends was also associated. So it would have been a good idea to think more about the "freetime" variables and what to measure.

Associations between categorical variables
Since almost all background variable were categorical, the possible number of independent tests were limited. For the nominal variables a classical test based on the chi-squared distribution was used but it was harder to find reliable tests that took into consideration the ordering of the ordinal variables. It was especially difficult when comparing ordinal and nominal variables. The ordinal variables had few levels, resulting in many ties. Kendall's tau-b was adjusted for ties and hopefully gave trustful results. Rank tests like for example Wilcoxon rank-sum test was also rejected because of the ties. Maybe rank test was more reliable than the generalized Cochran-Mantel-Haenszel test of independence. Perhaps more research would have been required to find better tests, especially for when there's a mixture of ordinal and nominal variables. It was also difficult to know if a result could be trusted if any of the assumptions for the tests were violated.

Logistic regression models
The models may have had multicollinearity problem. Since correlation for nominal and ordinal data was not measured in the same way, a correlation matrix for all variables couldn't be made. Instead the variables were compared pairwise but correlation coefficients were not calculated for all pairs. There may be more associations than the ones found here. What limit should be chosen for possible multicollinearity problem when different measurements of correlation are used? The method used for preparing the reduced data set can be improved. It is also possible that the results would differ if the dummy variables were transformed into other groups.

The fitting of the model could have been improved by increasing the training set or the evaluation could have been improved by increasing the test set. If the significance level in the backward elimination procedure had been increased to 20%, perhaps some other important variables would have been revealed.

There were some outliers, student with zero grades. These outliers were assigned to the group of failures. It's not known why the students received a zero grade. Did some of them receive a zero grade for missing an important exam or for being ill? If so, some of these students would probably have passed mathematics under normal circumstances and are therefore student with more similar characteristic to those who passed. Maybe the outliers are one of the problems with the low *specificity* and should have been removed.

One of the assumptions for logistic regression models was not checked. The assumption was "linearity of explanatory variables and the logit response function".

Evaluation
It was not possible to maximize both *Specificity* and *Sensitivity.* One had to decide what was more important. To choose a cutoff where the proportion of correctly classified pass grades was high (a high *Sensitivity* is needed) or a cutoff where the proportion of correctly classified fail grades was high (a high *Specificity* is needed). One could also choose a cutoff where the proportion of correctly classified pass and fail grades was high (a high ACC is needed). The models could not be compared using ACC, *Specificity* or *Sensitivity* since the optimal cutoff differed for the models. These measurement were only calculated for model 3 without influential observations and for model 4. Instead AUC was used to compare the models.

6 Conclusions

This thesis shows that a logistic regression model with only previous grade as explanatory variable (model 4) classified pass and fail grades in mathematics well. Similar results was found by Kotsiantis et. al. [18]. Logistic regression models using only variables describing student background characteristics did not classify into pass and fail grades as well. In addition, association between the background variables also caused problems with multicollinearity.

Using the variables from the *student performance data* set but excluding all previous grade variables (G1 and G2) gave poor logistic regression models. The best model (model 3) had four explanatory variables (all statistically significant, $p < 0.05$, before the model was adjusted): failures (failures), school support (schoolsup), in a romantic relationship (romantic) and going out with friends (goout). These variables all affected student performances in mathematics (measured in whether or not the student received a pass grade in mathematics) negatively. Removing influential observations resulted in a model with better predictive power. Still, the model's predicted passing rate was too high (80% compared to the true value of 71% in the test set) and it predicted too often a pass grade when the actual grade was fail (52% of the true fail grades were classified as pass grades). 78% of the grades were correctly classified. There was a 69% chance that this model was able to distinguish between pass and fail grades. The limit for an acceptable model is 70%.

Further studies
The investigated data set in this study is clearly lacking important factors that may affect student performance, like for example emotional health (stress, perfectionism). Mary E. Pritchard and Gregory S. Wilson performed a study in 2003 with purpose "to investigate the impact of student emotional health and social health on college student GPA and retention" [20]. They found that both emotional and social health factors affect student performance and retention. The variables used for this study based on the *student performance data set* was mainly focused on social health, family background and external school factors. It lacks important information about internal school factors. Class size, student to teacher ratio, computer to student ratio, experienced teachers, teacher shortage, activities in class, motivation, student participation in class, homework, math anxiety and social attitudes are all important factors that may affect student performance in mathematics. Do internal school factors affect students' mathematical performance in Portugal?

7 References

References

[1] Dua, D. & Graff, C. (2020). *UCI Machine Learning Repository.* http://archive.ics.uci.edu/ml/datasets/Student+Performance?ref=datanews.io. Irvine, CA: University of California, School of Information and Computer Science.

[2] Cortez,P. & Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance.* In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[3] Agresti, A. (2008). *An Introduction to Categorical Data Analysis.* 2. ed. New York: John Wiley & Sons Inc.

[4] Agresti, A. (2013). *Categorical Data Analysis.* 3rd. ed. New York: John Wiley & Sons Inc.

[5] Friendly, M. & Meyer, D. (2015). *Discrete Data Analysis with R.* New York: Productivity Press.

[6] Lantz, B. (2013). *Machine learning with R* Birmingham: Packt Publishing Ltd.

[7] Alm, S.& Britton, T. (2018). *Stokatisk. Sannolikhetsteori och statistikteori med tillämpningar.* Stockholm: Liber.

[8] Wackerly, D., Mendenhall III, W. & Scheaffer, R. (2019) *Mathematical Statistics with Applications.* 7th. ed. Ashford: Ashford Colour Press Ltd.

[9] Kutner, M.,Nachtsheim, C. & Neter, J. (2008) *Applied linear regression models.* 4th. ed. New York: McGraw-Hill Education.

[10] Hosmer, D. & Lemeshov, S. (2014) *Applied logistic regression* 2nd. ed. New York: John Wiley & Sons Inc.

[11] Sheskin, D.J. (2000) *Handbook of parametric and nonparametric statistical procedures.* 2nd. ed. Florida: CRC Press LLC.

[12] Sprent,P. & Smeeton, N.C. (2001) *Applied nonparametric statistical methods.* 3rd. ed. Florida: CRC Press LLC.

[13] Pennstate Eberly College of science (2018) *Analysis of Discrete Data.* https://online.stat.psu.edu/stat504/ [2020-01-28]

[14] AcaStat Software (2020) *Applied Statistics Handbook.* https://www.acastat.com/Pub/Docs/AppliedStatistics.pdf [2020-02-05]

[15] University of Florida (Athienitis, D.) (2020) *Categorical Data Analysis.* http://users.stat.ufl.edu/ athienit/STA4504/class_notes_4504.pdf [2020-03-04]

[16] Eurydice (European Commission) (2020) *Portugal Overview* https://eacea.ec.europa.eu/national-policies/eurydice/content/portugal_en [2020-03-23]

[17] OECD (2019) *Programme for international student assessment (PISA) results from PISA 2018.* http://www.oecd.org/pisa/publications/PISA2018_CN_PRT.pdf [2020-03-23]

[18] Kotsiantis S., Pierrakeas C. & Pintelas P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence* (AAI), 18, no. 5, 411–426.

[19] Pereira, M. *An analysis of Portuguese students' performance in the OECD Programme for International Student Assessment (PISA)* https://pdfs.semanticscholar.org/b9e9/8eefe42ebfd249cf58dc9e52e754107be860.pdf [2020-03-26]

[20] Pritchard,M. & Wilson, G. (2003). Using Emotional and Social Factors to Predict Student Success. *Journal of College Student Development*, 44, no. 1, 18-28.

A List of variables and their abbreviations

1. school – student's school, nominal (binary): 'GP' – Gabriel Pereira or 'MS'– Mousinho da Silveira.
2. sex – student's sex, nominal (binary): 'F' – female or 'M'– male.
3. age – student's age, quantitative discrete, interval: 15 to 22 years
4. address – student's home address, nominal (binary): 'U' – urban or 'R' – rural
5. famsize – family size, nominal (binary): 'LE3' – ≤ 3 or 'GT3' – > 3
6. Pstatus – parent's cohabitation status, nominal (binary): 'T' – living together or 'A' – living apart
7. Medu – mother's education, ordinal (five levels): 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
8. Fedu – father's education, ordinal (five levels): 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education
9. Mjob – mother's job, nominal (five levels): 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other'
10. Fjob – father's job, nominal (five levels): 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other'
11. reason - reason to choose this school, nominal (four levels): 'close to home', 'school reputation', 'course preference' or 'other'
12. guardian – student's guardian, nominal (three levels): 'mother', 'father' or 'other'
13. traveltime – home to school travel time, ordinal (four levels): 1 – $<$15 min, 2 – 15 to 30 min, 3 – 30 min to 1 hour or 4 – $>$1 hour
14. studytime – weekly study time, ordinal (four levels): 1 – $<$2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$10 hours
15. failures – n, number of past class failures, ordinal (four levels): if $0 \leq n \leq 2$, else 3
16. schoolsup – extra educational support, nominal (binary): 'yes' or 'no'
17. famsup – family educational support, nominal (binary): 'yes' or 'no'
18. paid – extra paid classes within the course subject (Math), nominal (binary): 'yes' or 'no'
19. activities – extra-curricular activities, nominal (binary): 'yes' or 'no'
20. nursery – attended nursery school, nominal (binary): 'yes' or 'no'
21. higher – wants to take higher education, nominal (binary): 'yes' or 'no'
22. internet – Internet access at home, nominal (binary): 'yes' or 'no'
23. romantic – with a romantic relationship,nominal (binary): 'yes' or 'no'
24. famrel – quality of family relationships, ordinal (five levels): from 1 – 'very bad' to 5 – 'excellent'
25. freetime – free time after school,ordinal (five levels): from 1 – 'very low' to 5 – 'very high'
26. goout – going out with friends, ordinal (five levels): from 1 – 'very low' to 5 – 'very high'
27. Dalc – workday alcohol consumption, ordinal (five levels): from 1 – 'very low' to 5 – 'very high'
28. Walc – weekend alcohol consumption, ordinal (five levels): from 1 – 'very low' to 5 – 'very high'
29. health – current health status, ordinal (five levels): from 1 – 'very bad' to 5 – 'very good'
30. absences – number of school absences, quantitative discrete, interval: from 0 to 93'
31. G1, first period grade, quantitative discrete, interval: from 0 to 20
32. G2, second period grade, quantitative discrete, interval: from 0 to 20
33. G3, final grade, quantitative discrete, interval: from 0 to 20

B R code

```
#---------------------------------------------------
# Load "student performance data set":
student <- read.csv("student-mat.csv",sep=";")

#------------------------------------------------------------------------
# ASSOCIATIONS. Pairwise comparison
#------------------------------------------------------------------------
library(vcd)
library(vcdExtra)

# Select variables, Medu/Mjob:
MeduMjob<-dplyr::select(student,Medu,Mjob)

# Two-way table with observed proportion (Table 6):
(Rowproptable<-round(prop.table(table(MeduMjob), 1),2))
# Two-way table with observed and expected counts (Table 5):
# The table was made manually in Latex using contTables below
library(jmv)
contTables(MeduMjob, rows=’Medu’, cols=’Mjob’,chiSq=TRUE,exp=TRUE,pcRow=TRUE)

#Pearson chi-squared test of independence and Pearson C
MMeduMjob = as.matrix(TMeduMjob)
assocstats(MMeduMjob)

# Select variables, studytime/traveltime:
studytimetraveltime<-dplyr::select(student,studytime,traveltime)

# CMH-test (Table 8)
Mstudytimetraveltime = as.matrix(table(studytimetraveltime))
CMHtest(Mstudytimetraveltime)

# Correlation, Kendall’s tau-b (Table 9)
cor(student$traveltime,student$studytime,method="kendall")

#------------------------------------------------------------------------
# MULTIPLE LOGISTIC REGRESSION MODELS
#------------------------------------------------------------------------
#----------------------------------------
# 1. Create training and test set.
#----------------------------------------
set.seed(200)
n<-nrow(student)
shuffled_stud <- student[sample(n), ]
train_indices <- 1:round(0.8 * n)
train <- shuffled_stud[train_indices, ]
test_indices <- (round(0.8 * n) + 1):n
test <- shuffled_stud[test_indices, ]
nrow(train)
nrow(test)

train$pass<-ifelse(train$G3>=10,1,0)
train$pass<-as.factor(train$pass)
test$pass<-ifelse(test$G3>=10,1,0)
test$pass<-as.factor(test$pass)
student$pass<-ifelse(student$G3>=10,1,0)
student$pass<-as.factor(student$pass)
```

```
#Passing rates, entire (Table 13)
mean(student$pass==1)
#Passing rates, train/test (Table 14)
mean(train$pass==1)
mean(test$pass==1)

# Tranformation of nominal variables to dummy variables
train$reasonD<-as.factor(ifelse(train$reason=="reputation","reputation",0))
train$guardianD<-as.factor(ifelse(train$guardian=="mother","mother",0))
train$FjobD<-as.factor(ifelse(train$Fjob=="teacher","teacher",0))
train$MjobD<-as.factor(ifelse(train$Mjob=="teacher","teacher",0))
test$reasonD<-as.factor(ifelse(test$reason=="home","home",0))
test$guardianD<-as.factor(ifelse(test$guardian=="mother","mother",0))
test$FjobD<-as.factor(ifelse(test$Fjob=="teacher","teacher",0))
test$MjobD<-as.factor(ifelse(test$Mjob=="teacher","teacher",0))

# Tranformation of age (five levels, one category for age > 19)
train$ageT<-NULL
train$ageT<-train$age
train$ageT<-ifelse(train$ageT>=19,19,train$age)
train$ageT<-as.factor(train$ageT)
levels(train$ageT)<-c(1,2,3,4,5)
train$ageT<-as.integer(train$ageT)
test$ageT<-NULL
test$ageT<-test$age
test$ageT<-ifelse(test$ageT>=19,19,test$age)
test$ageT<-as.factor(test$ageT)
levels(test$ageT)<-c(1,2,3,4,5)
test$ageT<-as.integer(test$ageT)

# Select variables (entire data set)
train_adjusted<-dplyr::select(train,-c(G1,G2,G3,reason,guardian,Fjob,Mjob,age))
test_adjusted<-dplyr::select(test,-c(G1,G2,G3,reason,guardian,Fjob,Mjob,age))

# Select variables (reduced data set)
train_adjusted2<-dplyr::select(train_adjusted,-c(MjobD,FjobD,Walc,school,Dalc))
test_adjusted2<-dplyr::select(test_adjusted,-c(MjobD,FjobD,Walc,school,Dalc))

#------------------------------------------
# 2. Single logistic regression model
#------------------------------------------
single.model1<- glm(pass ~ school, data = train_adjusted, family = binomial)
# Table 25 row 1 (variabel school)
summary(single.model1)

#------------------------------------------
# 3. Multiple logistic regression models. Backward selection.
#------------------------------------------
#--------------------
# MODEL 1: Entire 15%
#--------------------
library(StepReg)

stepwiselogit(data=train_adjusted, y="pass", exclude = NULL, include = NULL, selectio
select = "SL", sle = 0.15, sls = 0.15, goft = TRUE)
model.full15 <- glm(pass ~Pstatus+Medu+MjobD+reasonD+failures+schoolsup
+famsup+internet+romantic+freetime+goout+Walc+ageT, data = train_adjusted, family = b
# Table 19
(s.full15<-summary(model.full15))
```

```
AIC.full15<-AIC(model.full15)

#--------------------
# MODEL 2: Reduced 15%
#--------------------
stepwiselogit(data=train_adjusted2, y="pass", exclude = NULL, include = NULL, selecti
select = "SL", sle = 0.15, sls = 0.15, goft = TRUE)
model.reduced15<-glm(pass~sex+Pstatus+reasonD+failures+schoolsup+famsup+
higher+internet+romantic+goout+ageT,
data=train_adjusted2,family = binomial)
# Table 21
(s.red15<-summary(model.reduced15))

AIC.red15<-AIC(model.reduced15)

#--------------------
# MODEL 3: FULL/REDUCED (same)
#--------------------
stepwiselogit(data=train_adjusted, y="pass", exclude = NULL, include = NULL, selectio
select = "SL", sle = 0.05, sls = 0.05, goft = TRUE)
model.full5<-glm(pass~failures+schoolsup+romantic+goout, data=train_adjusted,family =
(s.full5<-summary(model.full5))

AIC.full5<-AIC(model.full5)

stepwiselogit(data=train_adjusted2, y="pass", exclude = NULL, include = NULL, selecti
select = "SL", sle = 0.05, sls = 0.05, goft = TRUE)
model.reduced5<-glm(pass~failures+schoolsup+romantic+goout,
data=train_adjusted2,family = binomial)
(s.red5<-summary(model.reduced5))

AIC.red5<-AIC(model.reduced5)

#--------------------
# MODEL 4: FULL G2
#--------------------
train_adjusted3<-dplyr::select(train,-c(G1,G3,reason,guardian,Fjob,Mjob,age))
test_adjusted3<-dplyr::select(test,-c(G1,G3,reason,guardian,Fjob,Mjob,age))

model.G2<-glm(pass~G2, data=train_adjusted3,family = binomial)
(s.G2<-summary(model.G2))

AIC.G2<-AIC(model.G2)

p_G2<-xtable(s.G2)
print(p_G2, floating=FALSE, include.rownames=T, include.colnames=T)

#--------------------
# AIC
#--------------------
Limit<-c(0.15,0.15,0.05,0.05)
AIC<-round(c(AIC.full15,AIC.red15,AIC.full5,AIC.G2),2)
# Table 24:
(tAIC<-as.data.frame(cbind(Limit,AIC)))

#------------------------------------------
# 4. EVALUATION
#------------------------------------------
```

```r
#--------------------
# Hosmer-Lemeshow Goodness of fit test
#--------------------
library(ResourceSelection)

# Model 1 (row 1 in Table 26)
pass<-as.numeric(levels(train_adjusted$pass))[train_adjusted$pass]
(hfull15 <- hoslem.test(pass, fitted(model.full15),g=16))

#--------------------
# Residual plots (Figure 9)
#--------------------
# MODEL 3: model.full5
# Pearson residuals
model.full5<-glm(pass~failures+schoolsup+romantic+goout, data=train_adjusted,family =
train_adjusted$rpifull5 <- residuals(model.full5,type="pearson")
train_adjusted$Pihatfull5<-fitted(model.full5)

# Plot Pearson residuals
Peifull5<-ggplot(train_adjusted,aes(x=Pihatfull5,y=rpifull5))+
geom_point() +
geom_smooth(method="loess",se=FALSE) +
ggtitle("Pearson␣residuals␣model␣3") +
xlab("Estimated␣probability")+
ylab("Pearson␣residuals")
Peifull5

# Studentized Pearson residuals
train_adjusted$hatfull5 <- hatvalues(model.full5)
train_adjusted$rspifull5 <- train_adjusted$rpifull5/sqrt(1-train_adjusted$hatfull5)

# Plot Studentized Pearson residuals
Prspifull5<-ggplot(train_adjusted,aes(x=Pihatfull5,y=rspifull5))+
geom_point() +
geom_smooth(method="loess",se=FALSE) +
ggtitle("Studentized␣Pearson␣residuals␣model␣3") +
xlab("Estimated␣probability")+
ylab("Studentized␣Pearson␣residuals")
Prspifull5

#--------------------
# Influential observations (Figure 10)
#--------------------
# MODEL 3: model.full5
# Cook's distance
cooksdistfull5 <- cooks.distance(model.full5)
# Add index to plot
index<-1:nrow(train_adjusted)
studfull5<-as.data.frame(cbind(index,cooksdistfull5))
stud_abovefull5<-dplyr::filter(studfull5,cooksdistfull5 > 0.06)
stud_abovefull5<-as.data.frame(stud_abovefull5)
indfull5<- stud_abovefull5$index
# Plot
PDifull5<-ggplot(studfull5,aes(x=index,y=cooksdistfull5))+
geom_line()+
geom_point(data=stud_abovefull5,pch=19, col="red")+
ggtitle("Cook's␣distances␣model␣3") +
xlab("Case␣index")+
ylab("Cook's␣distance")+
scale_y_continuous(limits = c(0, 0.14),breaks=seq(0,0.14,0.02))+
geom_label(data=stud_abovefull5,aes(x=index+0.01,y=cooksdistfull5+0.01,label=indfull5
PDifull5

#--------------------
# Error plot (Figure 11) and model 3 without influential observations (Table 31)
#--------------------
# MODEL 3 without: model.full5 without
# Remove influential observations
train_adjusted4<-train_adjusted[-c(80,234,277),]
model.full5w<-glm(pass~failures+schoolsup+romantic+goout, data=train_adjusted4,family
# Table 31
summary(model.full5w)
fitted.resultsfull5w<-predict(model.full5w,test_adjusted,type="response")

# Error plot
p<-0.1
pvec<-NULL
ACCvec<-NULL
sensitivityvec<-NULL
specificityvec<-NULL

while (p < 0.9) {
fitted.resultsfull5w<-predict(model.full5w,train_adjusted4,type="response")
tab<-table(fitted.resultsfull5w > p,train_adjusted4$pass)
sensitivity<-tab[2,2]/(tab[2,2]+tab[1,2])
specificity<-tab[1,1]/(tab[1,1]+tab[2,1])
fitted.resultsfull5wpass<-ifelse(fitted.resultsfull5w>p,1,0)
ACC<-round((mean(fitted.resultsfull5wpass == train_adjusted4$pass)),3)
print(paste("p:␣",p,"sensitivity:␣",round(sensitivity,3),"sensitivity:␣",
round(specificity,3),"ACC:␣",ACC))
sensitivityvec<-c(sensitivityvec,sensitivity)
specificityvec<-c(specificityvec,specificity)
ACCvec<-c(ACCvec,ACC)
pvec<-c(pvec,p)
p<-p+0.01
}

error<-as.data.frame(cbind(ACCvec,specificityvec,sensitivityvec))
names(error)<-c('ACCvec','specificityvec','sensitivityvec')
total<-as.data.frame(cbind(pvec,error))
x <- 0.52
y <- 0.722
coords = paste(x,y,sep=",")
data<-data.frame(x,y)

totalpoint<-dplyr::filter(total,pvec == 0.52 & ACCvec ==0.722 )
totalpoint<-as.data.frame(totalpoint)

library(ggplot2)

Errorplot<-ggplot(data=total,aes(x=pvec))+
geom_line(aes(y=ACCvec),linetype = "solid")+
geom_line(aes(y=specificityvec),linetype = "dashed")+
geom_line(aes(y=sensitivityvec),linetype = "dotted")+
geom_point(data=data,aes(x=x,y=y),pch=19, col="red")+
geom_label(data=data,aes(x=x,y=y+0.06,
label=coords),size = 3,label.size=0)+
labs(title = "ACC␣(solid␣line),␣Specificity␣(dashed␣line)␣and␣Sensitivity␣(dotted␣lin
x = "cutoff", y = "Value")+
scale_x_continuous(limit=c(0,1),breaks=seq(0,1,0.1))+
scale_y_continuous(limit=c(0,1),breaks=seq(0,1,0.2))+
theme_bw()
Errorplot

# Confusion matrix (Table 28)
p_cut<-0.52
fitted.resultsfull5w<-predict(model.full5w,test_adjusted,type="response")
(tab<-table(fitted.resultsfull5w > p_cut,test$pass))
tab
# Values for Table 29
# Model 3 without:
(sensitivity<-tab[2,2]/(tab[2,2]+tab[1,2]))
(specificity<-tab[1,1]/(tab[1,1]+tab[2,1]))
fitted.resultsfull5wpass<-ifelse(fitted.resultsfull5w>p_cut,1,0)
(ACC<-round((mean(fitted.resultsfull5wpass == test$pass)),3))
(error_rate<-1-ACC)

# ROC-curve, model 3 without (Figure 12)
ROCRpredfull5w = prediction(fitted.resultsfull5w, test$pass)
ROCRperffull5w = performance(ROCRpredfull5w, "tpr", "fpr")
plot(ROCRperffull5w)
# AUC, model 3 without (Table 27)
AUCpredfull5w <- performance(ROCRpredfull5w, measure = "auc")
(AUCpredfull5w<- AUCpredfull5w@y.values[[1]])

#------------------------------
# Final model evaluated on entire data set (without influential observations)
#------------------------------
# Remove influential observations
train_adjusted4<-train_adjusted[-c(80,234,277),]
# Add train and test set to an entire set without influential observations
studentfull_adjusted<-as.data.frame(rbind(train_adjusted4,test_adjusted))
# Fit the model
model_without_infnew <-glm(pass~failures+schoolsup+romantic+goout, data=studentfull_a
# Table 30
(s_mwinew<-summary(model_without_infnew))

#-----------------------------------------------------------------------
```