Factors affecting Student Performance

ISyE 6414 with Dr. Nicoleta Serban

4/27/15

Siddharth Shah Sudipta Mukherjee

Abstract

This study tries to find out which factors affect the academic performance of a student. Data was collected for students in two subjects – Math and Portuguese. This included directly study-related data such as hours spent studying, past failures and also data from a student's domestic life such as mother's job, parents' cohabitation status. The grades for these students in three tests for these two subjects were also collected. Using this data, we tried to find out which were the most useful predictors for grades. Using this information, teachers and schools may be able to intervene if they know students are heading towards bad grades and offer assistance in areas it is needed for the students. We found that while study-related variables such as hours spent studying and past failures are significant in predicting grades, factors such as a child having school support and how much time they spend going out also has a relationship with grades.

Table of Contents

Introduction

Education is an important element of society. Regression techniques, which allow high-level extraction of knowledge from raw data, can offer useful insights for the education domain. There are several interesting research questions in the area of school and college education that can be studied using regression techniques. Who are the students taking most credit hours? Who is likely to return for more classes? What type of courses can be offered to attract more students? What are the main reasons for student transfers? What are the factors that affect student achievement? This project will focus on the last question.

Modeling student performance is an important tool for both educators and students, since it can help get a better understanding of this phenomenon and ultimately improve it. For instance, school professionals could perform corrective measures for weak students (e.g. remedial classes). They could detect if a student was heading towards a bad grade on a class and help them get on track to a better grade. Further, there are two reasons supporting the choice of such a project: (a) there are multiple sources of data available (e.g. traditional databases, online web pages), and (b) there are diverse interest groups (e.g. students, teachers, administrators or alumni) interested in the insights offered by such a project.

A priori statement

In our project, we first expect to find specific factors that affect student performance. There could be multiple demographic, student involvement, and student grade related factors that could have an influence on student performance. While there are multiple demographic variables including age, sex, address, social background, family size, etc. we expect that only some of these variables would have a significant impact on student performance. Studies in the past have shown that the most important demographic factors that can impact student performance include gender, age, and social background. Thus, in line with these previous studies, we expect our study to show that gender, age, and social background may have a significant influence on student performance.

Some of the other factors that can also influence student performance include the reason why a student chooses to attend a particular school, number of past failures, the total amount of time that he/she utilizes for studying, whether he/she gets educational support from school and family, and Internet availability. In addition, extraneous student motivation related factors such as access to total free time after school, going out with friends, romantic relationships and alcohol consumption could also have some influence on student performance.

In addition, we expect past grades to have a significant influence on future grades. Student performance is significantly predicted by their past performance. We expect our analysis to offer similar insights regarding the prediction accuracy based on past grades. Finally, one of the most important indicators of student motivation and actual educational support received is whether a student chooses to attend school. Thus, we expect that the number of absences during a school year to have a significant influence on student performance.

Overall, we expect that we would able to elicit from the raw data factors that can affect student performance.

Problem Statement

The problems we will focus primarily on are (a) highlighting the factors significantly influencing performance in these two subjects (b) providing insights into how the two subjects differ in terms of the factors significantly influencing their performance

Data Source

The data that we will analyze consists of 649 responses obtained from two secondary education Portuguese schools for the year 2005-06. This data was collected using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics and Portuguese language. There are 33 data attributes. All but one of the independent variables are qualitative, such as students' home type (urban/rural), students' guardian (mother/father/other) etc., while only one is quantitative (number of absences). The full list of attributes is included in Appendix A. The dependent variables are the grades of the students in the three periods.

The choice of these two subjects is interesting. Mathematics requires rational "left-brain" skills while learning the Portuguese language requires creative "right-brain" skills. It is possible that factors affecting performance in these two subjects may be very different. For the math dataset we have 395 observations, while for Portuguese we have 649. So it is possible that Portuguese may give us better results because of the availability of more data.

It is important to note that while observing the data, we found that certain students had obtained a score of 0 out of a maximum possible score of 20 on G3. With the next highest score being 4 and above, it was unclear whether the students had actually taken the test or were absent on the day of the exam. We also found that all the students who scored a zero on G3 had a score of 10 or less on G1 and G2. However some students who had a score of less than 10 on G1 and G2 yet had a non-zero score for G3. This ambiguity affects our analysis, which we explore further in our report.

Exploratory Analysis

One of our first findings was that G3 was highly correlated with G1 and G2. This was one of our a priori. Below are scatterplot matrices of G1, G2 and G3 for Math and Portuguese.

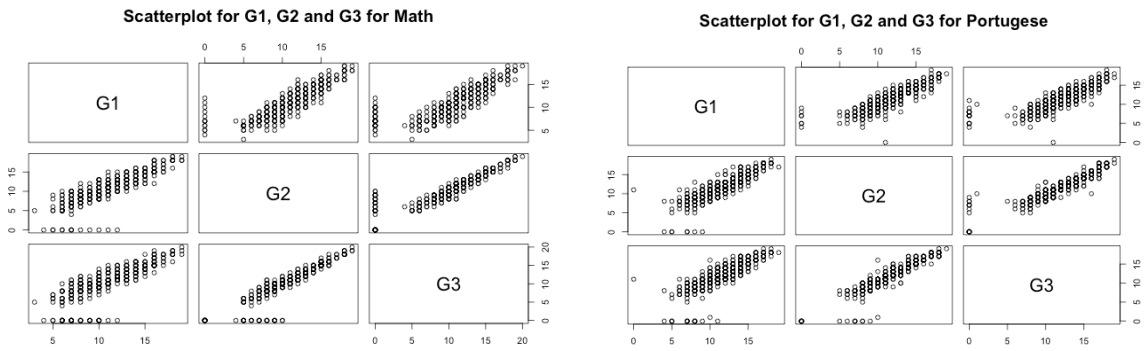


Figure 1: Scatterplots for G1, G2 and G3 for Math and Portueguese

From these scatterplots, it follows that if G1 and G2 are used as predictors, G3 becomes easier to predict. However, this analysis is much more useful is G3 is predicted without knowing G1 and G2 scores. Hence, for all our analyses, we will use G3 as the dependent variable.

We constructed box plots for all but one of the variables (since they are categorical). A scatterplot was constructed for absences. These plots can be found in Appendix B for Math and Appendix C for Portuguese. For Math, G3 seems to have a positive relationship with higher, and a negative relationship with absences and failures. Interestingly, studytime does not have a strong trend with G3, which is not what we expected a priori. For Portuguese, Medu and Fedu are two variables that show a linear trend, which was also not expected.

Models (for Math Dataset):

Full fitted model

We fitted a Poisson model for the math dataset using all the given predictors with G3 (grade in period 3) as the dependent variable. We chose to fit a Poisson model since our dependent variable G3 has values ranging from 0 to 20, which can be interpreted as count data (number of points).

The results for this full model and the coefficient estimates are noted in Appendix D. The significant variables were sex, age, studytime, failures, schoolsup, famsup, higher, romantic, goout and absences. Please refer to the list of variables in Appendix A to see what each variable means. This model gave us some interesting results. A priori we expected studytime to have an effect on G3. However, this was not the case.

Our next step was to fit a Poisson model using the significant variables found from the full model. This would be a starting point for our model selection procedures.

Model selection for predicting student performance ("Selected" Model)

Since k (number of parameters) was large, we could not use an all-subset method for finding the best model. Hence we used forward and backward stepwise model selection. We used

a model with only the significant variables found in the full model as a baseline (reduced model) and added variables with the full model as the end-point for the regression. For backward stepwise regression, we started with the full model and dropped variables with the reduced model as the end-point. Both these methods gave us the same model, which is given in Appendix E. The variables included in the model are sexM, age, address, famsizeLE3, Medu, Mjobhealth, Mjobother, Mjobservices, Mjobteracher, studytime, failures, schoolsupyes, famsupyes, higheryes, romanticyes, freetime, goout, health and absences. Of these, sexM, Medu, studytime, failures, schoolsupyes, famsupyes, higheryes, romanticyes, goout and absences were significant. This selected model gave us an AIC of 2501.

Note that certain variables that were insignificant such as famsize were also included in this selected model through forward and backward stepwise regression.

Residual Analysis and Alternate Model

Using this selected model, we performed analysis of the residuals as shown below. To check if there are any outliers, we also plotted Cook's distances on a histogram, as shown below.

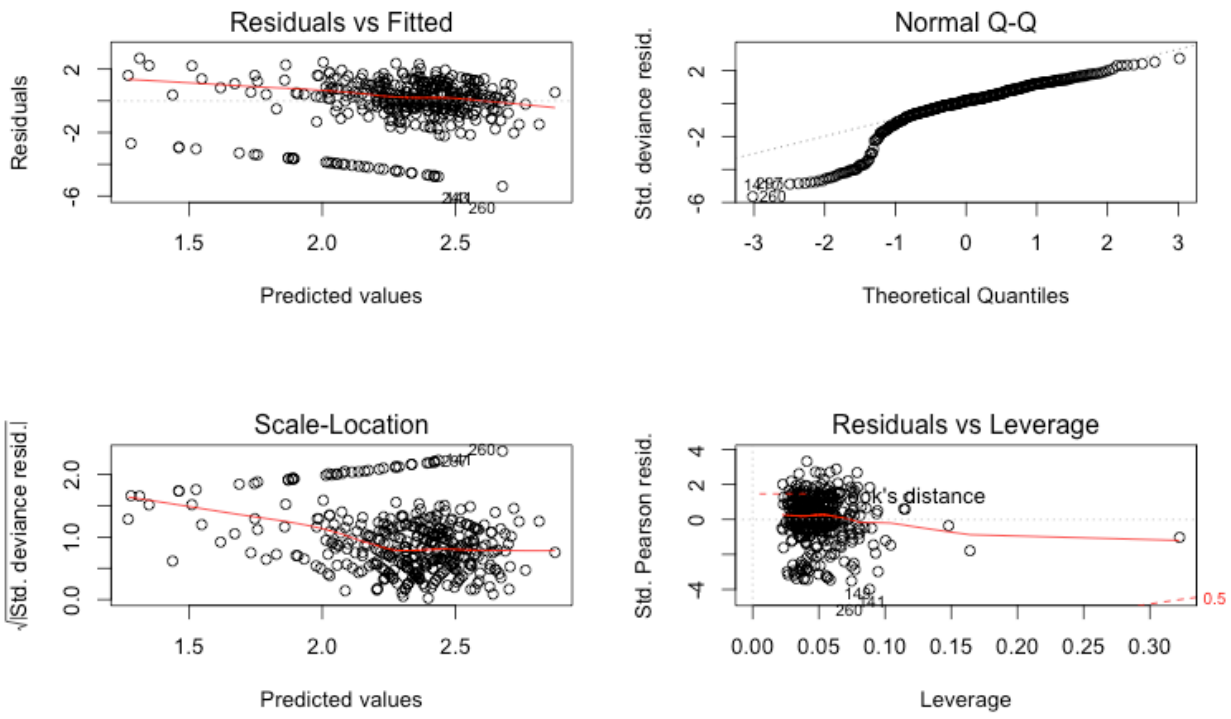


Figure 2: Analysis of Residuals for Selected Model

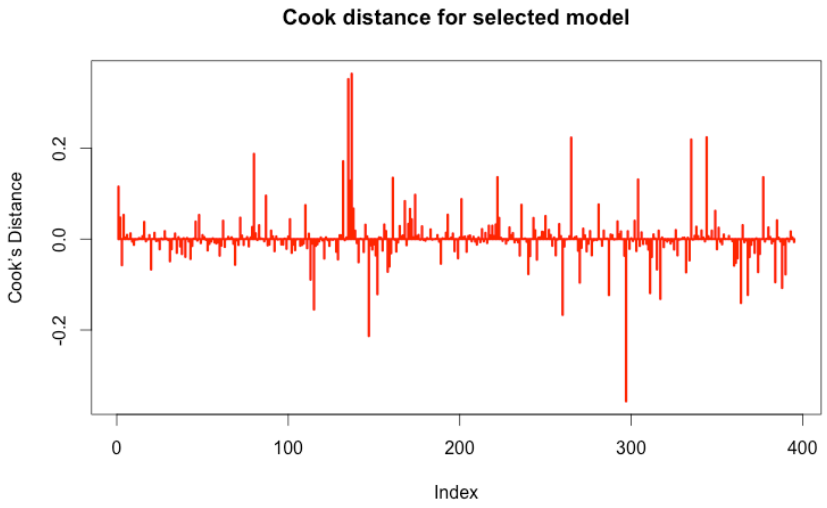


Figure 3: Histogram of Cook's distances for Selected Model

In the residuals v/s fitted plot, there is somewhat of a trend for the lower predicted values. This may be because of the zero scores for G3 that we wrote about previously. The Q-Q plot also has a skewed left tail, again because of the zeros present in G3. The Cook's distance histogram helped us identify points 135-138 and 297 as points of concern.

To solve for this issue, we removed the observations that had a score of 0 in G3 from the dataset. Using the variables we found in the model selected from model selection, we created an alternative model (Poisson) for this modified data-set. The points with high Cook's distance were also all points with zeros as G3 scores. So they are among the points removed as well. The model is in appendix (). Below is the residual analysis for this alternate model. The Q-Q plot is now a straight line and indicated normality. The residuals plot is more random as well.

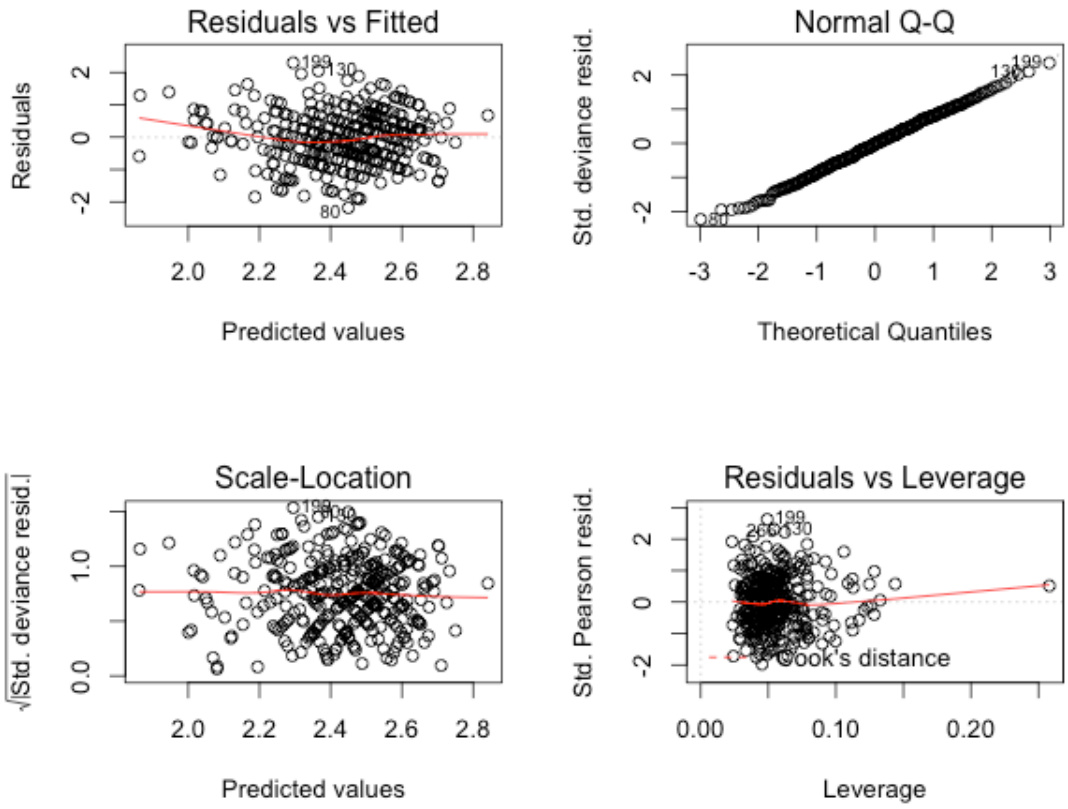


Figure 4: Residual Analysis for Alternate Model

This alternate model has a lower Residual deviance and AIC than the model selected from model selection. This model will do a good job of predicting scores of students who did not get a zero in G3, however it will not do as good a job of predicting scores of the entire set of students (including students who may get a zero in G3).

Overdispersion

In a Poisson distribution, the variance should equal the mean. However, for our selected model, we notice that the variance is greater than the mean. This phenomenon when the data shows greater variability than expected is called overdispersion. For our selected model, overdispersion is present. This may be because the variables recorded for this study and used in the model may not completely explain grades in G3. There may be other variables, such as inherent intelligence of the children that are not measured. The population of a school is made up of several children, each having their own Poisson with a different mean. At the school level, this would cause greater variability in the Poisson model.

To account for this overdispersion, we used a quasi-Poisson distribution that expresses the variance as proportional to the mean using a dispersion parameter. For our selected model in Appendix (), the Residual deviance is 941.66 with 375 degress of freedom. So the dispersion

parameter is 941.66/375 = 2.51. The standard error of the coefficients in the selected model is corrected by this factor but the estimates of the coefficients itself remain the same. This new model is called the Quasi-Poisson model, and the results are noted in appendix (). The p-values are also affected, and some of the variables that were significant in the selected model are now insignificant in the quasi-Poisson model. The variables Medu, studytime, school, familysup, and higher are now insignificant.

Portuguese Dataset

The same models "full model", "selected model" and "alternate model" were run for Portuguese dataset. Overdispersion did not exist for the selected model so a quasi-Poisson model was not run. A comparison is given in the next section. Please refer to Appendix () for full results. The variables included in the selected model for Portuguese data are sex, studytime, failures, famsup, schoolsup, higher, school, Dalc, Fedu and health.

Comparison of Selected Variables

Full Model Significance Comparison:

Variable	Math	Portuguese
schoolMS	N	Y
sexM	Y	Y
age	Y	N
studytime	Y	Y
failures	Y	Y
schoolsupyes	Y	Y
famsupyes	Y	N
higheryes	Y	Y
romanticyes	Y	N
goout	Y	N
absences	Y	N

Table 1: Comparison of full model for "Math" and "Portuguese"
Y=Significant. N=Not significant

Selected Model Variables Included Comparison:

Variable	Math	Portuguese
schoolMS	N	N
sexM	Y	Y
age	Y	N
address	Y	N
famsizeL3	Y	N
Medu	Y	N

Mjobhealth	Y	N
Mjobother	Y	N
Mjobservices	Y	N
Mjobteacher	Y	N
studytime	Y	Y
Failures	Y	Y
Schoolsupyes	Y	Y
Famsupyes	Y	Y
higheryes	Y	Y
romanticyes	Y	N
freetime	Y	N
goout	Y	N
health	Y	Y
absences	Y	N
Dalc	N	Y
Fedu	N	Y

Table 2: Comparison of variables included in selected model for "Math" and "Portuguese"
Y=Included. N=Not included

It is interesting to note that there is a differences in the variables in terms of the ability to impact student performance in math v/s Portuguese. While age, family support, romantic relationships, going out with friends, and absences have a significant impact on student performance in math, they don't affect student performance in Portuguese. This could be explained by the fact that Portuguese being a native subject, is less affected by extraneous student social life related variables such as family support, romantic relationships, and going out with friends. Similarly, the school attended has an impact on Portuguese performance but not math performance. This could be attributed to factors such as medium of instruction.

There are certain variables that impact student performance in both math and Portuguese. These are: gender, failure, study time, school support, and desire for higher education.

Findings

Our analysis reveals the following important insights about the predictors of student performance (using full models):

- **School attended**: We find that the school attended has a significant influence on student performance on Portuguese but not math. We find that by changing schools, the student's score in Portuguese could go up by 10%.
- **Sex**: We find that gender has a significant impact on both math and Portuguese scores. In line with widespread stereotype of female students being better in language subjects than male students and poorer in math than males, we find that female students are likely to receive 11% score less than male students in math and 5% score higher than male students in Portuguese.
- **Age**: We find that age has a negative impact on student score in math but not in Portuguese. For every year older, student score in math decreases by 3%.

- **Study Time**: We find that the time of study has a significant impact on both math and Portuguese. For every hour increase in the time spent in studying, the math score goes up by 5% and Portuguese score goes up by 3%.
- **Failures**: We find that failures have a significant influence on both math and Portuguese scores. For every increase in past class failure, the math score decreases by 22% and the Portuguese score decreases by 15%.
- **School support**: We find that school support also significantly influences performance in both math and Portuguese. The presence of school support increases math and Portuguese scores by 12% and 10% respectively.
- **Family support**: We find that family support also influences math performance but not Portuguese performance. The presence of family support increases math scores by 9%.
- **Higher Education**: The student's inclination to take higher education was also found to influence student performance in both math and Portuguese. Student inclination to pursue higher education increased math score by 20% and Portuguese score by 17%.
- **Romantic Relationships**: We find that romantic relationships can also significantly influence student performance in math but not in Portuguese. Being in a romantic relationship decreased math score by 10%.
- **Go out with friends**: We find that going out with friends also decreased student performance in math but not Portuguese. For every unit increase in time spent going out with friends, math score decreased by 6%.
- **Absence**: We find that absence from school significantly influenced student performance in math but not Portuguese. For every unit increase in the number of absences, the math score decreased by 0.6%.

Prior to the study, we felt that variables such as studytime, absences, failures that were more numeric in nature would be strong predictors for grades. It turned out that while these variables were significant, other qualitative variables such as the child's support system and even their sex (for Math) play a part in grades. Goout (time spent going out) is also significant. It seems like going out time is a direct substitute for study time, so students spend less time studying if they spend more time going out. In this case, better time management skills could be taught to the students

Further Study

Analysis is needed to understand why and how variables such as reason to choose school, parent's job, and alcohol consumption affect student performance. The relationships we have explored are only correlations, but causal studies need to be done to thoroughly understand the effect of each of these variables. It can be debated that there needs to some indicator of a child's inherent intelligence or "smartness". However, it can also be argued that this smartness is only a product of all these other variables that have been measured here. Either way, other qualitative variables could be added to this study. This can help in developing robust models that can be more accurate predictors of student performance.

Also, this study was based on off-line learning, since the regression techniques were applied after the data was collected. However, there is a potential to utilize online learning data

to predict student performance in real time and make the appropriate interventions required to improve student performance [1]

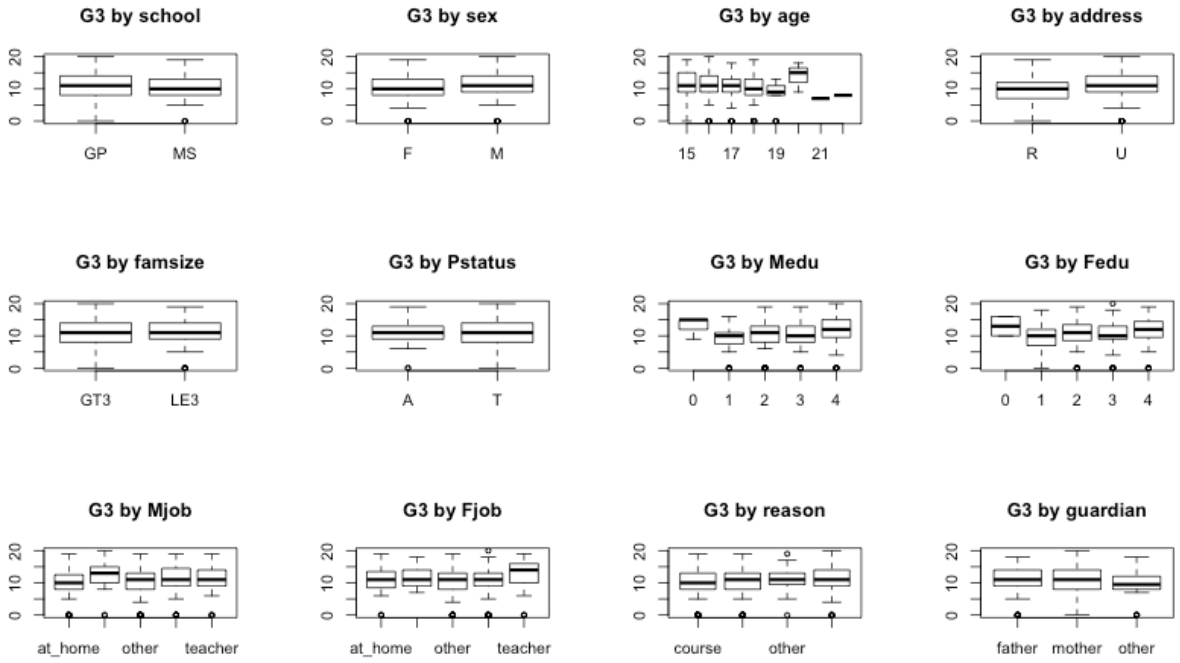
APPENDIX

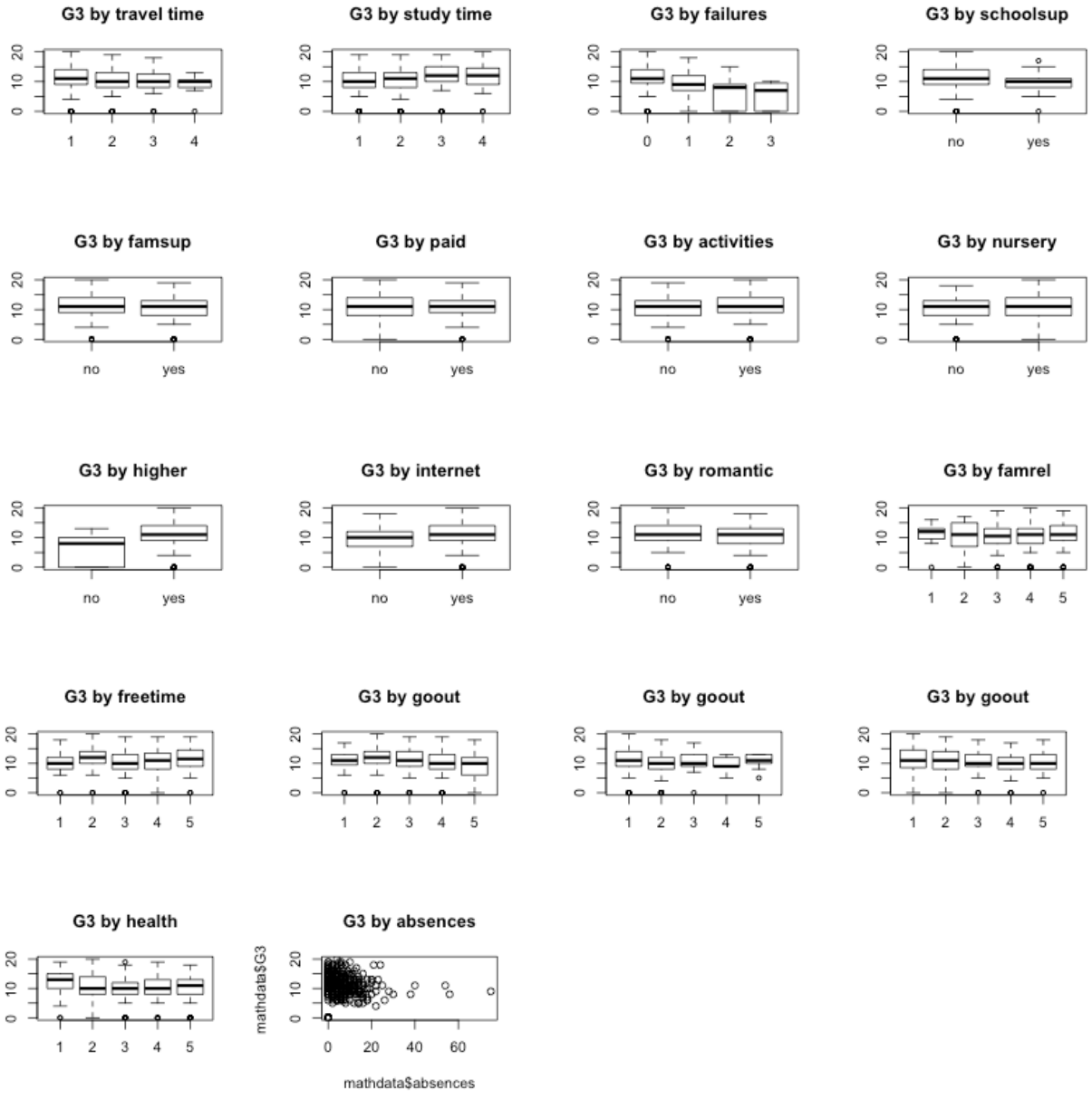
Appendix A: List of Variables

Variable	Explanation	Value
school	student's school	Binary; 1 ("GP" - Gabriel Perera) or 2 ("MS" - Mousinho da Silveira)
sex	student's sex	Binary; 1 ("F" - female) or 2 ("M" - male)
age	student's age	Numeric; 15-22
address	student's home address type	Binary; "U" (urban) or "R" (rural)
famsize	family size	Binary; "LE3" (less than or equal to 3) or "GT3" (greater than 3)
Pstatus	parent's cohabitation status	Binary; "T" (living together) or "A" (living apart)
Medu	mother's education	Numeric; 0 (none), 1 (primary - 4th grade), 2 (5th to 9th grade), 3 (secondary), 4 (higher)
Fedu	father's education	Numeric; 0 (none), 1 (primary - 4th grade), 2 (5th to 9th grade), 3 (secondary), 4 (higher)
Mjob	mother's job	Nominal; 1 (at home), 2 (healthcare related), 3 (other), 4 (civil services), 5 (teacher)
Fjob	father's job	Nominal; 1 (at home), 2 (healthcare related), 3 (other), 4 (civil services), 5 (teacher)
reason	reason to choose school	Nominal; 1 (course preference), 2 (close to home), 3 (other), 4 (reputation)
guardian	student's guardian	Nominal; 1 (father), 2 (mother), 3 (other)
traveltime	travel time to school	Numeric; 1 (<15 mins), 2 (15-30 mins), 3 (30 min-1hour), 4 (>1 hour)
studytime	weekly study time	Numeric; 1 (<2 hours), 2 (2-5 hours), 3 (5-10 hours), 4 (>10 hours)
failures	number of past class failures	Numeric: 0-3
schoolsup	extra educational support	Binary: 1 (Yes) or 2 (No)
famsup	family educational support	Binary: 1 (Yes) or 2 (No)
paid	extra paid classes	Binary: 1 (Yes) or 2 (No)
activities	extra-curricular activities	Binary: 1 (Yes) or 2 (No)
nursery	attended nursery school	Binary: 1 (Yes) or 2 (No)
higher	wants to take higher education	Binary: 1 (Yes) or 2 (No)
romantic	has a romantic relationship	Binary: 1 (Yes) or 2 (No)
famrel	quality of family	Numeric: 1 (very bad) to 5 (excellent)

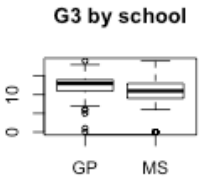
	relationships	
freetime	free time after school	Numeric: 1 (very low) to 5 (very high)
goout	going out with friends	Numeric: 1 (very low) to 5 (very high)
Dalc	workday alchohol consumption	Numeric: 1 (very low) to 5 (very high)
Walc	weekend alcohol consumption	Numeric: 1 (very low) to 5 (very high)
health	current health status	Numeric: 1 (very bad) to 5 (very good)
absences	number of school absences	Numeric: 0 to 93
G1	first period grade	Numeric: 0-20
G2	second period grade	Numeric: 0-20
G3	third period grade	Numeric: 0-20

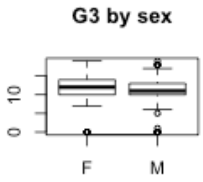
Appendix B: Box Plots of Categorical Variables for Math





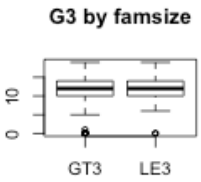
Appendix C: Box Plots of Categorical Variables for Portuguese

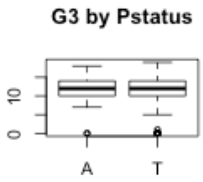

G3 by school
0
10
GP
MS

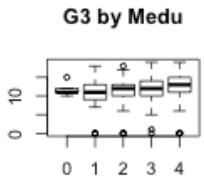

G3 by sex
0
10
F
M

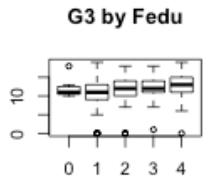

G3 by age
0
10
15
17
19
21

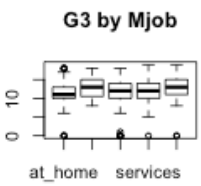

G3 by address
0
10
R
U

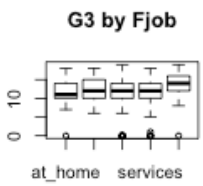

G3 by famsize
0
10
GT3
LE3

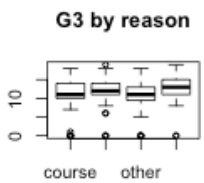

G3 by Pstatus
0
10
A
T

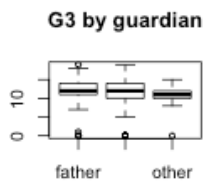

G3 by Medu
0
10
0
1
2
3
4


G3 by Fedu
0
10
0
1
2
3
4

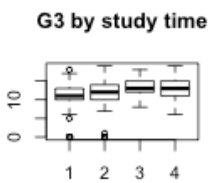

G3 by Mjob
0
10
at_home
services

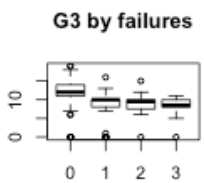

G3 by Fjob
0
10
at_home
services

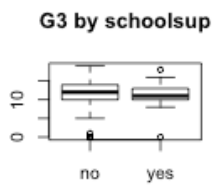

G3 by reason
0
10
course
other

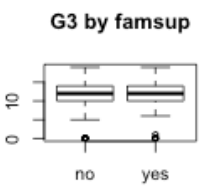

G3 by guardian
0
10
father
other


G3 by travel time
0
10
1
2
3
4


G3 by study time
0
10
1
2
3
4


G3 by failures
0
10
0
1
2
3

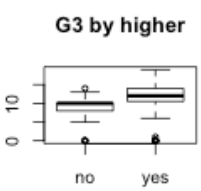

G3 by schoolsup
0
10
no
yes

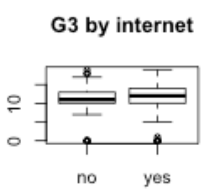

G3 by famsup
0
10
no
yes

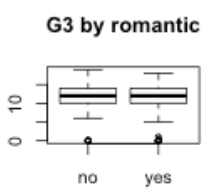

G3 by paid
0
10
no
yes

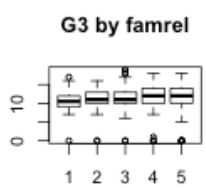

G3 by activities
0
10
no
yes

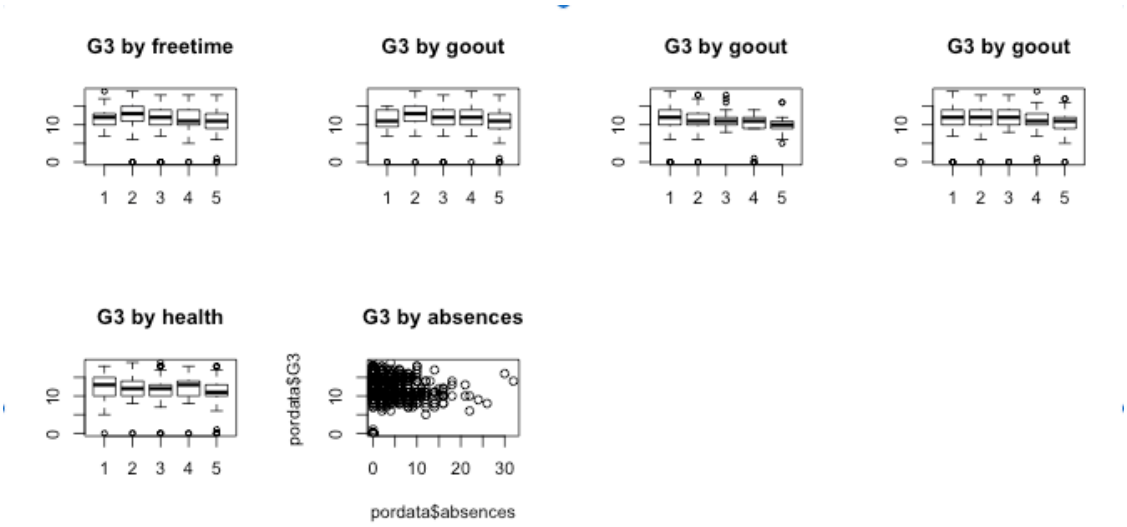

G3 by nursery
0
10
no
yes


G3 by higher
0
10
no
yes


G3 by internet
0
10
no
yes


G3 by romantic
0
10
no
yes


G3 by famrel
0
10
1
2
3
4
5



Appendix D: "Full" Model Estimates

Variable	Math		Portuguese	
	Estimate	Significance	Estimate	Significance
schoolMS	0.081723	0.179740	**-0.100681**	**0.000706*** **
sexM	**0.115098**	**0.001948** **	**-0.054827**	**0.043055***
age	**-0.035890**	**0.029815***	0.012108	0.27593
addressU	0.052338	0.247279	0.027353	0.345676
famsizeLE3	0.061290	0.090452	0.026878	0.30987
PstatusT	-0.026958	0.612921	0.011226	0.765126
Medu	0.042600	0.079349	0.001663	0.920554
Fedu	-0.008111	0.696419	0.015802	0.289658
Mjobhealth	0.089444	0.286747	0.072957	0.206724
Mjobother	-0.035508	0.533342	0.004305	0.899199
Mjobservices	0.070618	0.257325	0.033251	0.418266

Mjobteacher	-0.113275	0.153082	0.040910	0.450274
Fjobhealth	0.031155	0.771937	-0.052801	0.514005
Fjobother	0.060356	442592	-0.017513	0.728615
-				
Fjobservices	-0.043903	0.588111	-0.054036	0.307377
Fjobteacher	0.115347	0.230908	0.040343	0.57185
reasonhome	0.005817	0.891529	0.003244	0.9161
reasonother	0.074529	0.218939	-0.043619	0.295305
reasonreputation	0.053021	0.222223	0.015179	0.631954
guardianmother	-0.004933	0.903322	-0.028408	0.321528
guardianother	0.088615	0.257683	0.013515	0.819459
traveltime	-0.025305	0.347028	0.006514	0.710774
studytime	**0.052688**	**0.013155***	**0.032331**	**0.032061***
failures	**-0.224337**	**0.000000000000369*****	**-0.148940**	**0.0000000123*****
schoolsupyes	**-0.129216**	**0.013343***	**-0.109094**	**0.006951****
famsupyes	**-0.088907**	**0.013671***	-0.003441	0.889935
paidyes	0.035369	319721	-0.024491	0.635008
activitiesyes	-0.036272	0.282193	0.015655	0.519088
nurseryyes	-0.010361	0.80607	-0.018370	0.53699
higheryes	**0.195120**	**0.041927***	**0.173776**	**0.000194*****
internetyes	0.045241	0.354372	0.021649	0.481718
romanticyes	**-0.106043**	**0.00355****	-0.035293	0.159472
famrel	0.019952	0.285154	0.012315	0.340094
freetime	0.030301	0.088913	-0.011816	0.336605

goout	**-0.058131**	**0.000673*** **	-0.004999	0.670194
Dalc	-0.020912	0.416117	-0.018388	0.292362
Walc	0.023945	0.210566	-0.007880	0.54463
health	-0.017740	0.141821	-0.014370	0.085805
absences	**0.006355**	**0.003158** **	-0.002727	0.32047

Math full model deviances:

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1159.13  on 394  degrees of freedom
Residual deviance:  921.44  on 355  degrees of freedom
AIC: 2520.9

Number of Fisher Scoring iterations: 5
```

The drop in deviance is 237.69, distributed as a chi-square with 29 degrees of freedom. We fail to reject the null hypothesis that all predictors can be set to zero. But the full model does still not explain a high amount of null deviance. The deviance goodness of fit test of the model is rejected. Overdispersion factor of the model is 921.44/355 = 2.6

Portuguese full model deviances:

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 745.61  on 648  degrees of freedom
Residual deviance: 529.31  on 609  degrees of freedom
AIC: 3352

Number of Fisher Scoring iterations: 4
```

The drop in deviance is 216.3, distributed as a chi-square with 29 degrees of freedom. We fail to reject the null hypothesis that all predictors can be set to zero. But the full model does still not explain a high amount of null deviance. The deviance goodness of fit test of the model is rejected.

Appendix E: "Selected Model" Estimates for Math

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.392698   0.284866   8.399  < 2e-16 ***
sexM          0.105819   0.035194   3.007 0.002641 **
age          -0.020836   0.014325  -1.455 0.145801
addressU      0.061846   0.040275   1.536 0.124639
famsizeLE3    0.056581   0.034856   1.623 0.104535
Medu          0.045356   0.019996   2.268 0.023312 *
Mjobhealth    0.125641   0.077310   1.625 0.104128
Mjobother    -0.030870   0.053696  -0.575 0.565363
Mjobservices  0.091340   0.058739   1.555 0.119943
Mjobteacher  -0.084272   0.074680  -1.128 0.259133
studytime     0.049799   0.019632   2.537 0.011191 *
failures     -0.221305   0.029106  -7.603 2.89e-14 ***
schoolsupyes -0.124601   0.051391  -2.425 0.015326 *
famsupyes    -0.079053   0.033688  -2.347 0.018945 *
higheryes     0.200786   0.093024   2.158 0.030895 *
romanticyes  -0.099238   0.035503  -2.795 0.005187 **
freetime      0.032947   0.016934   1.946 0.051709 .
goout        -0.056412   0.015216  -3.707 0.000209 ***
health       -0.019408   0.011462  -1.693 0.090426 .
absences      0.006336   0.002011   3.150 0.001632 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1159.13  on 394  degrees of freedom
Residual deviance:  941.66  on 375  degrees of freedom
AIC: 2501.1

Number of Fisher Scoring iterations: 5
```

Significant Variables: Intercept, sexM, Medu, studytime, failures, schoolsupyes, famsupyes, higheryes, romanticyes, freetime, goout, absences.

Appendix F: "Alternate" Model for Math – using nonzero G3 data

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.665700   0.284623   9.366  < 2e-16 ***
sexM         0.050023   0.035012   1.429 0.153085
age         -0.012798   0.014375  -0.890 0.373328
addressU     0.053555   0.040131   1.334 0.182046
famsizeLE3   0.017793   0.034652   0.513 0.607622
Medu         0.031918   0.019840   1.609 0.107671
Mjobhealth   0.079281   0.078596   1.009 0.313109
Mjobother   -0.038020   0.054907  -0.692 0.488657
Mjobservices 0.079484   0.059986   1.325 0.185154
Mjobteacher -0.051331   0.075912  -0.676 0.498921
studytime    0.042854   0.019969   2.146 0.031872 *
failures    -0.097309   0.029228  -3.329 0.000871 ***
schoolsupyes -0.195629   0.051592  -3.792 0.000150 ***
famsupyes   -0.061657   0.033897  -1.819 0.068915 .
higheryes    0.027616   0.094541   0.292 0.770205
romanticyes -0.006201   0.035580  -0.174 0.861642
freetime     0.012481   0.016943   0.737 0.461349
goout       -0.046480   0.015471  -3.004 0.002662 **
health      -0.020272   0.011476  -1.766 0.077319 .
absences    -0.005164   0.002254  -2.291 0.021975 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 326.87  on 356  degrees of freedom
Residual deviance: 230.08  on 337  degrees of freedom
AIC: 1789.5
```

Appendix F: "**Quasi-Poisson" Model Estimates for Math**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.392698   0.376316   6.358 5.94e-10 ***
sexM         0.105819   0.046493   2.276  0.02341 *
age         -0.020836   0.018924  -1.101  0.27158
addressU     0.061846   0.053204   1.162  0.24580
famsizeLE3   0.056581   0.046046   1.229  0.21993
Medu         0.045356   0.026415   1.717  0.08679 .
Mjobhealth   0.125641   0.102129   1.230  0.21939
Mjobother   -0.030870   0.070934  -0.435  0.66368
Mjobservices 0.091340   0.077596   1.177  0.23989
Mjobteacher -0.084272   0.098654  -0.854  0.39353
studytime    0.049799   0.025934   1.920  0.05559 .
failures    -0.221305   0.038450  -5.756 1.80e-08 ***
schoolsupyes -0.124601   0.067889  -1.835  0.06724 .
famsupyes   -0.079053   0.044503  -1.776  0.07649 .
higheryes    0.200786   0.122888   1.634  0.10312
romanticyes -0.099238   0.046901  -2.116  0.03501 *
freetime     0.032947   0.022371   1.473  0.14166
goout       -0.056412   0.020100  -2.807  0.00527 **
health      -0.019408   0.015142  -1.282  0.20074
absences     0.006336   0.002657   2.385  0.01759 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.745121)

    Null deviance: 1159.13  on 394  degrees of freedom
Residual deviance:  941.66  on 375  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Significant Variables: Intercept, sexM, failures, **romanticyes, goout, absences.**

Appendix G – "Selected" and "Alternate" Model for Portuguese Data

Selected Model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.379158   0.066641  35.701  < 2e-16 ***
sexM        -0.043323   0.025282  -1.714  0.08661 .
studytime    0.037510   0.014348   2.614  0.00894 **
failures    -0.149481   0.024431  -6.118 9.45e-10 ***
famsupyes   -0.001966   0.024169  -0.081  0.93516
schoolsupyes -0.115869   0.038715  -2.993  0.00276 **
higheryes    0.183104   0.045185   4.052 5.07e-05 ***
schoolMS    -0.122052   0.025808  -4.729 2.25e-06 ***
Dalc        -0.032790   0.013753  -2.384  0.01712 *
Fedu         0.024343   0.010850   2.244  0.02486 *
health      -0.014687   0.007945  -1.848  0.06454 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 745.61  on 648  degrees of freedom
Residual deviance: 550.41  on 638  degrees of freedom
AIC: 3315.1

Number of Fisher Scoring iterations: 4
```

Significant Variables: studytime, failures, schoolsupyes, higheryes, schoolMS, Dalc, Fedu

Alternate Model (Non-zero G3’s) for Portuguese:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.396542   0.066662  35.951  < 2e-16 ***
sexM         -0.036941   0.025314  -1.459 0.144482
studytime     0.034845   0.014263   2.443 0.014563 *
failures     -0.119652   0.023998  -4.986 6.17e-07 ***
famsupyes    -0.024655   0.024085  -1.024 0.306003
schoolsupyes -0.107690   0.038720  -2.781 0.005415 **
higheryes     0.166363   0.044873   3.707 0.000209 ***
schoolMS     -0.073240   0.025737  -2.846 0.004432 **
Dalc         -0.027627   0.013777  -2.005 0.044923 *
Fedu          0.021309   0.010824   1.969 0.048984 *
health       -0.011059   0.007946  -1.392 0.164017
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 384.23  on 633  degrees of freedom
Residual deviance: 259.61  on 623  degrees of freedom
AIC: 3024.3

Number of Fisher Scoring iterations: 4
```

Significant Variables = studyime, failures, schoolsupyes, higheryes, schoolMS, Dalc. Fedu

References

1. P. Cortez and A. Silva. 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.
   [Web Link]
2. Wedderburn, R.W.M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method". *Biometrika* **61** (3): 439 447. doi:10.1093/biomet/61.3.439. MR 0375592.
3. Dorina, Kabakchieva. . Predicting Student Performance by Using Data Mining Methods for Classification. Cybernetics and Information Technologies. Volume 13, Issue 1, Pages 61–72, ISSN (Online) 1314-4081, ISSN (Print) 1311-9702, DOI: 10.2478/cait-2013-0006, March 2013
4. Ramaswami, M., R. Bhaskaran. A CHAID Based Performance Prediction Model in Educational Data Mining. - IJCSI International Journal of Computer Science Issues, Vol. **7**, January 2010, Issue 1, No 1, 10-18.