Effect of Student Characteristics on Math Performance in Portuguese Schools

BIOSTAT 653 Course Project

Ina Conrado Mandy Wong Ryan Ross Kevin Liao

January 8, 2022

1 Introduction

1.1 Background

Educational performance is often used as a predictor of long-term economic success for students. In 2006, Portugal performed worse than the rest of the Europe in measures of educational success. Drop-out and failure rates in Portugal are comparably quite high among secondary school students. (Eurostat, 2007). Compared to the European Union average of 15 percent, Portugal reported 40 percent drop out of secondary school students in 2006 (Cheng, 2017). According to Cortez and Silva (2006), student performance in the core classes of Mathematics and Portuguese are key in determining student achievement since they provide the fundamentals necessary in further subjects such as physics and history.

The dataset comes from a study done by Cortez and Silva (2008), where they looked at student achievement in secondary education of two Portuguese schools in the year 2005-2006. The two schools are Escola Secundaria Mouzinho da Silveria (MS) and Escola Secundaria Gabriel Pereira (GP), which are both located in the Alentejo region of Portugal. The data attributes include 395 students' grades, demographic, social, and school related features. It was collected by using school reports and questionnaires. Two datasets are available from the study regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In Cortez and Silva (2008), their analysis focused on prediction of future education scores utilizing machine learning methods. Educational success in Mathematics and Portuguese was modeled in three ways: binary classification of performance using pass/fall, five-level classification with I as very good performance and V as insufficient performance, and regression with the numeric grade score as the outcome. They tested four algorithms, decision trees, random forest, neural networks, and support vector machines, in order to determine which model had the best predictive accuracy.

For the purpose of our project, we were primarily interested in the Mathematics subject score trend. Our study differs from the previously published paper by Cortez and Silva because we are interested in the student characteristics affecting their score trend, while their anlaysis focused on developing models to predict future grades. Therefore, our analysis focused on implementing statistical methods for longitudinal data.

1.2 Study Objectives

The scientific question that we are addressing is whether the change in grade period score in Mathematics over the course of secondary school differs between the following covariate values: 1) School Attended (MS vs GP) 2) Gender 3) Romantic Partner Status 4) Health Status 5) Absences 6) Parental Cohabitation. In addition, we also included and adjusted for student baseline characteristics. Ultimately, We are interested in determining the effect of our covariates on the changes in the educational success of students over time.

2 Methods

2.1 Model Specification

Our outcome of interest was grade score in mathematics. This variable is continuous, so we modeled it using a linear mixed effects model. The model was fit to account for the correlation within subjects between our outcomes: the three grades for each of the 395 students that were measured throughout their secondary education. All other covariates of interest were captured at baseline and only measured at the beginning of the study. Students and grade were treated as random effects, while the remainder of the covariates are fixed effects in model.

Since we fit a mixed effects model with both random intercepts and slopes, we did not specify a covariance structure for each subjects responses. Rather we assume that conditional on a subjects random effects, their outcomes have constant variance and are assumed uncorrelated. Because mixed effects model condition on each student's initial grade, a correlation structure was effectively induced through our model.

2.2 Testing for Random Effects

In order to determine whether we needed to include random effects in our model, we first fixed our base model as specified below and fit the following three models.

1. A model with fixed effects only
2. A model with fixed effects and a random intercept
3. A model with fixed effects and both a random intercept and random slope

To formally test which random effects to include in our model, we used a likelihood ratio test, because our three models are nested. To determine our p value for each test, we used the 50-50 mixture of χ^2_1, χ^2_2 distributions as the null distribution.

2.3 Variable Selection

To eliminate unnecessary covariates from our model, we performed variable selection for our mean model. The reason we decided to perform variable selection is to find the most parsimonious model. By removing redundant parameters, we avoid any potential multi-collinearity and ensure degrees of freedom are not wasted. To compare each of our models, Maximum likelihood (ML) estimation was used instead of restricted maximum likelihood (REML) because we are comparing mean models. Backward variable selection using AIC for model comparison was performed. We begin with our

full model as specified below. At each step of the analysis, a variable was eliminated from the model and AIC was used as the criteria to determine whether or not removing the variable improved model fit. The variables in the base model were always kept and selection was done solely on the interaction variables. The *step* function in the package *lmerTest* was used for variable selection. We also verified our results to the package *MuMIn*, which shows the AIC of all specified models. The covariates of interest included in the model are described in more detail in table 1.

Full Model

$$\begin{aligned} GScore_{ij} = {} & b_{0i} + b_{1i}Grade_{ij} + \beta_{0i} + \beta_1 Grade_{ij} + \beta_2 School_i + \beta_3 RomPartner_i + \beta_4 Health_i + \beta_5 Absences_i \\ & + \beta_6 ParCohab_i + \beta_7 Gender_i + \beta_8 School_i * Grade_{ij} + \beta_9 Gender_i * Grade_{ij} \\ & + \beta_{10} Grade_{ij} * RomPartner_i + \beta_{11} Grade_{ij} * Health_i + \beta_{12} Grade_{ij} * Absences_i \\ & + \beta_{13} Grade_{ij} * ParCohab_i + \epsilon_{ij} \end{aligned} \tag{1}$$

$i = 1, 2, ...357$, $j = 1, 2, 3$

Base Model

$$\begin{aligned} GScore_{ij} = {} & b_{0i} + b_{1i}Grade_{ij} + \beta_{0i} + \beta_1 Grade_{ij} + \beta_2 School_i + \beta_3 RomPartner_i + \beta_4 Health_i + \beta_5 Absences_i \\ & + \beta_6 ParCohab_i + \beta_7 Gender_i + \epsilon_{ij} \end{aligned} \tag{2}$$

$i = 1, 2, ...357$, $j = 1, 2, 3$

Where $b_i \sim MVN_2(0, G)$, $e_i \sim MVN_3(0, R_i)$ and $R_i = \sigma_\epsilon^2 I_3$

Table 1: Covariates of Interest

Type	Variable Name	Variable Type	Description
Response	G Score	Numeric	Secondary grade score
Covariate	Sex	Binary	(female/male)
	School	Binary	School attended (MS/GP)
	Pstatus	Binary	Parents cohabitation status (living together/apart)
	Romantic	Binary	With a romantic relationship (yes/no)
	Health	Numeric	Current Health Status
	Absences	Numeric	Number of school absences

Table 1: Covariates of interest and description of data

2.4 Data Analysis

Statistical analyses were carried out using R version 3.5.1. The package *mgcv* was used to fit our model.

3 Results

3.1 Exploratory Analysis

Descriptive statistics of the covariates of interest were obtained (Table 1). Thirty-eight students were excluded from our data-set since they did not report a grade from all three grading periods. Therefore, our sample size used for our analysis was $n = 319$.

Table 2. Descriptive Statistics		
	N = 357*	**Percent**
School		
Gabriel Pereira	315	88.24
Mousinho da Silveira	42	11.76
Gender		
Female	185	51.82
Male	172	48.18
Parent's Cohabitation Status		
Living Apart	39	10.92
Living Together	318	89.08
Health Status		
Very Bad	45	12.61
Bad	38	10.64
Neutral	83	23.25
Good	58	16.25
Very Good	133	37.25
Romantic Relationship Status		
Not In a Romantic Relationship	245	68.63
In a Romantic Relationship	112	31.37
	Mean	**Std**
Absences	6.32	8.19
G1 Score	11.27	3.24
G2 Score	11.36	3.15
G3 Score	11.52	3.23

* 38 students were exluded from our dataset since they did not report a grade from all 3 grading period.

Table 2: Descriptive Statistics for the Covariates of Interest

Students math scores over their grades in secondary education were plotted using a spaghetti plot (Figure 1). The mean trajectory line for all students, shown in orange, indicates a shallow increase in mean grade score over time. However, looking by student we see high correlations within individual students grade. In addition, there appears to be quite a bit of between-subject heterogeneity as initial grades and how they change over time can very drastically.

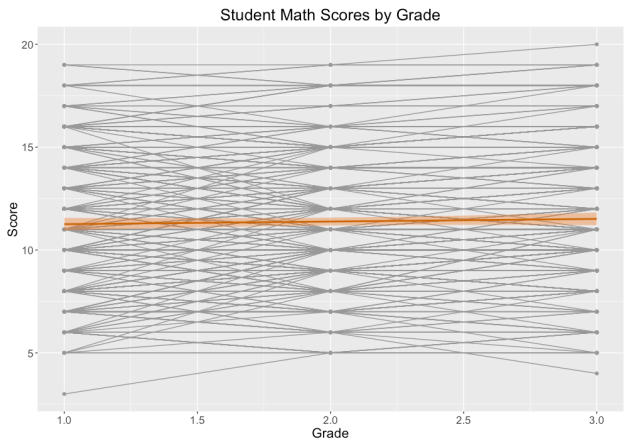

Figure 1: Spaghetti plots visualizing the 20 randomly sampled individuals' response over time

3.2 Model Specification

Figure 1 shows that there was both within subject correlation and between subject heterogeneity in our dataset which indicated a linear mixed model was appropriate. Since a mixed effects model with random intercepts and slopes was fit, a covariance structure was induced and did not require direct specification.

The model with both random slope and intercept performed better than the model with no random effects and the model with only a random intercept (Table 3). To compare between nested models, we performed likelihood ratio tests. Testing models 1 and 2 at an $\alpha = 0.05$ significance level, we reject the null hypothesis and conclude that having a random intercept improves our model fit ($p < 0.001$). Similarly, testing models 2 and 3 we also conclude that having both a random intercept and random slope further improves our model fit ($p < 0.001$).

Table 3: Evaluation of Need for Random Effects						
Model	AIC	BIC	LogLik	Test	Likelihood Ratio	P-value
1. No Random Effects	5453.0	5453.0	-2715.5			
2. Random Intercept Only	4108.2	4167.9	-2042.1	1 vs. 2	1346.8 (1 vs. 2)	<0.001
3. Random Intercept and Slope	4063.5	4138.1	-2016.7	2 vs. 3	50.7 (2 vs. 3)	<0.001

Table 3: Evaluation of need for Random Effects through comparison of three models

3.3 Variable Selection

Variable selection kept the following variables: grade, school, romantic partner, health, absences, parental cohabitation, gender, interaction between grade and school, and the interaction between grade and absences. The AIC of the final model chosen was 4040.3.

Final Model

$$GScore_{ij} = b_{0i} + \beta_{0i} + b_{1i}Grade_{ij} + \beta_1 Grade_{ij} + \beta_2 Schooli + \beta_3 RomPartner_i + \beta_4 Health_i + \beta_5 Absences_i + \beta_6 ParCohab_i + \beta_7 Gender_i + \beta_8 Grade_{ij} RomPartner_i + \beta_9 Grade_{ij} Absences_i + \epsilon_{ij} \quad (3)$$

$i = 1, 2, ...395$, $j = 1, 2, 3$

3.4 Data Analysis

The effect of each of the variables in the final model are reported in table 4. Two student characteristics were found to affect trends in grades score: presence of a school and number of absences.

Table 4: Fixed Effects Model Results					
Effect	Value	Std.Error	DF	t-value	p-value
(Intercept)	12.11	0.68	711	17.80	0.0000
Grade	0.29	0.05	711	5.65	0.0000
School: MS	-0.14	0.56	350	-0.25	0.8006
sex: Male	0.65	0.33	350	1.99	0.0472
Romantic: yes	0.15	0.36	350	0.43	0.6667
Health	-0.22	0.12	350	-1.88	0.0610
Absences	-0.03	0.02	350	-1.54	0.1244
Pstatus: Together	-0.37	0.52	350	-0.71	0.4780
Grade:schoolMS	-0.355102	0.12	711	-2.94	0.0034
XGrade:absences	-0.02	0.005	-4.12	711	0.0000

Table 4: Fixed Effects Model Results

Figure 2 shows predicted mathematics score for different schools, showing the population trend within each school, as well as trends for 20 selected subjects within each school. This plot shows that the overall trends in scores are different between schools, and that students in GP have a better outlook than students in MS, even adjusting for the other covariates. Because of the imbalance in the number of students within each school, we also looked at a model adding school as a another hierarchical cluster of random intercepts, allowing for different distributional assumptions for students within each group. We found that this model increased both AIC and BIC, and the residual and "qqnorm" plots did not indicate any evidence that the variances are different. Thus, we assumed the structure is the same for students at each school.

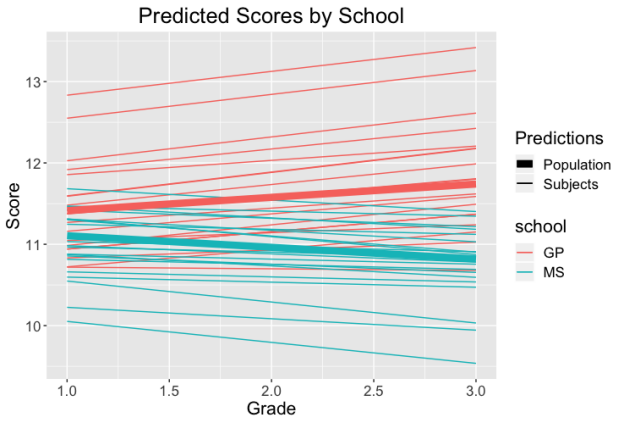

Figure 2: Predicted Scores by School

Figure 3 shows predicted mathematics score for different numbers of absences, showing the

population trend for each count, as well as trends for 5 selected subjects with that number of absences. This plot shows how the overall trends in scores are different as the number of absences increases.

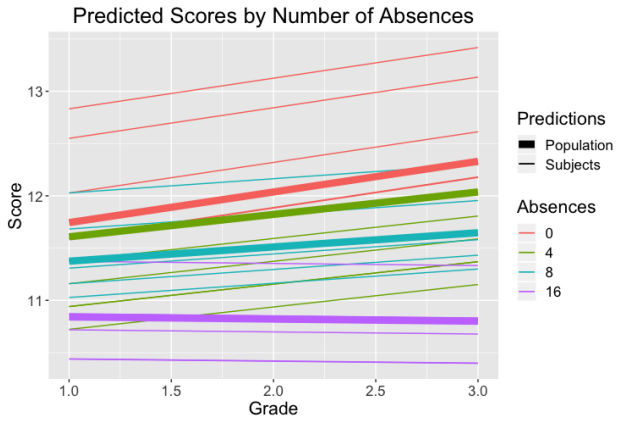

Figure 3: Predicted Scores by Absences

4 Discussion and Conclusion

While student performance in Portugal secondary schools have improved over time, their failure and drop out rates are still high compared to the rest of Europe (Cortez and Silva, 2008). Our results show that an increase in number of absences is associated with decreased mathematics score. The Ministry of Education in Portugal can work with parents and school administrators to come up with strategies to decrease number of absences, such as creating incentives to promote school attendance or developing outreach programs to teach about the importance of education.

Academic performance differences are evident between the two public secondary schools evaluated using this dataset. This difference between school performance is not unexpected, because higher mathematics score in Gabriel Pereria compared to Mousinho da Silveria has been reported in a related analysis (Cheng, 2017). Cheng (2017) reported an urban/rural difference between the two schools. Gabriel Pereria has students who come from mostly urban settings and are of higher socio-economic status compared to Mousinho da Silveria (Cheng,2017).

A study looking at student performance in mathematics using a 2009 dataset found that the school attended was the most predictive of mathematics score when looking at schools of differing size and quality (Faria and Portela, 2016). The Ministry of Education should focus educational outreach programs towards the schools where the need is greatest. There appears to be differing quality in the education provided by school, so efforts should be made to improve educational quality at schools which perform consistently lower than what is expected.

There are a few limitations of our analysis. For each student only three mathematics scores are available, one for each year of secondary school. As a result, we are only capturing end of year performance for each student. In addition, all variables were only measured at the final year and some of the predictors such as, relationship status, parental cohabitation, absences, and health status

may vary throughout schooling. Therefore, our analysis was a pseudo-longitudinal study because grade scores and student characteristics were only recorded in the final year. As a result, potential variation in time variant predictors were not accounted for.

There are several possible extensions to our study. Mathematics score could be measured multiple times per year for each student. In addition, measuring relationship status, parental cohabitation, absences, and health status each time mathematics score is measured would add additional complexity to our model and may better identify factors that affect mathematics performance over time. A possible extension to the current study would be including variables such as alcohol and drug use, participation in extracurricular activities, and more to understand their effects on student performance. In addition, if mathematics score is measured more than three times per student this may warrant the use of splines and smooths in our model.

Overall, our project showed that number of absences and school attended has an effect on academic performance in mathematics over time for Portuguese students in the two schools studied. Our analysis is an important first step in understanding student performance and can be used to advise policy makers interested in improving educational outcomes in Portugal.

References

1. Burnham, K. P. and Anderson, D. R (2002) Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York, Springer-Verlag.
2. Cheng, L. (2017) Exploring the Factors that Affect Secondary Student's Mathematics and Portuguese Performance in Portugal. Masters dissertation Dublin Institute of Technology, 2017. doi:10.21427/D7P33K
3. Cortez, Paulo, and Silva, Alice. USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE. 15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008, Apr. 2008, www3.dsi.uminho.pt/pcortez/student.pdf.
4. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
5. Eurostat. Early school-leavers. 2007, http://epp.eurostat.ec.europa.edu
6. Faria, S., Portela, M. C. (2016). Student Performance in Mathematics using PISA-2009 data for Portugal. Working Papers De Gestão (Management Working Papers). Retrieved from http://cemapre.iseg.ulisboa.pt/educonf/4e3/files/Papers/Faria.pdf
7. Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, *82*(13), 1-26. doi: 10.18637/jss.v082.i13 (URL: http://doi.org/10.18637/jss.v082.i13).
8. Wood, S.N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.