0.1 Executive Summary

Within the European Union, Portugal has the highest high school abandonment rate of 41%, compared to the EU average of 14%. Furthermore, although Portugal's students' test scores have improved over the last decades, the average academic performance is still lower than OECD countries' average. What could be the factors that are limiting students' performances? Could programs be implemented that improve grades? In this analysis, the data—collected by Paulo Cortez and Alice Silva in 2006—contains math scores from two secondary schools: Galriel Pereora and Mousiho da Silveira. Our analysis of the data concludes that although there is no significant relationship between alcohol use (both daily and weekend) and academic performance, there are significant factors that can model both number of failures, and first period and final grades. With these findings in mind, schools can create programs that focus on the significant factors that affect grades and possibly improve academic performance.

0.2 Hypothesis

We propose there are factors affecting grades and failures that can be implemented into social programs which can create improvements in academic performance. In this analysis, we are answering three questions: 1) Does alcohol consumption affect grades, 2) what factors affect class failure rates, and 3) what factors affect first period and final grades?

0.3 Data and EDA

The dataset used—which was downloaded from Kaggle—contains 395 students, their number of class failures along with first, second, and final math grades, and students' respective information like age, sex, address (urban or rural), and family size; other variables included are: mother's and father's education level and job, study time, school and family extra support, desire for higher education, free time, go out frequency, alcohol use (daily and weekend), health, and absences.

A look at the variables' correlation matrix (Fig. 1) reveals some unsurprising correlations among variables. Some strong positive correlations include among G1, G2, and G3 grades, between weekday and weekend alcohol consumption (Dalc, Walc) along with going out frequency, and the education level between parents (Medu, Fedu). Failures have a moderate level of negative correlation with grades along a weaker negative correlation with Medu and Fedu.

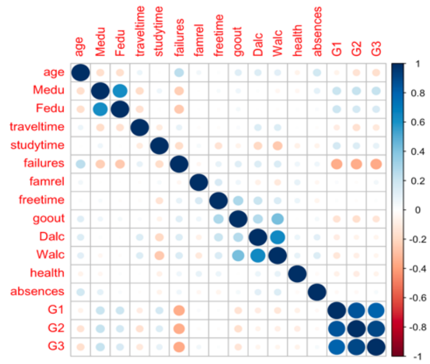

Figure 1: Correlation matrix of variables in the dataset

A histogram (Fig. 2) of 'failures' shows the majority (312) of students have never failed a class; 50 students have failed one class, 17 students have failed 2 classes, and 16 students have failed 3 classes. With an excess of 0 failures, a zero-inflated Poisson model may be used to model failures.

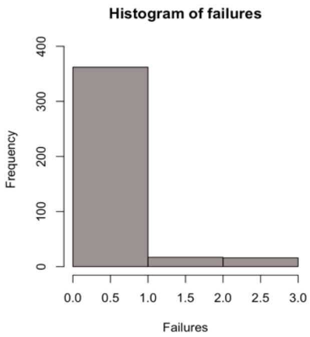

Figure 2: Histogram of failures

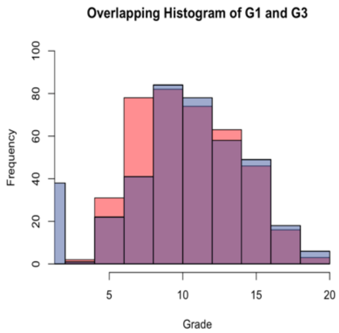

Figure 3: Overlapping histogram of first period grades and final grades. Red = G1, blue = G3.

An overlapping histogram (Fig. 3) of the first period and final grades shows that both grades are relatively normal, however there are 38 students who received a zero for their final grade. The mean grades are 10.91 and 10.42 for G1 and G3 respectively, while removing the zeros raises the mean G3 grade to 11.52.

0.4 Models

Because there are an excess number of zero failures, number of failures was modeled using a zero-inflated Poisson model. The final model for failures is:

$$log(\lambda) = 10.91 - 0.547(Age) - 0.486(Study\ time) - 0.382(Health) + 0.092(Absences)$$

This means that holding other variables constant, the mean number of failures changes by a factor of $e^{-0.547}$ = 0.58, or decreases by 42% for every additional year in age—similarly, a 38% decrease in failures for every category increase in study time, a 47% decrease for every category increase in health, and a 9.6% increase for every additional absence. Performing a Vuong test (null that the models are indistinguishable) between the zero-inflated Poisson model and a regular Poisson model of the same variables gives a p-value of 0.0006, indicating that the zero-inflated model is superior to the standard Poisson model.

The initial models for G1 and G3 included all variables and subsequent models were created using backward-elimination stepwise regression of p-values; both of the final models included variables that were all significant while delivering the best BIC values.

The final model for first period grades (G1) is:

G1 = 10.07 + 0.88(Sex if M) + 0.58(Study time) – 1.32(Failures) – 1.98(School support) – 0.63(Family support) + 1.5(Higher) – 0.35(Go out)

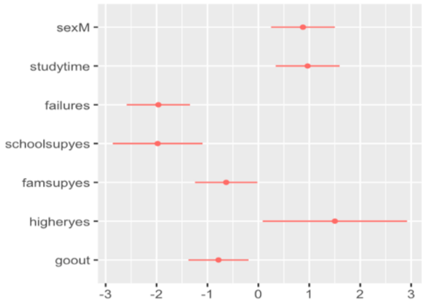

Figure 4: Whisker plot of G1 model coefficients and 95% CI

The coefficients (Fig. 4) show visually that some factors are quite close to 0, or having no effect. The factor with the biggest positive effect to first period grades is if a student has plans and aspirations for higher education while a surprising finding is that increased family and school support could decrease G1 grades by 3 points on average together; this is a huge impact as the highest grade possible is 20. Two other variables that lower grades unsurprisingly are going-out frequency and the number of failures, while being male and higher study times increase G1 grades. The initial model had a BIC value of 2130.34 while the final model was 2014.25, and model diagnostics show that model residuals are normal (Appx. Fig. 6).

The final model for final grades (G3) is:

G3 = 12.63 + 0.65(Sex if M) + 0.43(Study time) – 1.11(Failures) – 2.02(School support) – 0.43(Go out) – 0.06(Absences)

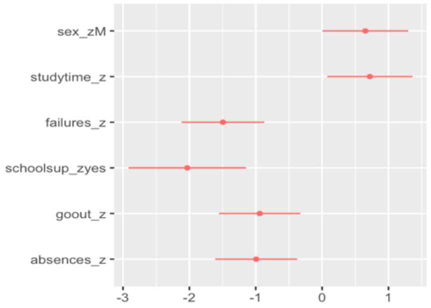

Figure 5: Whisker plot of G3 model coefficients and 95% CI

However, this model was created after 0's were removed from G3, so this model can only be interpreted given the student did not receive a zero; this choice resulted in better model diagnostics and more normal residuals (Appx. Fig. 7). The factors 'higher education' and 'family support' became no longer significant while 'absences' did. The initial model had a BIC value of 1921.38 while the final model was 1802.69, and model diagnostics show that model residuals are normal (Appx. Fig. 8).

0.5 Conclusion

Given the results, we can conclude that 1) alcohol consumption, either weekend or weekday, does not affect grades, 2) age, study time, health, and absences affect number of class failures, and 3) significant common factors affecting first period and final grades are sex, study time, failures, school support, and going out frequency.

In all models, none had a significant Dalc or Walc variable (daily and weekend alcohol consumption); their p-values were high (> 0.7) while additionally having very small coefficients (< 0.08). To lower number of failures, schools should keep an eye on students' health and number of absences. Study time was a significant factor in all three models so the two schools should encourage increased study time while discouraging excessive going out.

While it is tempting to apply these findings to all schools and students, because the sample data taken are from two specific schools (Galriel Pereora and Mousiho da Silveira), the possible applicable population are only those two secondary schools in Portugal. For further analysis, we recommend examination on why extra educational and family support could lower grades. This data was also collected in 2006 and updated data would be beneficial in providing a more recent look into academic performance in Portugal.

Appendix

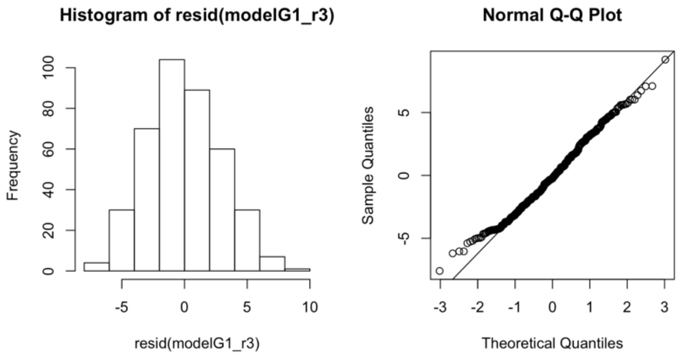

Figure 6: Model diagnostics of G1's model—histogram of residuals and QQ-plot

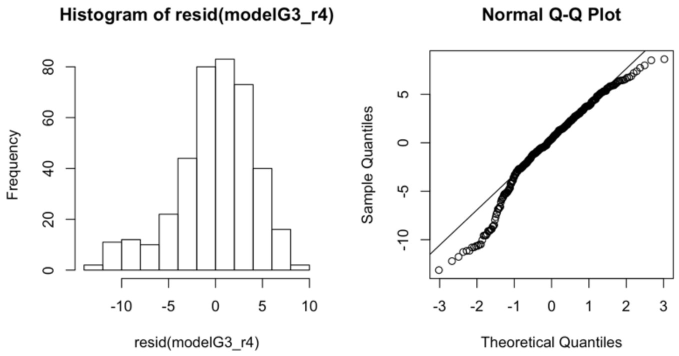

Figure 7: Model diagnostics of G3's model—histogram of residuals and QQ-plot

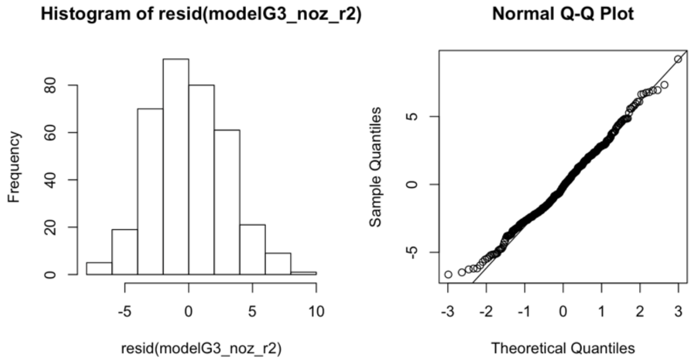

Figure 8: Model diagnostics of G3's model after removing failures—histogram of residuals and QQ-plot look much more normal compared to ones in Fig, 7