Predicting Students' Final Grades (G3) in the Portuguese Course

Hilary Xu[a], Zhiyi Yang[a], Emmy Ji[a], and Jingxian Zhao[a]

[a]School of Mathematics and Statistics, University of Sydney, NSW, Australia

This version was compiled on November 9, 2025

This study models students' final Portuguese grades (G3) using the UCI Student Performance dataset. Multiple linear regression assessed academic, behavioural, and demographic factors. In conclusion, G2 showed a very strong predictive power for G3, while other factors had moderate influences. GitHub: https://github.sydney.edu.au/mixu0768/L08G05.git

Multiple linear regression | Model selection | Educational

Introduction

This study investigates the key determinants of students' final grades (G3, 0–20) in the Portuguese secondary education dataset from the UCI Machine Learning Repository (Cortez, 2008).We employ multiple linear regression to balance accuracy and interpretability, examining academic (G2, failures, absences), behavioural (weekday alcohol use, school support), and demographic factors (school and gender). To mitigate multicollinearity, G1 is excluded from the final specification, and model performance is assessed using AIC and 10-fold cross-validation.

Dataset

The dataset contains 649 students and 33 variables from two schools, combining grades with survey measures of background and behaviour. Each row is a unique student (independence by design). Variables include G3, prior grades (G1, G2), behaviours (study_time, failures, absences), demographics/supports (sex, school …). We factorised categoricals and used absences_log = log1p(absences).

All analyses were conducted in R (R Core Team, 2024) using the tidyverse suite (Wickham *et al.*, 2024) for data wrangling and visualization. Model estimation employed the caret (Kuhn, 2024) and leaps (Lumley, 2024) packages, while diagnostics and correlation plots were produced with MASS (Ripley, 2024) and ggcorrplot (Kassambara, 2024). Tables were formatted using kableExtra (Zhu, 2024), and figures were composed with ggplot2 (Wickham, 2016) and patchwork (Pedersen, 2024).

Analysis

Independent variables Transformation. Binary and ordinal variables are converted into factors so the model recognizes them as categorical predictors, allowing coefficients to represent group differences instead of meaningless numeric changes.In EDA we find the absences was highly right-skewed with extreme values, a log transformation was applied to stabilize variance and make residuals more normally distributed in Figure 3.

Model Selection. To identify the most suitable model, we applied both backward and forward selection while testing whether including G1 or G2 improved model performance. Three models were compared, Model A (excluding G1 and G2), Model B (including both), and Model C (keeping G2 only). Since G2 is temporally closer to the final grade G3, we excluded G1 to avoid redundancy. Also, we then conducted 10-fold cross-validation to assess each model's generalization performance.

Model B shows the best overall performance based on AIC and cross-validation, but Pearson Correlation indicates strong multicollinearity between G1 and G2 mentioned in . The correlation between G1 and G2 is extremely high (0.86). Therefore, Model C is selected as the optimal model for further selection.

Though Exhaustive Model achieves slightly lower AIC (2139.62), the performance in 10-fold cross-validation (CV_R^2 is 0.86) is lower than Model C (CV_R^2 is 0.865), indicating weaker predictive stability. Therefore, we selected Model C as the final specification, as it balances goodness of fit and generalization.

Besides, we applied a log transformation to G3, but the model performance showed almost no difference, indicating that the transformation did not improve the model's predictive power.

Finally, we converted all ordinal variables into factors and applied the stepwise selection. How-

ever, the resulting model's AIC (2144.67) is higher Model C and cross-validation performance is worse, so we retained **Model C** as the final Model.

Assumption check (Figure 4).

- **Independence**: In this dataset, each observation represents an individual student, and their records are assumed to be independent of one another. There is no clustering or repeated measurement, so the independence assumption of the linear model is reasonably satisfied — after all, no two people in the world are exactly the same, and even twins differ in their personal habits.
- **Constant Variance**: The **Residuals vs Fitted** plot checks the linearity and constant variance assumption — the points are randomly scattered around zero, indicating no clear pattern.
- **Homoskedasticity**: The **Scale–Location** plot tests for homoscedasticity — residuals appear evenly spread, suggesting stable variance.
- **Normality**: The **Normal Q–Q** plot evaluates normality — most points align with the reference line, implying residuals are approximately normal.

Overall, Model C satisfies the main linear regression assumptions.

Results. Fitted model:

$$G3 = 0.625 + 0.983\,G2 - 0.258\,failures + 0.129\,absences\_log - 0.098\,weekday\_alc - 0.310\,schoolMS - 0.221\,sexM + 0.122\,travel\_time - 0.243\,school\_supportyes$$

To interpret the model, we focused on predictors that were statistically significant or theoretically meaningful in explaining academic performance and daily behaviors.

G2 (β = 0.983, p < 0.001) G2 has a very strong positive predictive effect on G3. Specifically, a one-point increase in the second-period grade is associated with an average increase of about 0.98 points in the final grade, by holding other variables constant. This indicates that consistent academic performance throughout the year is the most influential factor in determining final achievement.

Absences (β = 0.129, p = 0.015) Specifically, a 1% rise in absences corresponds to an estimated 0.0013-point increase in the final grade, indicating that attendance has only a minor but statistically significant association with academic outcomes.

Failures (β = -0.258, p = 0.004) Specifically, each additional past failure is associated with an average decrease of about 0.26 points in the final grade, indicating that students who have previously repeated a year or failed subjects tend to perform substantially worse in the final evaluation.

Schol_supportyes (β = -0.243, p = 0.140) Students with school support scored on average 0.24 points lower in G3, suggesting that extra assistance is typically offered to those facing academic challenge

Discussion and conclusion

Limitations.

1. First, the dataset covers only Portuguese students from two schools, which restricts cross-cultural or nationwide applicability.
2. The exclusion of G1 was necessary to avoid multicollinearity with G2, yet this may have reduced the model's explanatory power for early-stage performance.
3. The model assumes linear relationships between predictors and the outcome. However, variables such as study time and absences may exhibit non-linear effects that a linear model cannot capture.

Conclusion. Overall, the analysis revealed that G2 (second-period grade) has the strongest predictive power for G3 (final grade), while factors such as failures, absences, weekday alcohol use, and school context also exert meaningful effects. These findings suggest that consistent academic performance and responsible daily habits are key to student success. For future research, models that exclude G1 and G2 could be build to predict more impact of behavioural and contextual factors except former grades. Additionally, building non-linear model Random Forest or Ridge which was analysis by the author of this dataset.(Cortez and Silva, 2008). Moreover, integrating broader socio-psychological variables could further improve prediction accuracy and deepen understanding of what drives academic achievement.

Appendix

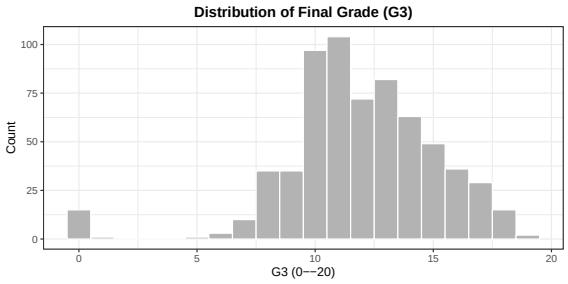

Fig. 1. (a) Distribution of G3 (b) Correlation among G1–G3 and key predictors

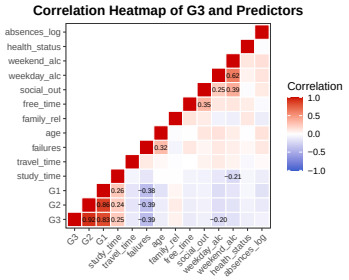

Fig. 2. (a) Distribution of G3 (b) Correlation among G1–G3 and key predictors

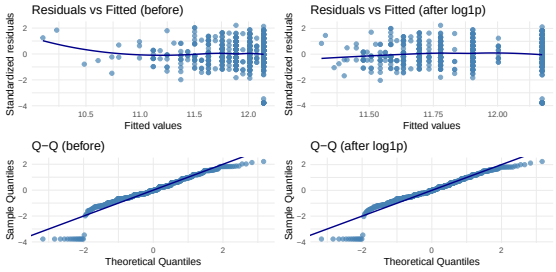

Fig. 3. Effect of log transformation on absences

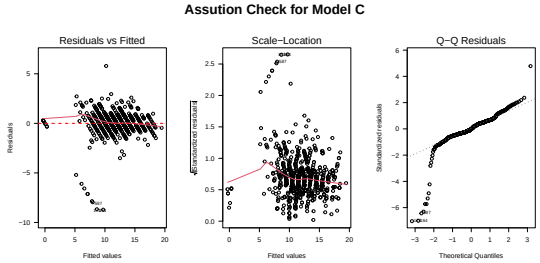

Fig. 4. Diagnostic plots for Model C

References

Cortez P (2008). "Student Performance." UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TG7T.

Cortez P, Silva A (2008). "Modeling Student Performance: A Case Study." *Proceedings of the 2008 International Conference on Educational Data Mining*, pp. 1–8. URL https://archive.ics.uci.edu/ml/datasets/student+performance.

Kassambara A (2024). *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*. R package version 0.1.4, URL https://CRAN.R-project.org/package=ggcorrplot.

Kuhn M (2024). *caret: Classification and Regression Training*. R package version 6.0-94, URL https://CRAN.R-project.org/package=caret.

Lumley T (2024). *leaps: Regression Subset Selection*. R package version 3.1, URL https://CRAN.R-project.org/package=leaps.

Pedersen TL (2024). *patchwork: The Composer of Plots*. R package version 1.2.0, URL https://CRAN.R-project.org/package=patchwork.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ripley B (2024). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-60, URL https://CRAN.R-project.org/package=MASS.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

Wickham H, *et al.* (2024). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 2.0.0, URL https://CRAN.R-project.org/package=tidyverse.

Zhu H (2024). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.4.0, URL https://CRAN.R-project.org/package=kableExtra.