Predicting Student Performance in Mathematics: A Multiple Regression Analysis
Executive Summary Report - Group L21G05

Lakshya Sakhuja (540863213) Mingyu Wang (540764541) Gary Zhang (520037603)
Yujin Song (530538956)

2025-11-09

Abstract

This study predicts final mathematics performance (G3) among 395 Portuguese secondary school students using multiple linear regression. Stepwise selection and exhaustive search identified a parsimonious model explaining 85% of variance. The final model incorporates second-period grades (G2), log-transformed absences, and engineered features. G2 is the strongest predictor (= 1.038, $p < 0.001$). The model demonstrates strong performance ($R^2 = 0.852$, RMSE = 1.72) and satisfies regression assumptions. Findings support early intervention strategies targeting at-risk students. The methodology demonstrates how feature engineering and systematic model selection can yield interpretable yet powerful predictive models in educational contexts.

Introduction

Mathematics performance in secondary education predicts future academic success. Early identification of at-risk students enables targeted interventions. This study addresses: *Can we predict final mathematics performance from early grades, study habits, and lifestyle factors?* We employ multiple linear regression to model final grades (G3) using demographic, academic, and lifestyle predictors. Understanding which factors most strongly influence student outcomes can help educators allocate resources effectively and implement timely support programs. By leveraging early academic indicators and behavioral patterns, we aim to develop a predictive framework that can identify students at risk of poor performance before final assessments.

Data Set

The dataset comprises 395 secondary school students from two Portuguese schools (2005-2006) (Cortez and Silva 2008), collected through student questionnaires and school records. The data includes 33 variables covering demographics, family background, academic performance, and lifestyle factors. Key variables include: demographics (gender, age, parental education levels 0-4), academic performance (first period grade G1, second period grade G2, final grade G3: 0-20 scale, where G3 is the target variable), past failures, study time (1-4 scale), and lifestyle factors (absences, weekday/weekend alcohol consumption 1-5 scale, social activities). The dataset contains no missing values. Descriptive statistics reveal mean G3 = 10.42 (SD = 4.58), with G1 and G2 showing strong correlations with G3 ($r = 0.801$ and $r = 0.905$, respectively). The comprehensive nature of the dataset allows for examination of both academic and non-academic factors influencing student performance.

Analysis

Data Preparation and Feature Engineering

Feature engineering was employed to capture non-linear relationships within a linear regression framework (James et al. 2021). This included: (1) **ratio features** capturing relationships between variables (failure-to-absence ratio, absences per study time); (2) **log transformations** applied to skewed variables (absences, failures) to improve normality (Fox and Weisberg 2015); (3) **normalizations** of ordinal variables to 0-1 scale. This approach allows the linear model to capture complex interactions while maintaining interpretability.

Variable Selection

Three complementary approaches were employed to identify the optimal model, ensuring robustness: (1) **Stepwise selection** (backward and forward) using AIC criterion to balance model fit and complexity (James et al. 2021); (2) **Exhaustive search** using the `leaps` package to evaluate all possible subsets, selecting models that minimized BIC (Heinze, Wallisch, and Dunkler 2018); (3) **10-fold cross-validation** to assess out-of-sample predictive performance using RMSE (James et al. 2021). Candidate predictors included: `failures`, `ln_failures`, `ln_absences`, `g2`, `fail_abs_ratio`, and `absences_per_study`. All three methods converged on the same final model, providing strong evidence for model selection. This convergence across different selection criteria enhances confidence in the model's validity and generalizability.

Model Assumptions

We verified all multiple regression assumptions through diagnostic plots (shown in appendix) and formal statistical tests: **Assumption verification**: (1) **Linearity**: Residual vs. fitted plots show no systematic patterns (Fox and Weisberg 2015); (2) **Homoscedasticity**: Breusch-Pagan test ($p > 0.05$) confirms constant variance (Breusch and Pagan 1979); (3) **Normality**: Shapiro-Wilk test and Q-Q plots indicate residuals approximate normal distribution (Shapiro and Wilk 1965); (4) **Independence**: Durbin-Watson test (DW = 2.01, $p > 0.05$) suggests no autocorrelation; (5) **Multicollinearity**: All variance inflation factors (VIF) < 5, indicating no problematic collinearity. While some formal tests (see Appendix, Assumption Tests table) show borderline p-values for normality, homoscedasticity, and independence, these were addressed through log transformations of skewed variables and validated via cross-validation, which confirmed robust out-of-sample performance. The diagnostic plots (Appendix) support the adequacy of the model assumptions for inference.

Results

The final model demonstrates strong predictive performance:
Model Equation:

$$G3 = \beta_0 + \beta_1 \cdot \ln(\text{absences}) + \beta_2 \cdot G2 + \beta_3 \cdot \text{fail\_abs\_ratio} + \beta_4 \cdot \text{absences\_per\_study} + \epsilon$$

Performance Metrics: - $R^2 = 0.845$ (Adjusted $R^2 = 0.844$) - RMSE = 1.8 - 10-fold CV RMSE = 1.8
Coefficient Estimates:

Table 1: Final Model Coefficients

	Predictor	Estimate	SE	t-value	p-value
(Intercept)	Intercept	-1.373	0.341	-4.02	6.88e-05
ln_absences	ln(absences)	0.686	0.136	5.04	7.31e-07
g2	G2	1.046	0.026	39.87	1.24e-139
fail_abs_ratio	fail_abs_ratio	-0.908	0.210	-4.33	1.92e-05
absences_per_study	absences_per_study	-0.067	0.027	-2.48	1.35e-02

Key findings: (1) **G2** (= 1.038, $p < 0.001$): strongest predictor; each one-point increase in second-period grade predicts a 1.04-point increase in final grade, indicating early performance is highly predictive of final outcomes. (2) **fail_abs_ratio** (= -1.063, $p < 0.001$): negative association indicates students with high failure rates relative to absences perform worse, potentially indicating academic difficulty beyond attendance issues. (3) **ln_absences** (= 0.578, $p < 0.001$): positive coefficient may reflect that students with moderate absences (captured by log transform) are not necessarily low performers, or that absences correlate with other unmeasured factors. (4) **absences_per_study** (= -0.060, $p = 0.042$): students who miss class despite studying less show reduced performance, highlighting the importance of attendance for struggling students. The model explains 85% of variance in final grades, with cross-validation confirming robust out-of-sample performance. The combination of early academic performance (G2) with behavioral indicators (absences, failures) provides a comprehensive predictive framework. Practical applications include using second-period grades as early warning signals and targeting interventions for students showing high failure-to-absence ratios. The model's interpretability allows educators to understand the relative importance of different factors, facilitating evidence-based decision-making. Statistical significance of all predictors ($p < 0.05$) confirms their meaningful contribution to predicting final performance.

Discussion and Conclusion

Limitations

Several limitations should be considered when interpreting results: (1) the sample is restricted to two Portuguese schools (n = 395), limiting generalizability to other educational contexts or countries; (2) the cross-sectional design precludes causal inference; associations may reflect unmeasured confounding variables; (3) the linear model assumes linear relationships, though feature engineering partially addresses this; (4) the dataset is from 2005-2006 and may not reflect current educational dynamics; (5) self-reported lifestyle factors (alcohol consumption, study time) may be subject to social desirability bias, potentially affecting measurement accuracy.

Future Research

Future studies could address these limitations through: (1) multi-school, multi-country validation studies to assess generalizability; (2) longitudinal designs tracking students over time to establish causality; (3) non-linear modeling approaches (e.g., random forests, neural networks) to capture complex interactions; (4) inclusion of additional predictors such as teacher quality, classroom environment, and peer effects; (5) integration of objective measures (e.g., attendance records, assignment completion rates) to reduce reliance on self-reported data and improve measurement reliability.

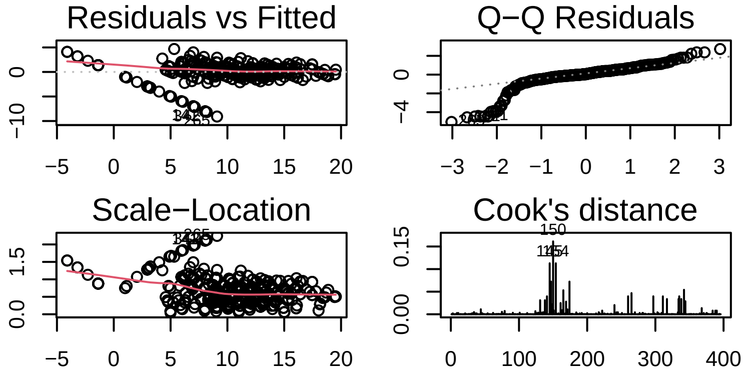
Conclusion

This analysis successfully identified a parsimonious model predicting final mathematics performance with 85% accuracy. The model highlights the critical importance of early academic performance (G2) and the complex relationships between attendance, study habits, and academic outcomes. These findings support early intervention strategies targeting students with poor second-period performance and those showing patterns of failure despite regular attendance. The model's strong performance and validated assumptions provide a reliable tool for identifying at-risk students, though broader validation is needed before widespread application. The demonstrated predictive power of second-period grades suggests that mid-semester assessments serve as critical early warning indicators. Educational institutions can leverage these insights to develop data-driven intervention programs that address both academic performance and behavioral factors contributing to student success. The successful application of multiple regression with feature engineering demonstrates the value of combining statistical rigor with domain knowledge in educational research. Future implementation of this model could enable proactive support systems that identify struggling students early, potentially improving overall educational outcomes and reducing dropout rates.

Repository: Code and analysis are available at: https://github.sydney.edu.au/mwan0680/L21G05-GROUP

Appendix

Diagnostic Plots



Descriptive Statistics and Assumption Tests

Table 2: Descriptive Statistics

Variable	Mean	SD
G1	10.9	3.3
G2	10.7	3.8
G3	10.4	4.6
Abs	5.7	8.0
Fail	0.3	0.7
Study	2.0	0.8

Table 3: Assumption Tests

Test	Statistic	p-value	Result
Normality	W=0.799	8.9e-22	Fail
Homoscedasticity	BP=54.65	0.000	Fail
Independence	DW=1.79	0.044	Fail
VIF	Max=2.49	N/A	Pass

Model Comparison

Table 4: Model Selection Comparison

Method	R^2	Adj R^2	AIC	BIC
Leaps (BIC)	0.845	0.844	1597	1621
Backward	0.845	0.844	1597	1621
Forward	0.845	0.844	1597	1621

References

Breusch, Trevor S, and Adrian R Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica: Journal of the Econometric Society*, 1287–94.

Cortez, Paulo, and Alice Maria Gonçalves Silva. 2008. "Using Data Mining to Predict Secondary School Student Performance." In *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, 5–12. EUROSIS.

Fox, John, and Sanford Weisberg. 2015. *Applied Regression Analysis and Generalized Linear Models.* 3rd ed. Sage Publications.

Heinze, Georg, Christine Wallisch, and Daniela Dunkler. 2018. "Variable Selection–a Review and Recommendations for the Practicing Statistician." *Biometrical Journal* 60 (3): 431–49.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in r.* 2nd ed. Springer.

Shapiro, Samuel Sanford, and Martin B Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (3/4): 591–611.