BIG DATA ECONOMETRICS

Manuel Manzo (s1106217)

The research analyzes the performance of 395 Portuguese secondary school students, examining data that includes grades, demographics, and educational context. The goal is to identify the determinants of the final grade G3, on a scale from 0 to 20, and to predict it through an econometric analysis of the available variables.

DATA CLEANING

Initially, a careful check was conducted to identify any missing values, duplicates, or anomalies within the data. It was found that there were no missing values or obvious irregularities; however, 38 observations were found where the G3 grade was equal to 0. It is highly unlikely that such a large number of students actually received a zero (especially since no student received a grade between 1 and 3). This value may indicate that the student did not take the test, so it was decided to eliminate them.

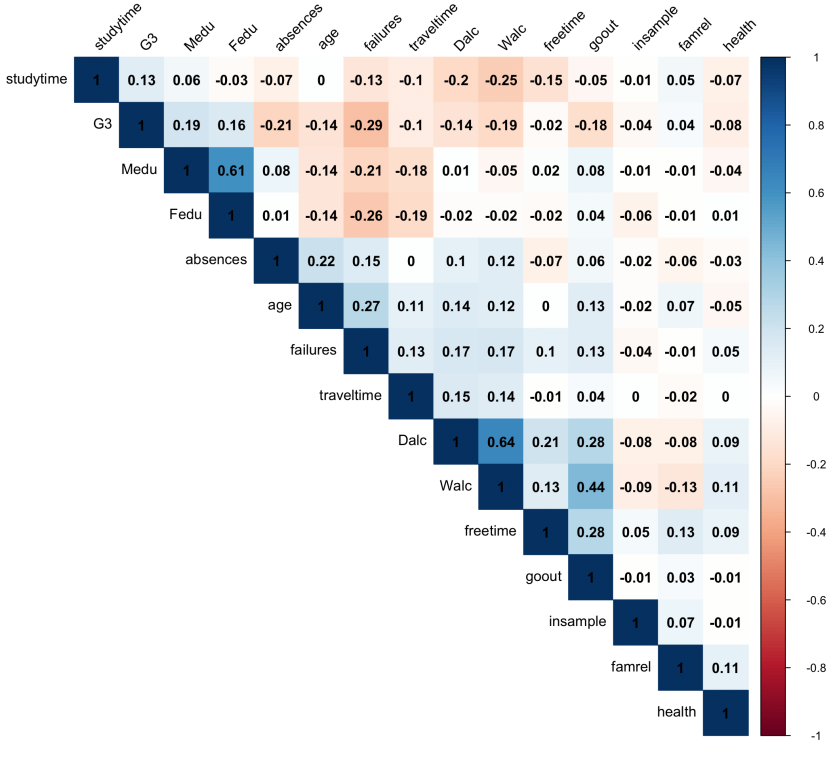
UNDERSTANDING THE VARIABLES

Descriptive Analysis

Using histograms and boxplots, it emerged that repeating students, those with many absences, or a lot of free time spent with friends tend to have lower grades in G3. Conversely, students with parents in certain work sectors, who live in cities, or with access to the internet tend to get better grades.

Correlation Analysis

The heatmap analysis, which shows the correlation between the numerical variables, revealed that the G1 and G2 grades, which represent intermediate results obtained during the school year, have a strong correlation with the final grade G3. These two intermediate variables are also strongly correlated with each other.



MODELS DEVELOPED

In the development of the final linear regression model, a strategic choice was made to exclude the variables G1 and G2 despite their high predictability for the final grade G3. This is because the aim is to explore broader and less immediate factors that influence academic performance. Similarly, the 'school' variable was omitted to ensure that the conclusions of the model can be extended beyond the two schools considered in the study.

Full and Stepwise Model

A regression model was developed to predict G3 using all variables except those previously mentioned. Applying the Stepwise algorithm, a model with acceptable performance on the training set was obtained, but it was significantly lower on the test set. This indicates an excessive complexity of the model (compared to the available data), which includes too many variables, leading to a clear overfitting.

In-sample Rsquared: 0.2779

Out-of-sample Rsquared: 0.144699

Final Model

The final model, selecting only some of the most significant variables from the stepwise model (the most significant ones), showed similar performance between in-sample and out-of-sample data. Specific tests confirmed that the residuals were normally distributed and homoscedastic, while the VIF function verified the absence of multicollinearity.

In-sample Rsquared: 0.2145

Out-of-sample Rsquared: 0.2008

CONCLUSIONS AND INTERPRETATIONS OF THE MODEL

In the next page more information regarding the final model is reported:

	Estimate	Std. Error	P-value
Intercept	13.00698	0.66957	< 2e-16 ***
failures	-1.09479	0.26035	3.52e-05 ***
schoolsupyes	-1.93431	0.48200	7.70e-05 ***
absences	-0.06410	0.02023	0.00170 **
goout	-0.50883	0.15480	0.00114 **
Mjobhealth	2.47522	0.76119	0.00129 **
Mjobother	0.48068	0.54518	0.37870
Mjobservices	1.81887	0.56731	0.00150 **
Mjobteacher	0.88638	0.65834	0.17927

The beta coefficient in a regression model indicates how much the dependent variable (e.g., a student's final grade) is expected to change for each one-unit increase of the independent variable (e.g., number of absences), provided that all other variables in the model remain unchanged. In our case, the analysis confirmed that variables such as failures, absences, and less time spent at school are associated with lower grades. Additionally, children of mothers who work in the medical sector or in services tend to have higher grades.

Out-of-sample goodness of the model

RMSE	Rsquared	MAE
2.8667062	0.2007771	2.3022956

Analyzing the coefficients of the obtained model, we see that the proposed model has limited effectiveness in estimating the final grade based on the other variables.

Including one of the two variables G1 or G2 would have resulted in a significantly better model for predictive purposes, but the model would have been much less useful for the purposes of our analysis

Extra: A full model including G2 was also provided in the R code, the model has an Rsquared of 0.947 out-of-sample. This model also satisfies the tests for normality of residuals and homoscedasticity