Executive summary

Emlyn Bilbie[1], **Aiden Davies**[1], **Joseph Isaacs**[1], **and Daniel Mechtersheimer**[1]

[1]T09oc_early_4, DATA2902, University of Sydney

This version was compiled on November 20, 2020

This report presents the findings of an investigation into predicting the final Mathematics grades of secondary school students in Portugal by performing multiple regression on the results of a survey given to students of Portuguese and Mathematics. Two candidate models were constructed through stepwise variable selection processes. The two models performed simmilarly out-of-sample, so the smaller model was selected, including earlier mathematics marks, family relationship quality, age, and absences from class as predictors. However, the model is prone to inaccuracy in predicting lower marks, and future investigation could be conducted into an 'early intervention' model that does not include past grades.

Introduction

Using a data set containing information about students in Portugal, and their marks in Mathematics and Portuguese, this project aims to develop a multiple rgression model to predict a student's Mathematics grade based on the other variables in the data set. The model will incorporate students' academic performance through the year, in order to provide estimated final marks that could allow students to plan future studies before receiving their final grade, or be used in the case of illness or misadventure. In this report, we will detail the process that was used to develop the model and discuss its effectiveness by examining model stability, and its performance, both in- and out-of-sample.

Data set

The data set was retrieved from the UCI Machine Learning Repository (2014), and contains information relating to students in Portugal, with each row representing a student and each column representing some attribute about the student.

The data was collected for a study by the University of Minho in 2005–06 from two public schools, using school reports and questionnaires (Cortez and Silva, 2008). Two data sets are included, containing information on 395 students of Mathematics and 649 students of Portuguese. Both contain 33 different attributes, with three being grades for the respective subjects and 30 sourced from the questionnaire given to students.

There are 382 students who are listed in both data sets, according to its metadata. We expected that a student's performance in Portuguese might be related to their performance in Mathematics, so for the purposes of our analysis, we merged the two data sets, including only these students.

However, there were several columns which were near-identical between the two data sets, with only a few differing observations. Since these were variables that would not sensibly differ based on the particular subject described, we surmised that the differences may be due to the survey for each subject being administered at slightly different points in time. To avoid choosing between these near-identical columns in the model selection, we filtered our dataset to only contain rows with the same values for these columns, ending up with 320 rows and 38 columns. The five subject-specific columns were the student's three grades (`G1`, `G2`, `G3`), whether or not they paid for extra classes in the subject (`paid`) and the number of absences from the subject (`absences`). We also re-coded some numeric variables that did not have linearly increasing categories as categorical . These were the two parents' levels of educational attainment (`Medu` and `Fedu`), as well as `traveltime` and `studytime`.

Analysis

Assumption checking. In Fig. 1 it appears that the residuals of the model are distributed randomly above and below zero, with the exception of a problematic patch of much lower values. So, if we momentarily disregard these points, the linear model does not appear to be unsuitable. There is also no evidence of fanning or change in variability, so homoskedasticity is fulfilled.

In the Q–Q plot, the straight line is closely followed for the majority of the points, again with the exception of the lower values which deviate quite notably. Despite this, there are enough values for the central limit theorem to hold, so inferences based on the normality assumption should be valid.

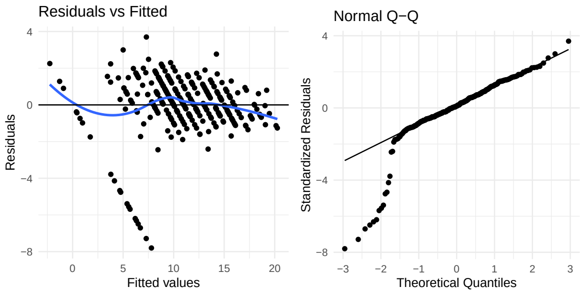

Fig. 1. Residual and Q–Q plots for full model

Observing the pairs of plots in Fig. 5 (see Appendix), it is clear that these troublesome lower points are resultant of a cluster of students who received a mark of 0 for the final mathematics assessment, despite doing well in other assessments. It is possible that this is due to missing values, but since there are real-world explanations for this (such as non-attempts or academic dishonesty), we decided to retain the values. This does mean, however, that caution should be taken when the model predicts smaller values (those below 7 or 8), keeping in mind that they may be inaccurate. However, for predicting higher values, the assumptions are met.

Finally, we assume that all students' responses are independent of each other, although we would need to confirm this with more information on the students sampled.

Model selection. To construct a model for the final mathematics mark (`G3_mat`) by means of multiple regression, we first carried out automated backwards and forwards variable selection using the `step()` function in R, which aims to minimise the model's Akaike information criterion (AIC) value.

The automated backwards selection returned a model containing 12 variables. Using forward selection instead, the process returned a model with 11 variables, of which 9 were common with the backwards model. Both models had similarly high r^2 values of approximately 0.86, indicating that they were strong.

Further fine-tuning of both models was conducted manually. Adding variables to either model did not produce any significant results. Removing variables, however, proved to be more useful. By iteratively assessing the result of removing a variable with

R's `drop1()` function, we managed to remove several insignificant variables from both models. From our original 'backwards' model, we ended up with only five variables, while from our original 'forwards' model, we ended up with seven (see Table 1 in Appendix). The smaller models allow accurate prediction (with similarly high r^2 values), without over-fitting by including too many predictor variables.

Figs. 6 and 7 (see Appendix) show that the behaviour for our two new models is very similar to the full model, so our assumptions still hold for higher values, with the same caveat for lower-performing students.

Model stability. In order to analyse the stability of our models, we started with a model stability plot for the 12 variables used in our original backwards model (Fig. 2). There are no other models which appear particularly dominant, with all groupings showing relatively equal applicability.

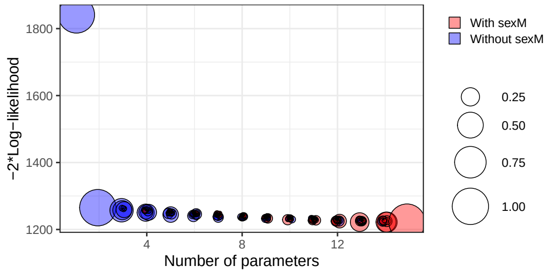

Fig. 2. Model stability plot

A variable inclusion plot (Fig. 3) clearly supports the necessary inclusion of `G2_mat` as a strong predictor variable. The other predictors in our model are more varied, although there is also evidence for their inclusion. So, we will retain both models found with stepwise selection and assess their performance.

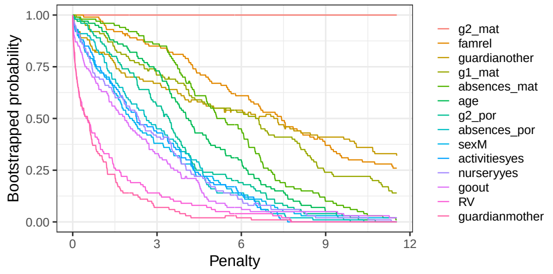

Fig. 3. Variable inclusion plot

Results

In-sample assessment. Our larger and smaller models have very similar AIC values: 1252.96 and 1256.63 respectively. They also have similar r^2 values: 0.849 compared to 0.846. While it is promising to see such strong in-sample performance, the close results do not give us sufficient evidence to select between the two models, or to make any conclusions as to the models' actual effectiveness. Evaluation of the two models' out-of-sample performance is needed to draw firmer conclusions.

Out-of-sample assessment. In order to assess out-of-sample performance, we conducted 10-fold cross-validation estimation, using the caret package for R. For the smaller model, we attained a root mean square error (RMSE) of 1.649 and a mean absolute error (MAE) of 1.024. For the larger model, our RMSE was 1.651 and our MAE was 1.042.

As evidenced by the bootstrapped confidence intervals in Fig. 4, these two errors are not significantly different, meaning the two models performed very similarly overall. Since the more complicated model does not provide any improvements to performance, it makes sense to select the smaller, simpler model for easier computation.

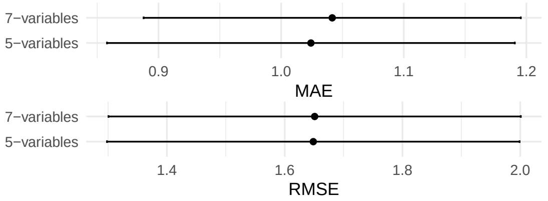

Fig. 4. Bootstrapped confidence intervals

Final model. Since our selected model does not include any variables specific to Portuguese, the model can be fitted against the original, larger Mathematics data set. For this model, we get,

$$\begin{aligned}\mathrm{G3} = &-0.08 + 0.16(\mathrm{G1}) + 0.98(\mathrm{G2}) + 0.04(\mathrm{absences}) + \\ &0.36(\mathrm{famrel}) - 0.2(\mathrm{age}) + \epsilon\end{aligned}$$

As expected, the final mathematics grade (`G3`) tends to increase with the first and second period mathematics grades (`G1` and `G2`). This relationship is stronger with `G2`. This is reasonable since this is the mark that is closest in time to the final mark, so students that do well in second period can be expected to do similarly well in the final assessment. `G3` also increases with better quality of family relationship (`famrel`), indicating that satisfaction with the home environment and family support play significant roles in success. Increases in `age` were associated with a lower grade, possibly as a result of increasing difficulty in later school grades. Further analysis of results across each grade would be required to determine if this is the case. An unexpected result was that more absences from class predicted higher marks. While one would expect that missing class would negatively affect a student's performance, it may also be possible that many absences from class motivate students to study independently, positively impacting their grades.

Discussion and conclusion

As expected, the strongest predictors were students' other marks, but also their family relationship quality, age and absences. Our out-of-sample performance is strong; however, our model may be prone to inaccuracies for lower marks. Further information about the `G3` marks recorded as 0 could help resolve this for future studies.

To apply this model across all schools in Portugal, a future study should be conducted with data from a larger range of different schools, located in different regions of the country. The geographical similarities between the two schools in the data set may misrepresent the relationship between certain variables when compared to students in different regions. For example, if both schools surveyed were in high-income areas then the model may be inaccurate in low-income areas.

A future study could also attempt to build a model without the `G1` and `G2` marks as predictor variables, in order to identify students who are at risk of under-performing at the start of the school year. However, there are ethical issues that may be raised by identifying these students based on personal characteristics such as sex and family relationships.

References

Anderson, D. and Heiss, A. (2020). *equatiomatic: Transform Models into LaTeX Equations*. R package version 0.1.0.

Cortez, P. and Silva, A. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th FUture BUsiness TEchnology Conference*, pages 5–12. EUROSIS.

Eddelbuettel, D. and Balamuta, J. (2020). *pinp: 'pinp' is not 'PNAS'*. R package version 0.0.10.

Hlavac, M. (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. R package version 5.2.2.

Kuhn, M. (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Crowley, J. (2020). *GGally: Extension to ggplot2*. R package version 2.0.0.

Tarr, G., Mueller, S., and Welsh, A. H. (2020). *mplot: Graphical Model Stability and Variable Selection Procedures*. R package version 1.0.4.

Tarr, G., Müller, S., and Welsh, A. H. (2018). mplot: An R package for graphical model stability and variable selection procedures. *Journal of Statistical Software*, 83(9):1–28.

UCI Machine Learning Repository (2014). Student performance dataset. https://archive.ics.uci.edu/ml/datasets/student+performance. Accessed: 20 November 2020.

Wickham, H. (2019). *tidyverse: Easily Install and Load the Tidyverse*. R package version 1.3.0.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Appendix

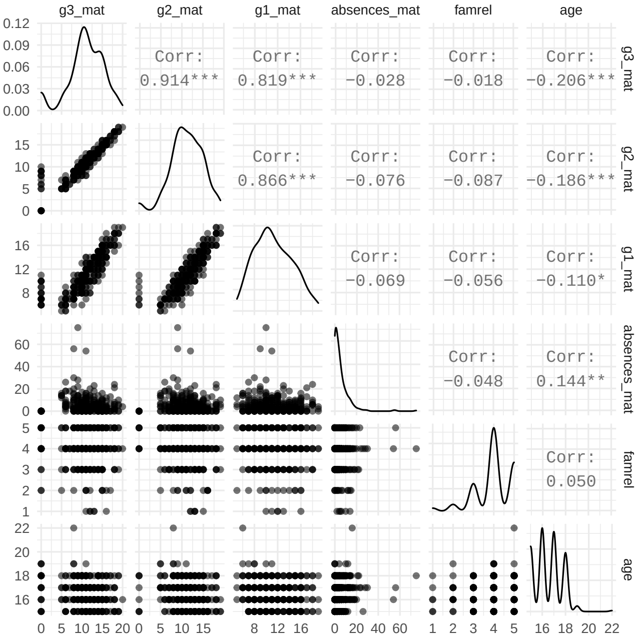

Fig. 5. Pairs for seleted numeric variables

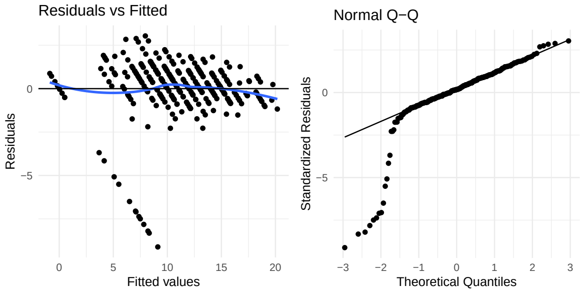

Fig. 6. Residual and Q-Q plots for 5-predictor model

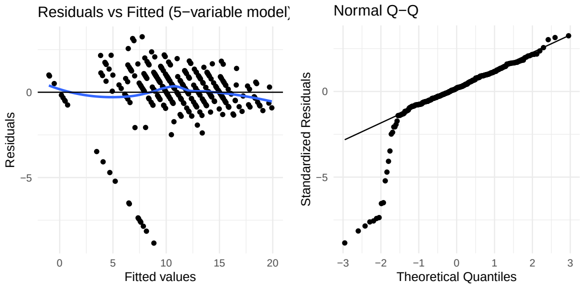

Fig. 7. Residual and Q-Q plots for 7-predictor model

Table 1. Updated models

	Dependent variable:	
	g3_mat	
	(1)	(2)
g2_mat	0.933***	0.952***
	(0.053)	(0.053)
famrel	0.343***	0.313***
	(0.108)	(0.108)
g1_mat	0.134**	0.157***
	(0.061)	(0.060)
Walc	0.177**	
	(0.079)	
age	−0.254***	−0.198**
	(0.088)	(0.086)
absences_mat	0.027**	0.029**
	(0.013)	(0.012)
g2_por	0.100**	
	(0.051)	
Constant	0.119	0.504
	(1.542)	(1.548)
R^2	0.849	0.846
Adjusted R^2	0.846	0.843
Akaike Inf. Crit.	1,252.957	1,256.634
Note:	*p<0.1; **p<0.05; ***p<0.01	