Reveal high correlated factors which can affect the grades of Portuguese language course of students by using multinomial regression

Bohao Zou

Department of Statistic, UC Davis

Introduction

Grade is one of the most important things for a student. In this project, I will use the data from the UCI Machine learning repository named “Student Performance Data set” to find *which factors can affect the grades of Portuguese language of those students.* The result of this project may give some tips for guardians to improve the grades of their children.

The sample size of this data set is 649 and there are 30 predictors. There are 3 grades G1, G2 and G3 for a student for different periods. For constructing a category response variable, I will first calculate the mean grades of those 3 grades. Then sorted those mean grades and split them into 3 parts with equal number of subjects. At the end, there are three categories of Y response variable. Its are "Low", "Medium" and "High" respectively.

Method

Proportional Odds Model

The response categories coded as $j = 1, 2, ..., M$, we can define a new response z_{im} with $z_{im} = \sum_{j=1}^{m} y_{ij}$.

The $z_{im} = 1$ if $y_{ij} = 1$ for $j \leq m, 1 \leq m \leq M - 1$, otherwise $z_{im} = 0$. With $\mu_{im} = E(z_{im})$ and with link function $g(x) = logit(x)$, the model is

$$g(\mu_{im}) = \beta_{0m} + X_i\beta$$

with ordered intercepts $\beta_{01} \leq \beta_{02} \leq ... \leq \beta_{0,M-1}$. This model can work for ordered and unordered categorical data but it is particularly useful if the data are ordered.

Baseline Odds Model

In the baseline odds model, we need to select a baseline category at first. Then we can assume the baseline category is 1, then the linear predictor in this model is

$$\frac{\mu_{ij}}{\mu_{i1}} = exp(\eta_{ij}), \quad \eta_{ij} = X_i\beta_j, \quad 2 \leq j \leq M$$

This model can also be used with ordered and unordered data. However, this model has lots of parameters. The baseline needs to be selected carefully because it can affect the interpretation of the result.

Model Choice

Because we have those two different models. We need to compare these two models by using this data set and some criteria first. Then choose one which has better performance and use it for this analysis. I build a pipeline for this comparision. The compared pipeline is:

1. Build a Proportional Odds Model and a Baseline odds model with same linear predictors. The baseline category of Baseline odds model is *"low"*. The linear predictor for those two model is

$$\eta_{ij} = \beta_0 + x_1\beta_1 + ... + x_{30}\beta_{30}$$

   This means all 30 predictor variables are in those model.
2. Check if there exist lack of fit in those two models. If one model has lack of fit, this means we should use another model.
3. We can compare the AIC value of those two models. The better model has smaller AIC value.
4. We can treat those two models as two different classifiers. Then uses 10 fold cross validation for model choice and model evaluation. The better model has more accuracy result.

Model Selection

There are 30 predictors in our data set that can be treated as candidates of predictors. We will use some criteria to help us for model selection like AIC or BIC. The lower bound of model selection is only containing the interception of regression model. The upper bound of model selection is all the 30 predictors.

Because the number of predictor candidates in upper bound are not too much (30 predictors). For tending to find a medium and suitable model, ***we will use AIC as the criteria***. The direction of model selection is ***"forward & backward"***, which means variable can be added or removed from model by the AIC criteria.

Model Diagnostic

Model diagnosis is the key part for an analysis project. It can significantly influence the result and the interpretation of the model. We should treat it carefully.

Check Lack of Fit

We will use Pearson residuals to check if lack-of-fit exist in our model. We can draw *Pearson residuals vs. Fitted value plot* first and then add a overlaying smoothing splines to fit those points in plot. If the spline fluctuate around 0 slightly, it indicate that there does not exist lack-of-fit in our model. Otherwise, there exist lack of fit in model.

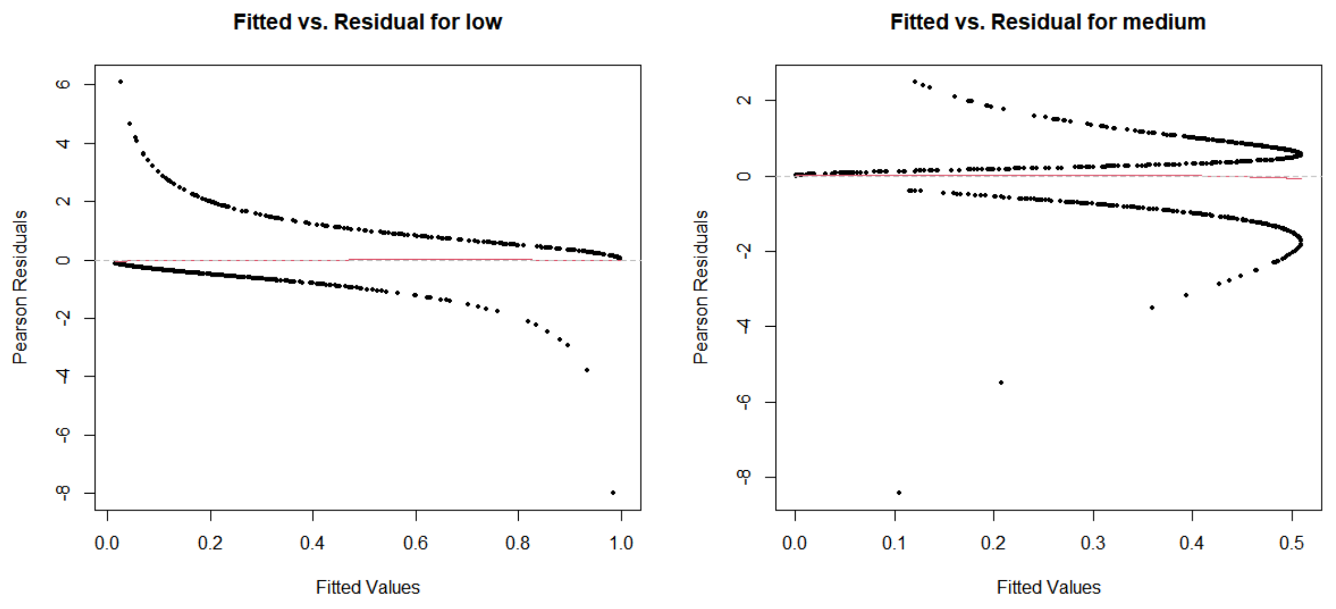
Model Fitting Information

We can use log-likelihood ratio test(LRT) to test if there exist significant regression relationship between response variable Y and entire set of X variables. The null model for LRT is the model with only a constant in it. If the p-value of LRT is smaller than 0.05, which means our model fits significantly better than null model and there exist significant regression relationship between response variable Y and entire set of X variables.

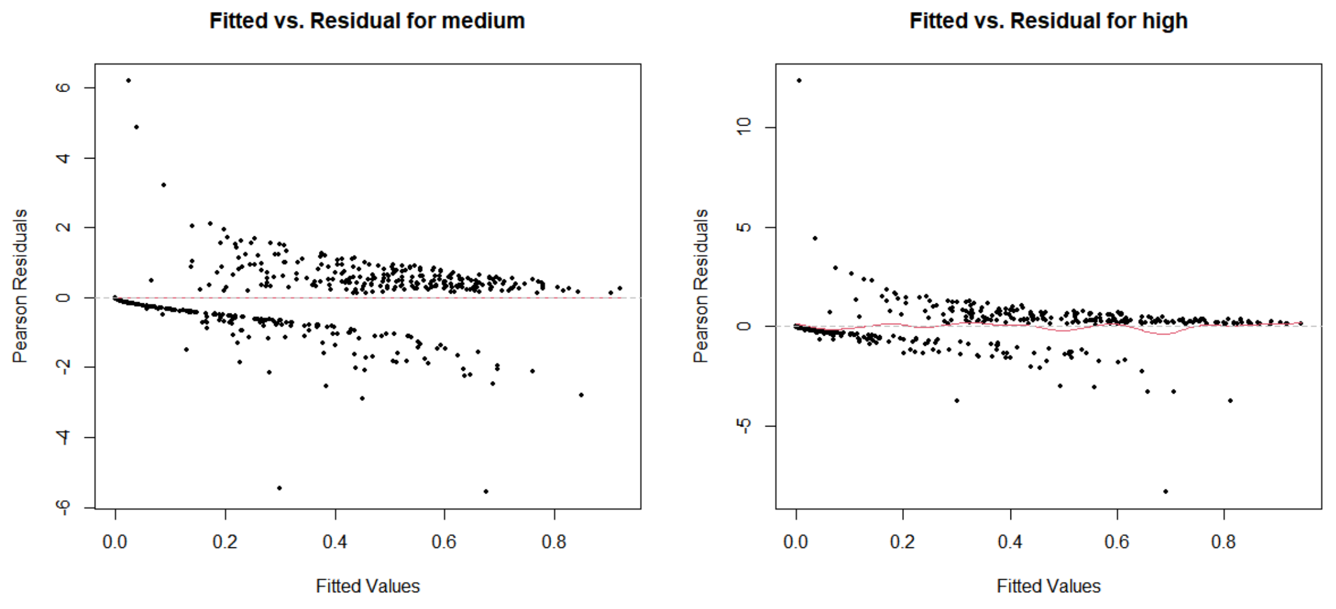
We can also use the 10-fold Cross Validation to check the fitting information of our model. If our model fits good, the mean accuracy of 10-fold Cross Validation will be higher than previous model.

Result of Model Choice

The Pearson residuals vs. Fitted values for proportional odds model for categories "low" and "medium". The plot is showed below:



The Pearson residuals vs. Fitted values for baseline odds model for categories "$\frac{medium}{low}$" and "$\frac{high}{low}$". The plot is showed below:

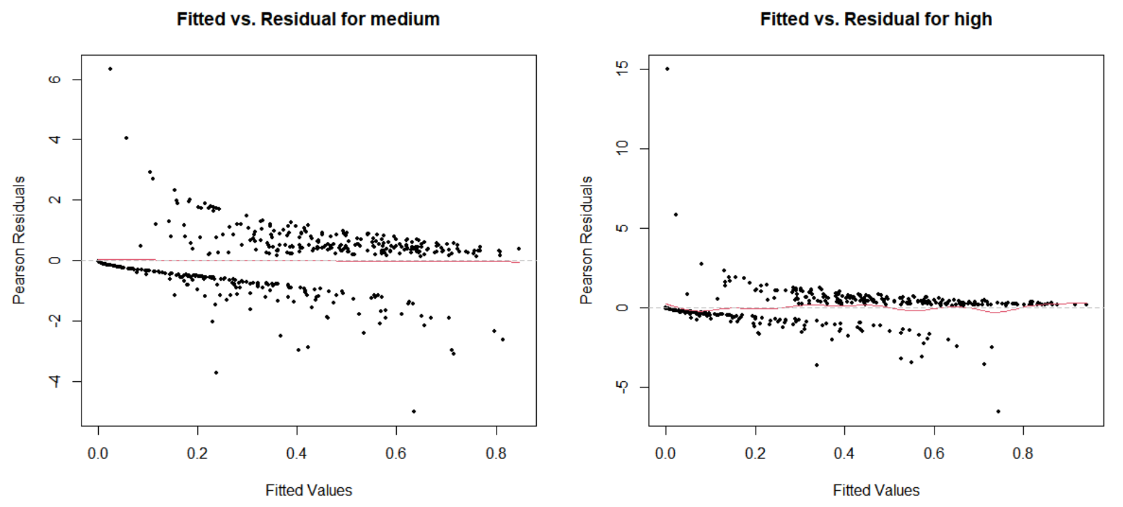


The AIC of proportional odds model is 1179.817, the AIC of baseline odds model is 1170.583. We can know the baseline odds model is better than proportional odds model at this step comparison.

After 10-fold Cross Validation, the mean accuracy of proportional odds model is 0.5375. The mean accuracy of baseline odds model is 0.5608. The baseline odds has better performance at this comparison. In the end, ***we will choose baseline odds model, the baseline category is*** $"low"$.

Result of Model Selection

After model selection by AIC, there are 16 predictors in our model. We name this model as selected model. The the Pearson residuals vs. Fitted values plot for selected model is



Result of Model Fitting Information

The result of LRT test tells us that there is a significant regression relationship between Y and entire set of X variables and the selected model fits significantly better than the null model.

The mean accuracy of selected model that build with 16 predictors is 0.5762. A significance improve compares with previous baseline odd model. This means selected model is better.

	Medium vs. Low			High vs. Low		
	Coefficient	Std.Error	P-Value	Coefficient	Std.Error	P-Value
Intercept	0.385	0.816	0.636	-1.592	1.153	0.167
school,MS	-1.732	0.282	8.20e-10	-1.425	0.307	3.52e-6
sex,Male	-0.847	0.275	2.08e-3	-1.207	0.301	6.06e-5
Mother's Edu	-0.077	0.140	0.581	0.224	0.152	0.140
Father's Edu	0.340	0.151	0.025	0.249	0.162	0.125
Father's Job, health	-2.497	0.829	0.002	-0.903	0.839	0.281
Father's Job, other	-0.447	0.484	0.354	-0.172	0.574	0.763
Father's Job, services	-0.945	0.515	0.066	-0.386	0.599	0.518
Father's Job, teacher	-1.143	0.867	0.187	0.626	0.896	0.484
Reason, home	0.863	0.330	0.009	0.900	0.363	0.013
Reason, other	0.276	0.382	0.469	-0.026	0.441	0.95
Reason, reputation	-0.088	0.343	0.797	0.596	0.350	0.089
Guardian, mother	-0.938	0.300	0.001	-0.656	0.327	0.045
Guardian, other	-0.096	0.619	0.876	0.225	0.760	0.767
Study time	0.161	0.161	0.315	0.439	0.171	0.011
Failures	-1.772	0.321	3.55e-8	-2.61	0.540	1.38e-6
School support, yes	-0.748	0.376	0.004	-1.968	0.463	2.14e-5
Family support, yes	0.253	0.252	0.315	-0.228	0.272	0.402
Activities, yes	0.626	0.251	0.012	0.717	0.272	0.008
Higher Edu, yes	1.247	0.396	0.001	2.814	0.783	0.000
Workday Alcohol	0.015	0.132	0.907	-0.365	0.170	0.032
Health	0.105	0.086	0.224	-0.070	0.091	0.445
Absences	-0.061	0.026	0.019	-0.08	0.029	0.007

Discussion

In the model of Medium vs. Low and model of High vs. Low, The students who have the factors of *1. Mousinho da Silveira(MS) school, 2.Sex, Male, 3.Father's job is health care related, 4.Guardian is mother, 5.number of past class failures are high, 6.Have extra educational support, 7.Have higher number of school absences, 8.Workday alcohol consumption* tends to have low grade of Portuguese.

The students who have the factors of *1.Father's educational level is high, 2.Reason why he/she choose this school is close to home, 3.Have extra-curricular activities, 4.Wants to take higher education, 5.Have longer study time* tends to have medium or high grade.

The limitation of this analysis is that we do not know the factors which may affect the grades between "Medium" and "High". This is because we used "Low" as baseline category. By this setting, we can not show the attribute "order" of response variable Y-grade. The improvement of this analysis is that we can set the baseline category as "Medium" but not "Low".

References

1. STA 223-LectureNotes-w21-1.pdf
2. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.