BST 223, Project 2.
Reveal high correlated factors which can affect the grades of Portuguese language course of students by using multinomial regression.

Bohao Zou
(917 796 070)

3/8/2021

Department of Statistics
University of California, Davis

1 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. 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. Finally, if a student falls into the 1-th part, he/she will be graded as "low" ; if he/she falls into the 2-th part, I will assign "medium". For the 3-th part, he/she will be graded as "high". In this data set, the predictors are ***1.school-binary, 2.sex-binary, 3.age-numeric, 4.address-binary, 5.family size-binary 6.parent's cohabitation status-binary, 7.mother's education-numeric, 8.father's education-numeric, 9.mother's job-nomial, 10.father's job-nomial, 11.reason to choose this school-nomial, 12.student's guardian-nomial, 13.home to school travel time-numeric, 14.weekly study time-numeric, 15.number of past class failures-numeric, 16.extra educational support-binary 17.family educational support-binary 18.extra paid classes within the course subject-binary, 19.extra-curricular activities-binary, 20.attended nursery school-binary, 21.wants to take higher education-binary, 22.Internet access at home-binary, 23.with a romantic relationship-binary, 24.quality of family relationships-numeric, 25.free time after school-numeric, 26.going out with friends-numeric, 27.workday alcohol consumption-numeric, 28.weekend alcohol consumption-numeric, 29.current health status-numeric, 30.number of school absences-numeric***.

In this project, if the p-value of some statistics are lower than 0.05, then we will treat them as statistical significance.

2 Method

2.1 Model Choice

The response variable Y of this data set is a multinomial category variable. We have two models for solving this multinomial regression problem.

2.1.1 Proportional Odd 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.

2.1.2 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.

2.1.3 Pipeline Build

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. The compared pipeline is

1. Build a Proportional Odd 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.

By using this pipeline, we can choose one model from Proportional odds model and Baseline odds model.

2.2 Model Selection

There are 30 predictors in our data set that can be treated as candidates of predictors. It is a huge number for us to pick and analyze them one by one if we also want to consider the interaction items between those predictors. In this section, 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.

Compare the formula of AIC and BIC

$$AIC: \quad AIC_p = nlog(\frac{SSE_p}{n}) + 2p, \quad k = 2$$
$$BIC: \quad BIC_p = nlog(\frac{SSE_p}{n}) + log(n)p, \quad k = log(n)$$

where p represents the number of predictors in model, n represents the number of subjects in data. We can know BIC penalties more on the model complexity. 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.

2.3 Model Diagnostic

2.3.1 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.

2.3.2 Model Fitting Information

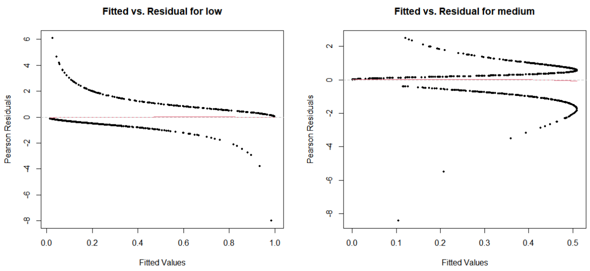
We can use log-likelihood ratio test(LRT) to test if there exist significant regression relationship between response variable Y and entire set of X variables. The null model for LRT is the model with only a constant in it. The null hypothesis H_0 is that there is no difference between null model and final model. 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.

3 Results

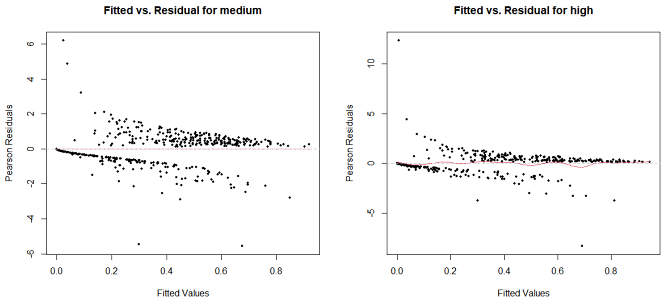
3.1 Model Choice

Following the pipeline, which builds in section 2.1.3, we need to use all 30 predictors to build two different model, proportional odds model and baseline odds model respectively. After built two models, We need to check if there exist lack of fit in proportional odds model or baseline odds model. By the method in the section 2.4.1, we should use the Pearson residuals vs. Fitted values plot to check lack of fit. The first plot is the Pearson residuals vs. Fitted values for proportional odds model for categories "low" and "$medium$". The plot is showed below :



From those plot we can know there does not exist lack of fit in proportional odds model.

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



From those plot we can know there does not exist lack of fit in baseline odds model either. By comparing if there exist lack of fit in those two models, we do not know which model performance better.

In the next step, we need to compare the AIC of each model. The AIC of proportional odds model is 1179.817, the AIC of baseline odds model is 1170.583. The AIC of baseline odds model is lower than proportional odds model even though the parameter in baseline odds model is more than proportional odds model. From those information we can know the baseline odds model is better than proportional odds model at this step comparison.

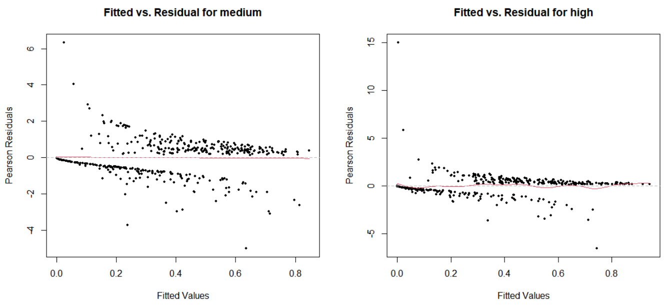
The final step is to use 10-fold cross validation for model evaluation. The table of accuracy of those two models in different i-th CV are showed below (ACC means accuracy) :

The i-th CV	1	2	3	4	5	6	7	8	9	10
ACC of Proportional	0.406	0.615	0.508	0.615	0.585	0.477	0.508	0.662	0.492	0.508
ACC of Baseline	0.531	0.569	0.569	0.492	0.538	0.554	0.554	0.615	0.569	0.615

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*".

3.2 Model Selection

After model selection by AIC, there are 16 predictors in our model. Its are ***1.school-binary 2.sex-binary, 3.mother's education-numeric, 4.Father's education-numeric, 5.Father's job-nomial, 6.reason to choose this school-nomial, 7.guardian-nomial, 8.weekly study time-numeric, 9.number of past class failures-numeric, 10.extra educational support-binary, 11.family educational support-binary, 12.extra-curricular activities-binary, 13.wants to take higher education-binary, 14.workday alcohol consumption-numeric, 15.current health status-numeric, 16.number of school absences-numeric***. We name this model as selected model. The the Pearson residuals vs. Fitted values plot for selected model is



From those two plots we can know that there is no lack-of-fit in this selected model. At present, this model is concise enough and also has great performance.

3.3 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 difference of log-likelihood between null model and the selected model by AIC is 386.4075. The P-Value is 0. This 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.

We can also do 10-fold cross validation for this selected model and make a comparison with the baseline odds model showed above, which built with all 30 predictors. The table of accuracy of those two models in different i-th CV are showed below (ACC means accuracy) :

The i-th CV	1	2	3	4	5	6	7	8	9	10
ACC of Baseline model	0.531	0.569	0.569	0.492	0.538	0.554	0.554	0.615	0.569	0.615
ACC of Selected model	0.500	0.585	0.569	0.553	0.569	0.569	0.523	0.677	0.600	0.615

The mean accuracy of baseline odds model that built with all 30 predictors is 0.5608. However, the mean accuracy of selected model that build with 16 predictors is 0.5762. A significance improve. This means selected model is much better than any previous models. The final table which contains estimated coefficients, standard errors and corresponding P-Values of selected model is showed below :

	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

4 Discussion

From the coefficient table, we can get some conclusions. ***In the model of Medium vs. Low***, 1.Students who are in Mousinho da Silveira(MS) school tend to have low Portuguese language grades against students who are in Gabriel Pereira(GP) tends to have medium grades. 2. Male student tends to have low Portuguese language grades against female students tends to get medium grades. 3.The students who father's educational level is higher tends to have medium grades. 4.Compared with students whose father's job is at home, the students whose father's job is health care related tends to have low grades. 5.Compared with students whose reason why he/she choose this school is course preference, the students whose reason why he/she choose this school is close to home tends to have medium grades. 6. Compared with students whose guardian is father, the students who guardian is mother tends to have low grade. 7. The students who number of past class failures are higher tends to have low grade. 8. Compared with students who do not have extra educational support, the students who have extra educational support tends to have low grade. 9.Compared with students who do not have extra curricular activities, students who have extra-curricular activities tends to have medium grade. 10. Compared with students who do not want to take higher education, the students who wants to take higher education tends to have medium grade. 11. Students who have higher number of school absences tends to have low grade. ***In the model of High vs. Low***, most factors have the same conclusion with the model of Medium vs. Low. The different factors are 1.Students who have longer study time tends to get high grade of Portuguese language. 2. The students who have workday alcohol consumption tends to have low grade of Portuguese language.

The limitation of this analysis is that we do not know the factors which may affect the grades between "Medium" and "High" and we can also find most of factors which can affect grades between "Medium vs. Low" and "High vs. Low" are the same. 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".

5 Reference

1. STA 223-LectureNotes-w21-1.pdf
2. STA 223-ClassProjects-w21.pdf
3. UCI Machine Learning Repository : http ://archive.ics.uci.edu/ml/datasets/Student+Performance#
4. 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.