Improving High School Math Education in Portugal in GLM Framework

Heqiao Ruan
Department of Statistics, University of California, Davis

Background

- Portugal, a country located in southwestern Europe
- Statistics keeps high school education quality in Portugal at tail end in Europe
- Especially serious problems in students' **failure rate and dropout rate**
- Officials want to improve math education
- Our goal: identify important factors influencing students' math grade
- Does they match common sense?
- Propose corresponding advice to help government to overcome difficulties

Dataset

- **Student Performance Dataset**
- Depicts students' achievement collected using school reports and questionnaires
- 395 observations,33 features
- Response: G1(First Stage),G2(2rd Stage), G3(Final Grade)
- 30 Predictors: Most Categorical
- Distribution of Grade Level(G3):

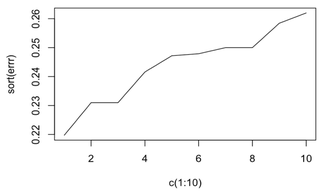
A(G3≥16)	$B(13 < G3 < 16)$	$C(11 < G3 < 14)$	$D(9 < G3 < 12)$	$F(G3 < 10)$
40	60	62	103	130

A-D Pass, F Fail

Method

- **Logistic Regression Model**: Fail(0,F)~Pass(1), 2-category classification /Probability of Passing on G3
- Prediction Error, 10-fold Cross Validation
- Model Selection by AIC

Predic/Level	Fail	Pass
Fail	60	25
Pass	70	240



Prediction Error 24.0% CV Error: 25.6%

Final Model after stepAIC:

$$Logit(E[Y|X]) \sim \beta_0+\beta_1 I(sexM)+\beta_2 age+\beta_3 I(Mjobhealth)+\beta_4 I(Mjobother)+\beta_5 I(Mjobservice)+\beta_5 I(Mjobteacher)+\beta_6 failures+\beta_7 I(schoolsupYes)+\beta_8 I(famsupYes)+\beta_9 I(higherYes)+\beta_{10} goout+\beta_{11} health+\beta_{12} failures * I(schoolsupYes)$$

Why add interaction? Schoolsupport effect contradict common sense

Method

Multinomial Model with 5-category: Proportional Odds Model, Baseline Odds Model on G3

Prediction Error:52.4% for Baseline Odds Model, 56.2% for Proportional OddsModel

Merge Categories

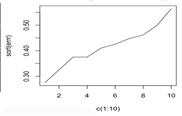
A:High, BCD:Medium, F:Low 3-category

Low(G3<10)	Medium($10 \leq G3 < 16$)	High($G3 \geq 16$)
130	225	40

Baseline Odds Model: Severe Lack of Fit. **Choose Proportional Odds Model**

Prediction Error:36.2% CV Error(10-fold):40.2% Performance Measurements:

Predic/Level	Low	Medium	High
Low	57	32	0
Medium	73	193	38
High	0	0	2



Accuracy	Precision	Recall	F-Score	F-Score($\beta = 0.5$)
0.638	0.649	0.703	0.676	0.683

Final Model after model selection :

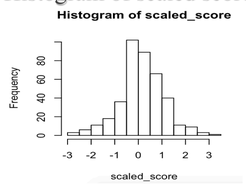
$$logit(P(Y \leq k)) = \beta_{0,k}+\beta_1 I(sexM)+\beta_2 age+\beta_3 I(PstatusT)+\beta_4 I(Mjobhealth)+\beta_5 I(Mjobother)+\beta_6 I(Mjobservice)+\beta_7 I(Mjobteacher)+\beta_8 studytime+\beta_9 failures+\beta_{10} I(schoolsupYes)+\beta_{11} I(famsupYes)+\beta_{12} I(higheryes)+\beta_{13} freetime+\beta_{14} goout+\beta_{15} health+\beta_{16} failures * I(schoolsupYes)$$

Transformed Model: Want to extract information from G1,G2 as well as G3

Perform PCA on cbind(G1,G2,G3)->Extract first principal component Y-> Scale Y to percent(0-100%) ->logit transformation on Y->Scaled Score

Scale-Score=logit(PERCENT(0.4629G1+0.5614G2+0.6859G3))

Histogram of scaled score: Approximately Normal, do **ordinary lm** on Scale-Score!



Summary of PCA: Table 10:PCA of G1,G2,G3:

Factor	PC1	PC2	PC3
Proportion of Variance	0.9095	0.06162	0.02892
Cumulative Proportion	0.9095	0.9711	1
G1	0.4629	0.8024	-0.3764
G2	0.5614	0.0632	0.8251
G3	0.6859	-0.5933	-0.4212

Final model after model selection:

$$E[Y|X] = E[ScaledScore|X] = \beta_0+\beta_1 sexM+\beta_1 I(Mjobhealth)+\beta_2 I(Mjobother)+\beta_3 I(Mjobservice)+\beta_4 I(Mjobteacher)+\beta_5 studytime+\beta_6 failures+\beta_7 I(schoolsupYes)+\beta_8 I(famsupYes)+\beta_9 I(higherYes)+\beta_{10} goout+\beta_{11} failures * schoolsupyes+\beta_{12} freetime+\beta_{13} health$$

Result(Important Predictors)

3.Logit Transformed Model after model selection

Predictors	Coefficient	Std.Error	t value	pvalue
sexM	0.251	0.092	2.735	6.53e-3
studytime	0.147	0.053	2.784	5.64e-3
failures	-0.418	0.0643	-6.506	2.5e-10
schoolsupyes	-0.465	0.137	-3.401	7.45e-4
famsupyes	-0.212	0.0872	-2.433	0.0154
romanticyes	-0.218	0.0890	-2.446	0.015
goout	-0.119	0.039	-3.098	0.002
health	-0.063	0.0299	-2.087	0.038
higher	0.750	0.632	1.186	0.236
failures:schoolsupyes	0.403	0.167	2.408	0.017

Result

1.Logistic Regression Model:

Predictors	Coefficient	Standard Error	z value	P Value
age	-0.217	0.108	-2.002	0.045
sexM	0.569	0.268	2.126	0.033
failures	-1.233	0.226	-5.460	4.76e-8
schoolsupyes	-1.334	0.385	-3.462	5.36e-4
goout	-0.346	0.114	-3.039	2.37e-3
higheryes	0.965	0.588	1.641	0.100
failures:schoolsupyes	1.412	0.475	2.982	2.87e-3

2.Proportional Odds Type Model after model selection

Predictors	Coefficient	Standard Error	z value	PVALUE
sexM	0.562	0.239	2.345	0.0195
age	-0.196	0.095	-2.065	0.396
failures	-1.278	0.221	-5.77	1.62e-8
schoolsupyes	-1.404	0.361	-3.888	1.19e-4
goout	-0.346	0.107	-2.047	0.041
higheryes	0.904	0.572	1.58	0.057
failures:schoolsupyes	1.478	0.447	3.309	1.027e-3

2(I):Intercepts

Prediction	Value	Std.Error	t value	pvalue
1|2	-4.706	1.863	-2.526	0.012
2|3	-1.163	1.842	-0.633	0.5270

Conclusion and Discussion

- Pay more attention to students who have fail before and provide extra support
- Pay more attention to Girls' study
- Arouse students' motivation to pursue higher degree, potentially help reduce dropout rate
- Extra Family support on math impose negative effect on students' grade. Avoid parents intervening
- Study more time, don't party too much

REFERENCES

[1].Paulo Cortez and Alice Silva. Using Data Mining to Predict Secondary School Student Performance. University of Minho Guimaraes, Portugal

[2].Hans-Georg Mueller. Generalized Linear Models Lecture Notes. UC Davis Winter 2018

Acknowledgement

I would like to express my thankfulness to Prof Hans Georg Mueller and TA Yaqing.