Improving High School Math Education in Portugal in GLM Framework

Heqiao Ruan Instructor: Hans-Georg Mueller

December 31, 2018

1 Abstract

Portugal is a country located in southwest Europe and its educational level has improved over the last decades. However the statistics keep the Portugal at the Europe's tail end in education due to its high student failure rate and dropping out rate in fundamental subjects such as Math and Portuguese in secondary school. So the official has realized this serious problem and it becomes pretty inspirational to identify what improve the students' math grade so that we can give corresponding advice to improve the math education in Portugal. Here we use the student performance dataset in UC Irvine machine learning dataset repository to conduct our research. Current research typically applied some machine learning algorithms such as ANN and SVM to perform binary/multi-label classification. We plan to extend their research by conducting techniques in generalized linear model and transformed model. Finally we will try to provide some insightful suggestions to improve middle school math education in Portugal standing in school's position.
For details of variables in the dataset,refer to Appendix A.

2 Background and Introduction

The Generalized Linear Model which is also named GLM is to examine the non-linear relationship between the response variables and predictors. The GLM has form $g(E[Y|X=x]) = \beta_0+\beta_1 X_1+..+\beta_p X_p = X\beta = \eta$ where g is the link function and $\eta = X\beta$ is the linear predictor. What's more, $\mu = E[Y]$ is another component. So it concludes the three components.
The dataset we use depicts the Portugal secondary school students' performance in math. It consists of 395 observations and 33 features. In the 33 features, the first 30 describes the students' status during the semester and the other three are the students' math grades for stage I, stage II and the final stage which is denoted as G1,G2 and G3 respectively. Here we apply various models in the framework of GLM and identify what influences the performance of the student in math exam. We point out that among the 30 predictors, most of them are categorical and for these predictors we use indicator variables to fit.

3 Methods

3.1 Logistic Regression

Previous research claims that the serious problem in Portugal education partly comes from the students' high failure rate in a couple of key subjects. So we fit the logistic regression model to identify what determines the students' failure rate in math. As we all know, logistic regression can be treated as a binary classification we also perform ten fold cross validation to examine the model's prediction ability as well checking the fit.

3.2 Multinomial Regression

Multinomial Regression stands for the GLM dealing with multiple response.As previous research states, students' grade are divided into five levels:A,B,C,D,F**(table[1])**. First we try to build two kinds of multinomial models. However, the prediction performance error of proportional odds model and baseline odds model is pretty high**(table[3] and table [4])**, we merge the BCD as medium and A as high and F as low to explore the relationship**(See table [5])**. Then we fit both baseline odds model and proportional odds models and compare their performance and decide a final model after performing model selection identifying the important elements decide students' performance.

3.3 Transformed Model

Based on the nature of the data we can see that the grade is divided into three categories: G1,G2 and G3.From previous research we know that it will be more efficient to predict G3 with G1 and G2. So we would like to extract information from the three components and we perform principal component analysis here and extract the first principal component which explains most of the variability in the data. As the first component represents more than 90% of the total variance(table[10]), this technique is well-rounded in this case.

We scale the combined grade(weighted average of G1,G2 and G3) into percentile which range from ϵ to $1-\epsilon$(continuous). Then we apply logit transformation on the combined grade which is denoted as Y for the distribution of $logit(Y) = log(\frac{Y}{1-Y})$ is more like normal distribution(**graph[2]**)**(It indeed help alleviate the heavy tail)**. Then we apply the general linear regression model(special case of GLM with identity link) which has the form $logit(Y) = \beta X$. Note that this is a transformed model, not a generalized linear model.

4 Main Results

First from the **Logistic Regression** model, we get the classification error of 24.4%**(Table 16)** which is a pretty decent classification.(after cross validation, still decent,**Graph 3(II)**) Then for the model fitting, we observe that **Age,Failures, Schoolsupport ,FamilySupport,Goout** are significant. We can see that students' passing chance are negatively related

to age,goout, familysupport,failures. However, as we all know, schoolsupport will highly different whether the students' fail before. So we add the interaction terms **Schoolsup:Failures** and observe that it is significant. Then from **Table 2** we can see that students' probability of fail in math exam is positively related to their **past failures, time of going out with friends** and negatively related to the students' age. Then from **Table 2** we also find the interaction effect between school support and failure is significant. What's more, outlier analysis including leverage value and cook distance can be seen in **(Table 15 and Graph 8)**
. **Then we fit the multinomial regression model**, if we fit the 5 category proportional odds model and baseline odds model respectively, the prediction error are 56.2% and 53.9% respectively.(**See Table 3 and Table 4**). So we merge the categories as three,(**See Table 5**), low, medium and high. So after fitting the proportional odds model and baseline odds model, the prediction errors are 36.2%,31.9%(see **Table6,Table7**) respectively. From **Graph 5** we can see that there are no obvious evidence of lack of fit in the proportional odds model while from **Graph 6** we can see that there is somewhat lack of fit in the baseline odds model(Pearson residuals are messy). What's more, various metrics measuring the multi-label classification performance is shown in **Table 1 (II)** which all indicates that the prediction fitting is fairly decent. Then we perform model selection by AIC criterion and the final model is shown in **Table2(I)**.
Then to compare the proportional odds model and the baseline odds model, we can see that in the baseline odds model none of the predictors in $Medium|Low$ is significant level 0.05, and the number of parameters in baseline odds model are pretty high,**(table 9)** what's more the baseline odds model shows obvious evidence of lack of fit**(Graph 6)**. **So we tend to prefer proportional odds type model to baseline odds model.(From Table 1(III),Graph 5,proportional odds type model fits decently)** Then we perform 10-fold cross validation to predict, the mean CV error is 41.6% which is normal for multi-label classification.**(Graph 3)**
So from the proportional odds type model we can see that **Age, Fjob, Failures, goouts, school support** are significant in determining the probability of getting a better math grade(Especially A). So we can see that the probability of students getting a better math grade is positively related to Fjobteacher and are negatively related to Age, Failures, Goouts. **For school support, its sign is not consistent to our common sense, so we add interaction term** $Failures : Schoosupport$ **and it's significant.** Then we can see that What's more we can also draw various conclusions. First, boys tends to perform better than girls and surprisingly we can see that students live in rural areas performs no worse than students living in urban areas, secondly, the students studying longer tend to get a better grade. What's more, students those mother has a higher education with internet access at home are more probable to get a higher math grade.
Then we want to include the influence of grade in the first two stages then we perform the logit transformed model. Then after fitting the model, we use AIC criterion to do model selection and from **Graph 7** we can see that no obvious pattern in residuals and the qq plot shows roughly normal which means a somewhat decent fit.Then we can see that **sexM, studytime, failures, schoolsupyes, famsupyes, goout,schoolsupyes:failures** are significant and the

interaction effect between school support:failure is significant as well. We can see that the result of transformed model is similar to the two previous models which means it is pretty meaningful to track students along the whole process.

5 Conclusion

The **significant predictors and the sign of coefficient** is shown as below in the three model(logistic regression, proportional odds model,logit transformed model) are shown as below.**(For detail, please refer to Appendix B)**

Logistic Model	Coef	ProportionOdd Model	Coef	Transform Model	Coef
sexM	0.569	sexM(BOY)	0.563	SexM(BOY)	0.006
failures	-1.233	Failures	-1.278	failures	-0.418
schoolsupyes	-1.334	schoolsupyes	–1.404	schoolsupyes	-0.465
goout	-0.346	goout	-0.346	goout	-0.1195
famsupyes	-0.615	age	-0.196	famsupyes	-0.2123
age	-0.217	health	0.161	studytime	0.1466
scsup:fail	1.429	scsup:fail	1.478	scsup:fail	0.403

We can see that the significant factors have the same sign and similar scale of coefficients which means the models are consistent in terms of significant predictors. It indeed validate the conclusion in [1] that predicting G3 will be more efficient with the information of G1 and G2.We can see that young men performs significantly better than young girls and young students tend to perform better in math **(It is indeed explanable that 15-16 is a adequate age for high school so older students may have difficulties in study even before high school)**. What's more, the students who fail more times before, go out to party too frequently tend to get a worse grade which is pretty straight forward to understand. Then for the influence of schoolsupport, we can see that for students with no failure history, students' grade are negatively related with school support(**Trivial by the definition of Indicator variable**) while for students with serious failure history, school support will have a better effect to students. Then some of the models indicate that healthier students are more probable to get a better grade which fits our common sense.
What's more, there are some other predictors we may particularly interest in which can help us to adjust the policy shown in the below table(Although may not that significant):

Logistic Model	Coef	ProportionOdd Model	Coef	TransformModel	Coef
higheryes	0.750	higheryes	0.874	higheryes	0.367
famsup	-0.4626	famsup	-0.439	famsup	-0.221
health	-0.148	health	-0.161	health	-0.062

From the above table we can see that students in a better health state are more likely to get a good grade. What's more, we can see that students willing to pursue a higher degree are more probable to get a good grade. It is pretty insightful that **math is an important**

prerequisite subject for most area in science and technology so the students willing to pursue a higher degree are more motivated so they will not only spend more time in studying(see **Table 12**) but also getting a higher grade**(see Table 13)** we can see that all students get A are willing to pursue a higher degree. Then for the influence of family support, we can see that the students' probability getting a better grade is negatively related to the family support and this is somewhat explanable as most parents are not experts in math education.**(See Table 14 which fit famsup individually)**.
So here we can see that the significant(important) predictors in all of the three models are similar.

6 Discussion and Corresponding Advice

In this investigation about **1/3 of all students fails in math which is one of the most important subject in high school** so we want to find what is significant in resulting to a higher grade and then propose corresponding advice to improve education quality in Portugal. From the point view of school, first we should guide students to arrange their time independently if they are competent **(school sup coef neg for no fail)** and provide help only for those who have failed before as well as prevent students from distracting from study during the process**(goout coef neg)**.More importantly, to decrease the failure rate of students and ensure everyone keep up with the course, we should particularly focus on those students who have failed before.Obviously,if you fail to follow math course this quarter, you will never understand it in the next quarter which will potentially lead to even higher failure rate in the future.For parents, we advice them not intervene on children's study as most of them they aren't expert in this area**(famsup coef neg,Table14)**. Another issue we have to point out is that in a well-developed country, university is one indispensable part of education so students should not only have motivation to pursue higher degree than secondary school but also the qualification to keep up with college level study. Then to arouse students' motivation to study math**(higheryes positive)** is pretty important too. What's more, as young men tend to perform better than young women, we should pay more attention to the girls' study especially for those have some difficulties in studying.**(SexM coef Positive)**. Finally we have to point out that the data we use to fit the model is somewhat limited for it ignores the differential influence of time. So further study may require the longitudinal type of data to repeatedly measure the students' performance and the predictors which may varies by time along the whole process of their study.

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

7 Appendix

7.1 Appendix A:Description of Datasets

Predictor Variables:
1.School- Student's School (Binary:"GP"-Gabriel Pereira or "MS"-Mousinho da Silveira)
2.Sex- Student's sex (Binary:"F"-female or "M"-male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: "U" - urban or "R" - rural)
5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)
9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "*athome*" or "other")
10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "*athome*" or "other")
11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12 guardian - student's guardian (nominal: "mother", "father" or "other")
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if $1 <= n < 3$, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very good to 5 - very bad)
30 absences - number of school absences (numeric: from 0 to 93)

Response Variables:
G1:Math Grade for first stage
G2:Math Grade for second stage
G3:Math Grade for final stage

7.2 Appendix B:Chosen Model for three types

1.Logistic regression Model:

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

2.Proportional Odds Type Model:

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

3.Logit Transformed Model:

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

For detailed coefficients, please refer to Appendix E.

7.3 Appendix C:Significant Predictors

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

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

7.4 Appendix D:Tables

Table 1: Distribution of levels

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

Table1(II):Measure of prediction performance of proportional odds type model with respect to merged categories:

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

Table1(III):Runs test for proportional odds type model of merged categories:

```
> runs.test(rdi_high)

        Runs Test - Two sided

data:  rdi_high
Standardized Runs Statistic = -1.1586, p-value = 0.2466

> runs.test(rpi_high)

        Runs Test - Two sided

data:  rpi_high
Standardized Runs Statistic = -1.2594, p-value = 0.2079

> runs.test(rdi_medium)

        Runs Test - Two sided

data:  rdi_medium
Standardized Runs Statistic = 0.45355, p-value = 0.6502

> runs.test(rpi_medium)

        Runs Test - Two sided

data:  rpi_medium
Standardized Runs Statistic = -0.15101, p-value = 0.88
```

Table 2(I): Full coefficient table of Logistic Regression Model after model selection:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)          5.55492    2.07353   2.679 0.007385 ** 
sexM                 0.56958    0.26795   2.126 0.033528 *  
age                 -0.21678    0.10828  -2.002 0.045281 *  
Mjobhealth           0.77209    0.57963   1.332 0.182844    
Mjobother           -0.18179    0.36658  -0.496 0.619949    
Mjobservices         0.72741    0.41422   1.756 0.079075 .  
Mjobteacher         -0.47103    0.44765  -1.052 0.292702    
failures            -1.23304    0.22583  -5.460 4.76e-08 ***
schoolsupyes        -1.33381    0.38524  -3.462 0.000536 ***
famsupyes           -0.46258    0.26588  -1.740 0.081891 .  
higheryes            0.96506    0.58799   1.641 0.100740    
goout               -0.34590    0.11381  -3.039 0.002372 ** 
health              -0.14809    0.09191  -1.611 0.107115    
failures:schoolsupyes 1.41762   0.47541   2.982 0.002865 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3:Proportional Odds Model(5 Category) Prediction Result

Prediction/Level	A	B	C	D	F
A	7	5	2	3	1
B	16	24	10	10	4
C	0	0	0	0	0
D	16	27	27	56	39
F	1	4	23	34	86

Table 4:Baseline Odds Model(5 Category) Prediction Result

Prediction/Level	A	B	C	D	F
A	21	7	2	4	1
B	7	28	9	12	9
C	1	6	14	7	8
D	7	9	17	48	25
F	4	10	20	32	87

Table 5: Merged Categories

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

Table 6:Merged Proportional odds Model Prediction:

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

Table 7:Merged Baseline Odds Model Prediction:

Predic/Level	low	Medium	High
Low	63	27	1
Medium	67	193	26
High	0	5	13

Table 8:Merged Proportional Odds Model after model selection:

	Value	Std. Error	t value	PVALUE
sexM	0.5623557	0.23980239	2.3450795	1.953820e-02
age	-0.1960426	0.09493438	-2.0650328	3.959949e-02
PstatusT	-0.5747717	0.35072678	-1.6388019	1.020843e-01
Mjobhealth	0.8086347	0.47814796	1.6911810	9.162431e-02
Mjobother	-0.1971477	0.33410197	-0.5900824	5.554870e-01
Mjobservices	0.9190661	0.36646843	2.5078999	1.256228e-02
Mjobteacher	-0.1408073	0.40782700	-0.3452624	7.300887e-01
studytime	0.2214303	0.13714921	1.6145209	1.072467e-01
failures	-1.2778355	0.22131517	-5.7738271	1.615042e-08
schoolsupyes	-1.4039719	0.36112172	-3.8878081	1.194119e-04
famsupyes	-0.4383335	0.23090447	-1.8983325	5.841172e-02
higheryes	0.9040662	0.57168980	1.5813930	1.146224e-01
freetime	0.1858861	0.11710992	1.5872786	1.132835e-01
goout	-0.3463535	0.10666424	-3.2471379	1.269591e-03
health	-0.1608674	0.07860043	-2.0466475	4.138206e-02
failures:schoolsupyes	1.4775380	0.44654813	3.3087990	1.026689e-03
1|2	-4.7060509	1.86272531	-2.5264331	1.192908e-02
2|3	-1.1662618	1.84196615	-0.6331613	5.270102e-01

Table 9:Baseline Odds Model(We ignore this model due to severe lack of fit):

(Intercept)	4.58362118	6.269436e-06	-10.94312390	0.0000000000
schoolMS	0.24737426	8.047565e-01	0.07045897	0.9438666257
sexM	0.46728876	6.405699e-01	0.77622095	0.4381171248
age	-0.24232833	8.086608e-01	-0.22082475	0.8253513563
addressU	0.31105183	7.559373e-01	-0.86530717	0.3874339513
famsizeLE3	0.10179867	9.189719e-01	0.43674153	0.6625549525
PstatusT	-0.53891908	5.902686e-01	-0.67917484	0.4974540256
Medu	0.02700642	9.784693e-01	0.74426068	0.4571937551
Fedu	0.15287676	8.785793e-01	-0.41261896	0.6801260881
Mjobhealth	0.46518106	6.420769e-01	0.79692363	0.4260092315
Mjobother	-0.31661063	7.517186e-01	-0.90162694	0.3678447884
Mjobservices	0.38457061	7.007777e-01	1.55723716	0.1202737404
Mjobteacher	-0.87951071	3.796984e-01	-0.90831459	0.3643064810
Fjobhealth	-0.18182527	8.558200e-01	0.93059982	0.3526706731
Fjobother	0.13071636	8.960711e-01	-0.06167560	0.9508546571
Fjobservices	-0.05621325	9.552024e-01	-0.25308321	0.8003452330
Fjobteacher	0.36884464	7.124556e-01	2.53042415	0.0118086660
reasonhome	0.28085369	7.789805e-01	0.58058047	0.5618785883
reasonother	0.33304118	7.392929e-01	0.42719634	0.6694861686
reasonreputation	0.50900033	6.110570e-01	0.26454778	0.7915059582
guardianmother	-0.08830207	9.296846e-01	-0.72307686	0.4700919795
guardianother	0.11987435	9.046481e-01	1.03981192	0.2991099729
traveltime	0.11225697	9.106809e-01	-0.34150253	0.7329202177
studytime	0.19401019	8.462749e-01	0.35721050	0.7211389868
failures	-1.09482063	2.743113e-01	-3.10179403	0.0020719934
schoolsupyes	-1.17825789	2.394551e-01	-3.43449575	0.0006613172
famsupyes	-0.63168210	5.279868e-01	-0.61313037	0.5401688827
paidyes	0.39371663	6.940183e-01	-1.14543582	0.2527730193
activitiesyes	-0.15250935	8.788688e-01	-0.60960252	0.5425013335
nurseryyes	-0.45366562	6.503369e-01	0.56047009	0.5754998811
higheryes	0.62012420	5.355598e-01	14.03167383	0.0000000000
internetyes	0.13872557	8.897429e-01	1.49103779	0.1368082297
romanticyes	-0.23149785	8.170567e-01	-0.67681433	0.4989489631
famrel	0.15941320	8.734308e-01	0.32870540	0.7425653398
freetime	0.12155331	9.033191e-01	0.13800656	0.8903107160
goout	-0.50608123	6.131027e-01	-0.39721859	0.6914365538
Dalc	-0.04955655	9.605027e-01	-0.20889932	0.8346424198
Walc	0.22800481	8.197691e-01	0.19654899	0.8442889900
health	-0.10216668	9.186800e-01	-0.38282725	0.7020688773
absences	-0.01130261	9.909881e-01	-0.05874724	0.9531853203
failures:schoolsupyes	1.33983821	1.811247e-01	-17.63425728	0.0000000000

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

Table 11:Logit Transformed model after AIC model selection:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.149374	0.360263	-0.415	0.678655	
sexM	0.251252	0.091857	2.735	0.006531	**
famsizeLE3	0.141826	0.091605	1.548	0.122419	
Medu	0.083697	0.052076	1.607	0.108859	
Mjobhealth	0.369433	0.204967	1.802	0.072291	.
Mjobother	-0.061151	0.131829	-0.464	0.643015	
Mjobservices	0.232519	0.148542	1.565	0.118353	
Mjobteacher	-0.201697	0.192928	-1.045	0.296494	
Fjobhealth	0.075793	0.269443	0.281	0.778642	
Fjobother	-0.150223	0.191309	-0.785	0.432816	
Fjobservices	-0.049219	0.198940	-0.247	0.804731	
Fjobteacher	0.328818	0.242108	1.358	0.175240	
studytime	0.146624	0.052661	2.784	0.005638	**
failures	-0.418060	0.064262	-6.506	2.5e-10	***
schoolsupyes	-0.464579	0.136612	-3.401	0.000745	***
famsupyes	-0.212256	0.087231	-2.433	0.015433	*
higheryes	0.367196	0.199698	1.839	0.066748	.
romanticyes	-0.217933	0.089099	-2.446	0.014909	*
freetime	0.064686	0.044093	1.467	0.143208	
goout	-0.119513	0.038582	-3.098	0.002099	**
health	-0.062546	0.029975	-2.087	0.037602	*
absences	0.009724	0.005226	1.861	0.063566	.
failures:schoolsupyes	0.403040	0.167363	2.408	0.016518	*

Table 12: Students' willing to pursue higher degree vs studytime:

Highyes/studytime	1	2	3	4
No	12	8	0	0
Yes	93	190	65	27

Table 13: Students graded A vs Willing to pursue higher degree:

Highyes/studytime	1	2	3	4
No	0	0	0	0
Yes	11	15	9	5

Table 14: Family Support coefficient with only predictor:

```
Call:
polr(formula = factor(level2) ~ famsup, data = stude_three)

Coefficients:
            Value Std. Error t value
famsupyes -0.2117     0.2032  -1.042

Intercepts:
    Value   Std. Error t value
1|2 -0.8430  0.1659    -5.0802
2|3  2.0586  0.2040    10.0920
```

Table 15: Outliers identified by leverage and cook's distance in logistic model:

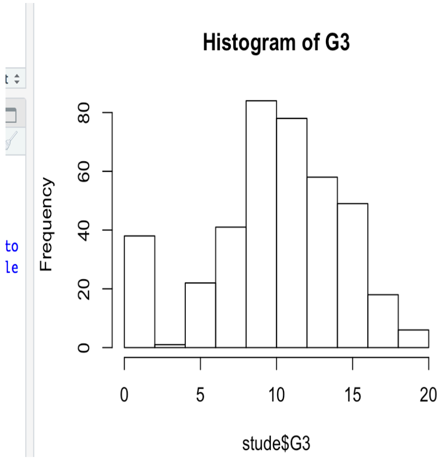
```
> intersect(index,index2)
[1]   3  62 184 277
```

Table 16:Logistic Regression Classification confusion matrix:

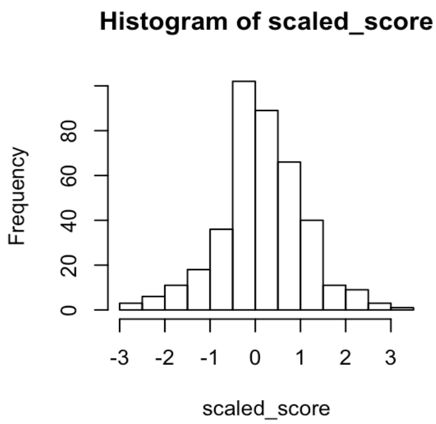
Predic/Level	Fail	Pass
Fail	60	25
Pass	70	240

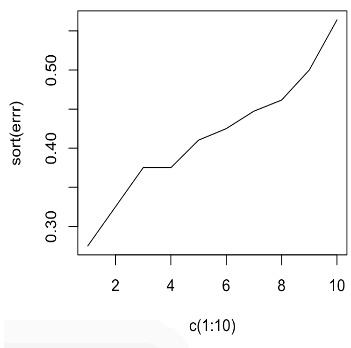
7.5 Appendix E:Graphs

Graph 1:Histogram of G3

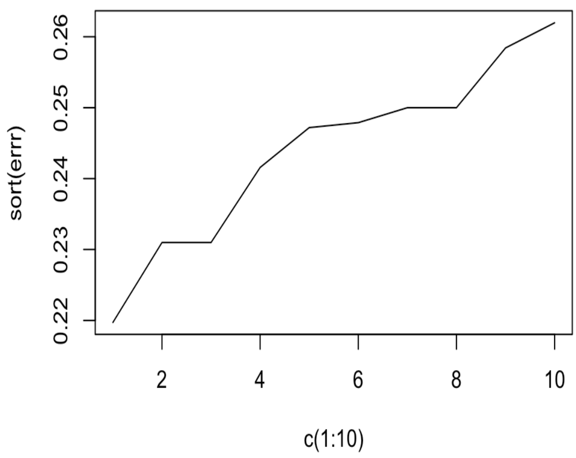


Graph 2:Histogram of Transformed First Principal Component



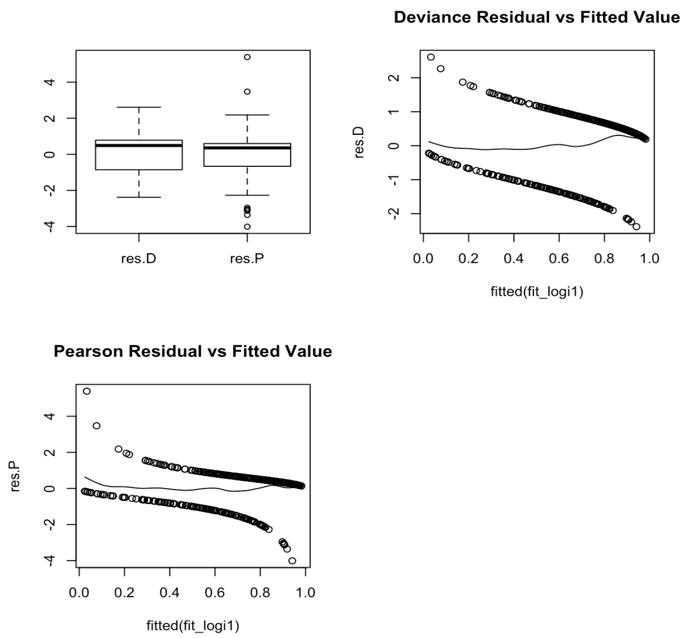


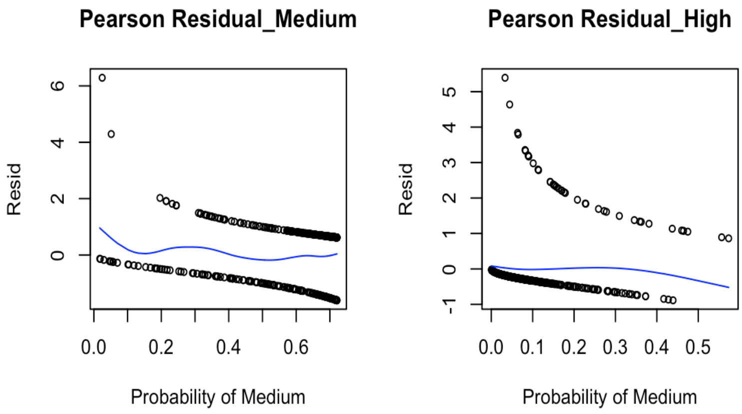
Graph 3:10-fold Cross validation error of proportional odds type model



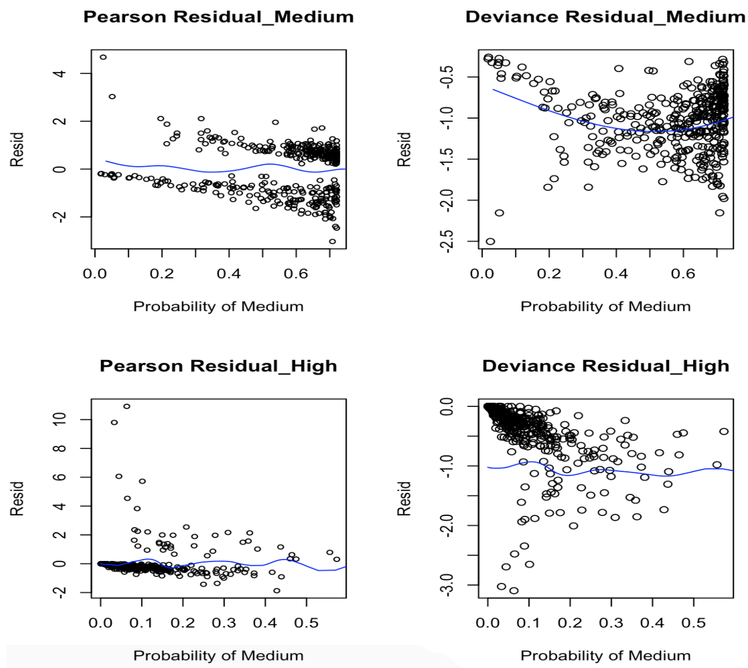
Graph 3(II):10-fold Cross validation error of logistic regression model

Graph 4:Diagnostic Plot for Logistic Regression

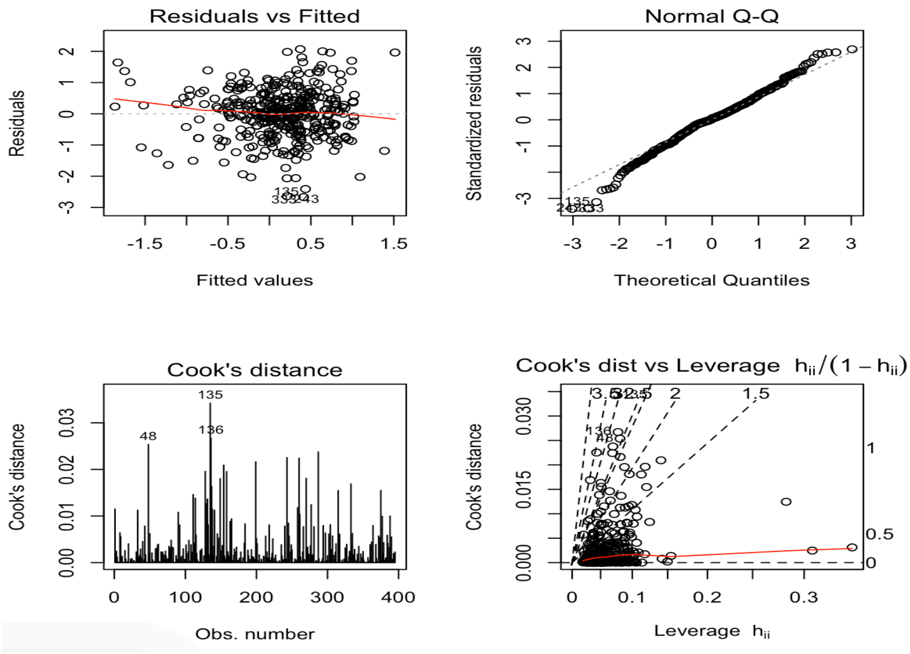




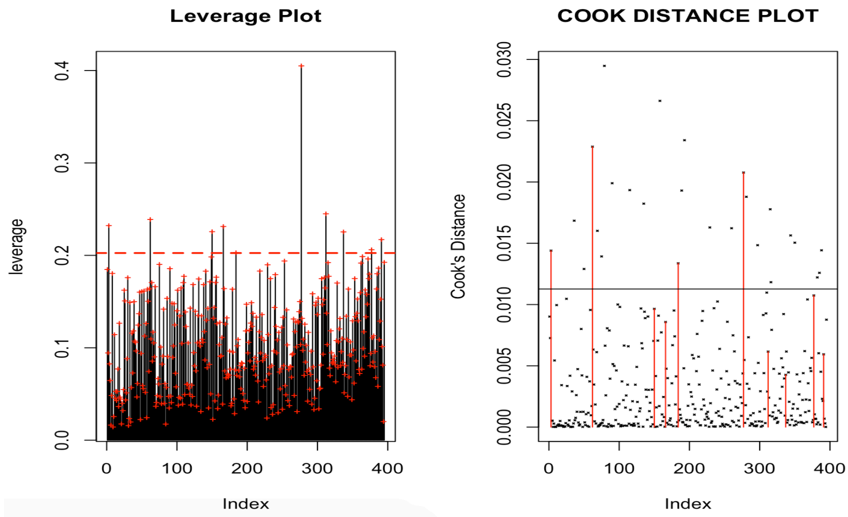
Graph 5:Diagnostic Plot for Merged Proportional Odds Model



Graph 6:Diagnostic Plot for Merged Baseline Odds Model



Graph 7:Diagnostic Plot for transformed linear model



Graph 8:Leverage and Cook Distance Plot in logistic model.