**USING LOGISTIC REGRESSION TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE**

Wid Sogata, Chris Ficklin, and James Tsai – Spring 2016, MSDS6372 – Experimental Statistics II

# KEYWORDS

Ordinal Logistic Regression

# INTRODUCTION

Analysis of student academic performance is often challenging. Multiple factors such as personal, socio-economic, psychological, and other environmental and non-environmental variables can potentially affect student success. For this study we will examine student performance collected in recent real-world data sets from two Portuguese secondary schools. Grade and attendance data from two core classes— Portuguese Language and Mathematics—was captured along with 29 questionnaire responses covering various demographic, social and school-related attributes. While there are many potentially informative facets to this topic, we aim to focus on aspects of the data that will help us model student motivation and the degree in which each of the attributes predict their final grade. To that end, it is our hope that our findings will add to the understanding of overall academic performance and perhaps assist administrators in improving academic policy and direction. Given Portugal’s high failure rate and low academic standing among European countries, this study is of particular importance.

# PROBLEM STATEMENT

Final grades and absences were recorded as well as study time and desire for higher education from a questionnaire. Our hypothesis is: reducing the number of absences, coupled with increased study time and the desire for higher education, will lead to improved odds of a student earning a higher grade. Grades in Portugal are assessed on a 20-point scale where 10+ is considered passing (<https://en.wikipedia.org/wiki/Academic_grading_in_Portugal>). The European Erasmus conversion standard for Portugal offers more granular classification as defined in the study “Using Data Mining to Predict Secondary School Student Performance” (Cortez et al. 2008). We will model the logistic regression based on the following classification:

* 5-Level classification – 1: excellent (16-20), 2: good (14-15), 3: satisfactory (12-13), 4: sufficient (10-11), 5: fail (0-9)

Logistic regression will be performed on both the Mathematics and Portuguese Language data sets. Analysis will be performed on the odds, maximum likelihood, and Wald confidence intervals in both datasets in order to identify and quantify the most relevant features.

# CONSTRAINTS AND LIMITATIONS

This analysis was completed on various students that were observed from 2 different schools. There are 649 Portuguese and 395 Math students that were picked randomly from these schools. There is no missing data. Given only 2 schools involved in the study, we cannot project the conclusion from this analysis to the general population. There are several (382) students that belong to both datasets. Searching for identical attributes can identify these students.

# DATA DESCRIPTION

There are 33 attributes for both Math (Math course) and Portuguese (Portuguese language course). The explanatory variables we are concerned with in our study are highlighted in yellow.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Usage** | **Description** | **Type** | **Range** |
| School | Identifier | Student's school | Binary | “GP” – Gabriel Pereira  “MS” – Mousinho da Silveira |
| Sex | Explanatory | Student's sex | Binary | “F” – female  “M” – male |
| Age | Explanatory | Student's age | Numeric | 15 to 22 |
| Address | Explanatory | Student's home address type | Binary | “U” – urban  “R” - rural |
| Famsize | Explanatory | Family size | Binary | “LE3” – less or equal to 3  “GT3” – greater than 3 |
| Pstatus | Explanatory | Parent's cohabitation status | Binary | “T” – living together  “A” – apart |
| Medu | Explanatory | Mother's education | Numeric | 0 – none  1 – primary (4th grade)  2 – primary (5th to 9th grade)  3 – secondary  4 – higher |
| Fedu | Explanatory | Father’s education | Numeric | 0 – none  1 – primary (4th grade)  2 – primary (5th to 9th grade)  3 – secondary  4 – higher |
| Mjob | Explanatory | Mother’s job | Nominal | “Teacher”  “Health” – care related  “Services” – administrative or police  “At home”  “Other” |
| Fjob | Explanatory | Father’s job | Nominal | Same as Mjob |
| Reason | Explanatory | Reason to choose this school | Nominal | Close to “home”  School “reputation”  “Course” preference  “Other” |
| Guardian | Explanatory | Student's guardian | Nominal | “Mother”  “Father”  “Other” |
| Traveltime | Explanatory | Home to school travel time | Numeric | 1 – “less than 15 min”  2 – “15 to 30 min”  3 – “30 min to 1 hour”  4 – “greater than 1 hour” |
| Studytime | Explanatory | Weekly study time | Numeric | 1 – “less than 2 hours”  2 – “2 to 5 hours”  3 – “5 to 10 hours”  4 – “greater than 10 hours” |
| Failures | Explanatory | Number of past class failures | Numeric | 1<=n<3, else 4 |
| Schoolsup | Explanatory | Extra Educational Support | Binary | Yes or No |
| Famsup | Explanatory | Family educational support | Binary | Yes or No |
| Paid | Explanatory | Extra paid classes within the course subject (Math or Portuguese) | Binary | Yes or No |
| Activities | Explanatory | Extra-curricular Activities | Binary | Yes or No |
| Nursery | Explanatory | Attended nursery school | Binary | Yes or No |
| Higher | Explanatory | Wants to take higher education | Binary | Yes or No |
| Internet | Explanatory | Internet access at home | Binary | Yes or No |
| Romantic |  | With a romantic relationship | Binary | Yes or No |
| Famrel | Explanatory | Quality of family relationships | Numeric | 1 – very bad to 5 – excellent |
| Freetime | Explanatory | Free time after school | Numeric | 1 – very low to 5 – very high |
| Goout | Explanatory | Going out with friends | Numeric | 1 – very low to 5 – very high |
| Dalc | Explanatory | Workday alcohol consumption | Numeric | 1 – very low to 5 – very high |
| Walc | Explanatory | Weekend alcohol consumption | Numeric | 1 – very low to 5 – very high |
| Health | Explanatory | Current health status | Numeric | 1 – very low to 5 – very high |
| Absences | Explanatory | Number of school absences | Numeric | 0 to 93 |
| G1 | Explanatory | First period grade | Numeric | 0 to 20 |
| G2 | Explanatory | Second period grade | Numeric | 0 to 20 |
| G3 | Response | Final grade | Numeric | 0 to 20, output target |
| Grade | Response | Final grade based on G3 | Numeric | 1 – excellent to 5 – fail |
| *Figure 1. List of all variables.* | | | | |

# EXPLORATORY DATA ANALYSIS

In Figure 2., we highlight our categorical response variable which contains 5-Levels based on the variable G3 for both Mathematics and Portuguese. We confirm the high frequency of low scores (Level 4 and 5) in both Mathematics and Portuguese. In Figure 3., we highlight the simple statistics for the variables that we are interested in our model. Note the binary variables are not included in this summary table. In Figure 4., we show the variables, which have a high Spearman’s correlation to the response variable G3 for Mathematics. Note that study time is highly correlated with G3, but absences are not in this analysis. Similarly, in Figure 5., we show the variables, which have a high Spearman’s correlation to the response variable G3 for Portuguese. Both study time and absences are significant in this analysis. In Figure 7., we show the Analysis of Maximum Likelihood Estimates for a backward selection for the explanatory variables that we are interested in for both Mathematics and Portuguese. Interestingly, only absences and higher explanatory variables were chosen for Mathematics, but absences, higher, and study time were all chosen for Portuguese. We proceed with only absences and higher explanatory variables for Mathematics, but we proceed with absences, higher, and study time for Portuguese.

|  |  |
| --- | --- |
| **Mathematics** | **Portuguese** |
|  |  |
| *Figure 2. Histograms of Levels (Response) for Mathematics and Portuguese.* | |
|  |  |
| *Figure 3. Simple statistics for Mathematics and Portuguese.* | |
| **Mathematics** | |
|  | |
| *Figure 4. Spearman Correlation for Mathematics.* | |
| **Portuguese** | |
|  | |
| *Figure 5. Spearman Correlation for Portuguese* | |
| **Mathematics** | **Portuguese** |
|  |  |
|  |  |
| *Figure 6. Backward Variable Selection.* | |

# ASSUMPTIONS

Before we proceed with logistic regression analysis, we make some assumptions about that data:

* The true conditional probabilities are a logistic function of the independent variables.
* The independent variables are measured without error.
* The observations are independent.
* The independent variables are not linear combinations of each other.

# LOGISTIC REGRESSION ANALYSIS

Following the variable reduction results, we perform logistic regression analysis for the five grade levels retaining the explanatory variables for absences and interest in perusing higher education. With the model convergence criterion satisfied and the Type 3 Analysis of Effects indicating that both remaining predictor variables add significant explanatory weight to the model, we next view the Analysis of Maximum Likelihood Estimates for each parameter. An important assumption for use of a reduced ordinal logistic model is that of proportional odds. In order to be met, the odds ratios can be assumed to be the same across all grade levels. This allows use of a reduced model in which the β (slope) coefficients for each explanatory variable are the same for all levels of the response. The Score Test for the Proportional Odds Assumption tests the hypothesis that there is significant evidence to indicate non-proportional odds. See Figure 7.

# Checking the Proportional Odds Assumption in Our Models

|  |  |
| --- | --- |
| **Mathematics** | **Portuguese** |
|  | |
| Full Model: - Because there are 5 levels of Grade  Reduced Model: – Assuming odds ratios are the same across levels of Grade. | Full Model: - Because there are 5 levels of Grade  Reduced Model: – Assuming odds ratios are the same across levels of Grade. |
|  |  |
| *There is not enough evidence to reject the proportional odds assumption for both Mathematics (p-value = 0.2123) and Portuguese (p-value = 0.2918).* | |
| *Figure 7. Score Test for the Proportional Odds Assumption.* | |

|  |  |
| --- | --- |
| **Mathematics** | **Portuguese** |
|  |  |
|  |  |
| *\* Note that the order of the grade levels as reported by SAS is reversed, with the lowest ordered value being the highest grade. These interpretations account for that and report the results using conventional terms rather implying that a grade below any fixed level is worse instead of better.* | |
| *Figure 8. Logistic Regression for Mathematics and Portuguese.* | |

# Interpretation for Mathematics

**Odds Ratios**

With preference for higher education held constant, each additional absence decreases the odds of earning a grade aboveany fixed level by .969 See *Figure 10* . Conversely, in the same condition, each additional absence increases the estimated odds of earning a grade below any fixed level by 3.2%.

For a given number of absences, a student having an interest in higher education (higher = 1) has estimated odds of earning a category grade higher than any fixed level times the estimated odds of students without a preference for higher education ( See *Figure 10.* Similarly, those same students with an interest in higher education enjoy 75.8% decreased odds of earning a grade below any fixed level, again holding absences fixed).

**Cumulative Probabilities**

Figure 9 shows the predicted probabilities of achieving either an excellent score (I) or passing (I, II, III, or IV) for each of four conditions: with and without an interest in higher education and at two quantities of absences, none and 4 (the median).

Several interesting things stand out. First is the drastically lower probabilities of success in either case for students who do not express an interest in higher education. In any scenario an interest in pursuing higher education accounts for a 3-4 times increase in probability of the success condition.

Also of note is the predicted effect of having either 0 or 4 absences. As expected, additional absences do to predict lower probabilities of success in all cases, but with the difference of effect in these examples ranging 0.4-2.6%, the practical significance should be considered. The school administrators may be better able to gauge the meaning of that finding.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mathematics** | **Absences** | **Higher = 0** | **Higher = 1** |
| - *Pass* | 0 | 0.3923 | 0.7271 |
|  | 4 | 0.3628 | 0.7015 |
| - *Excellent* | 0 | 0.0323 | 0.1212 |
|  | 4 | 0.0286 | 0.1084 |
| *Figure 9. Calculated probabilities for absences fixed at values 0 and 4 for passing and excellent grades.* | | | |

# Interpretation for Portuguese

**Odds Ratios**

As seen in the odds ratio estimate table (*figure 10*), with preference for higher education and study time held constant, each additional absence decreases the odds of earning a grade aboveany fixed level by a factor of .961.

For a given number of absences and study time, a student having an interest in higher education (higher = 1) has estimated odds of earning a category grade higher than any fixed level times the estimated odds of students without a preference for higher education (*.*

Finally, with a given preference for higher education and level of absences, additional study time increases the odds of earning a grade aboveany fixed level by a factor of 1.55*.*

# Wald’s Confidence Limits in Our Models

|  |  |
| --- | --- |
| **Mathematics** | **Portuguese** |
|  |  |
| *Figure 10. Odds Ratio Estimates.* | |

# CONCLUSION

We set out to create a model to judge whether motivation in students had a significant effect on their academic performance by counting the number of absences, coupled with increased study time and the desire for higher education. Interestingly, after the variable selection, only absences and desire for higher education were significant for Mathematics, but absences, desire for higher education, and study time were all significant for Portuguese. It is difficult to identify the reasons for this, but one explanation is that mathematics on average is a more difficult subject to comprehend and an incremental increase in study time doesn’t have a significant influence on performance. It was also interesting to note that the desire for higher education had a much greater influence on academic performance as compared to the number absences.

We saw this in the comparison of probabilities in Figure 9 for mathematics, in which there is a 1.85x increase in the odds of getting a passing grade for a student having desire for higher education with zero absences as compared to a student not having desire for higher education with zero absences.

Judging student academic performance is complex, and we have only studied a small facet of this complex subject. Based on this study, we did find that student motivation, based on desire for higher education and lack of absences does increase the probability that a student will receive a passing grade.

# APPENDIX LOGISTIC REGRESSION CALCULATIONS

Mathematics:

For Pass/Fail

4 absences, higher=0

= .0542

= .3628 - .1586 = .2042

🡨 failing grade

4 absences, higher = 1

0 absences, higher = 0

0 absences, higher = 1

|  |  |  |
| --- | --- | --- |
| - *Pass* | Higher= 0 | Higher = 1 |
| 0 absences | .3923 | .7271 |
| 4 absences | .3628 | .7015 |

For Grade I - Excellent

4 absences, higher=0

4 absences, higher = 1

0 absences, higher = 0

0 absences, higher=1

|  |  |  |
| --- | --- | --- |
| - *Excellent* | Higher=0 | Higher=1 |
| 0 absences | .0323 | .1212 |
| 4 absences | .0286 | .1084 |

Portuguese:

# APPENDIX SAS CODE

PROC IMPORT OUT=WORK.MAT

DATAFILE="/home/jjtsai0/MSDS6372/Project3/student-mat.csv"

DBMS=DLM REPLACE;

DELIMITER=';';

GETNAMES=YES;

DATAROW=2;

RUN;

PROC IMPORT OUT=WORK.POR

DATAFILE="/home/jjtsai0/MSDS6372/Project3/student-por.csv"

DBMS=DLM REPLACE;

DELIMITER=';';

GETNAMES=YES;

DATAROW=2;

RUN;

DATA WORK.MAT2;

SET WORK.MAT;

IF G3 >= 10 THEN PASS=1; ELSE PASS=0;

IF (G3 < 10) THEN GRADE = 5;

IF (G3 = 10 or G3 = 11) THEN GRADE = 4;

IF (G3 = 12 or G3 = 13) THEN GRADE = 3;

IF (G3 = 14 or G3 = 15) THEN GRADE = 2;

IF (G3 >= 16) THEN GRADE = 1;

RUN;

DATA WORK.POR2;

SET WORK.POR;

IF G3 >= 10 THEN PASS=1; ELSE PASS=0;

IF (G3 < 10) THEN GRADE = 5;

IF (G3 = 10 or G3 = 11) THEN GRADE = 4;

IF (G3 = 12 or G3 = 13) THEN GRADE = 3;

IF (G3 = 14 or G3 = 15) THEN GRADE = 2;

IF (G3 >= 16) THEN GRADE = 1;

RUN;

TITLE "MATHEMATICS: 5-LEVEL CLASSIFICATION";

PROC SGPLOT DATA=WORK.MAT2 NOBORDER;

VBAR LEVEL;

RUN;

TITLE "PORTUGUESE: 5-LEVEL CLASSIFICATION";

PROC SGPLOT DATA=WORK.POR2 NOBORDER;

VBAR LEVEL;

RUN;

PROC MEANS DATA=WORK.MAT;

RUN;

PROC MEANS DATA=WORK.POR;

RUN;

PROC CORR DATA=WORK.MAT SPEARMAN;

RUN;

PROC CORR DATA=WORK.POR SPEARMAN;

RUN;

PROC LOGISTIC DATA=WORK.MAT2;

CLASS HIGHER(REF="no") / PARAM=REF;

MODEL GRADE=ABSENCES HIGHER STUDYTIME / SELECTION=BACKWARD;

RUN;

PROC LOGISTIC DATA=WORK.POR2;

CLASS HIGHER(REF=”no”) / PARAM=REF;

MODEL GRADE=ABSENCES HIGHER STUDYTIME / SELECTION=BACKWARD;

RUN;

PROC LOGISTIC DATA=WORK.MAT2 PLOTS=ALL;

CLASS HIGHER(REF=”no”) / PARAM=REF;

MODEL GRADE=ABSENCES HIGHER;

OUTPUT PREDPROBS=L;

RUN;

PROC LOGISTIC DATA=WORK.POR2 PLOTS=ALL;

CLASS HIGHER(REF="no") / PARAM=REF;

MODEL GRADE=ABSENCES HIGHER STUDYTIME;

OUTPUT PREDPROBS=L;

RUN;