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 (4^{th} grade) 2 – primary (5^{th} to 9^{th} grade) 3 – secondary 4 – higher
Fedu	Explanatory	Father's education	Numeric	0 – none 1 – primary (4^{th} grade) 2 – primary (5^{th} to 9^{th} 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

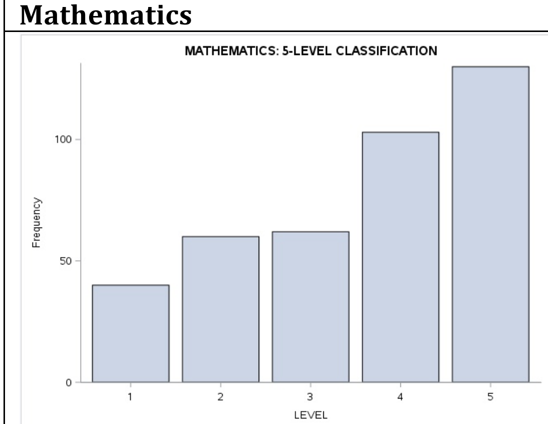
Dalc	Explanatory	Workday alcohol consumption	Numeric	1 – very low to 5 – very high
Walc	Explanatory	Weekend alcohol consumption	Numeric	1 – very low to 5 – very high
Health	Explanatory	Current health status	Numeric	1 – very low to 5 – very high
Absences	Explanatory	Number of school absences	Numeric	0 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

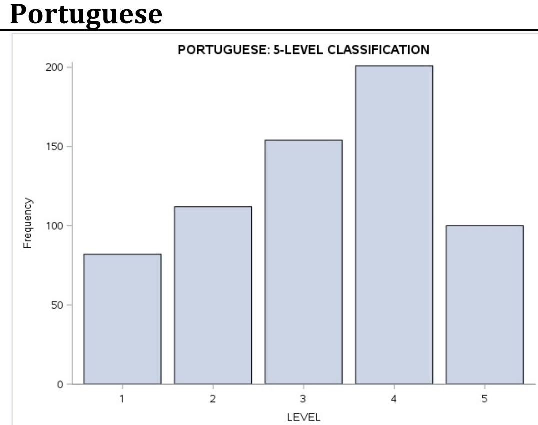


Figure 2. Histograms of Levels (Response) for Mathematics and Portuguese.

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
age	395	16.69620	1.27604	17.00000	15.00000	22.00000
Medu	395	2.74937	1.09474	3.00000	0	4.00000
Fedu	395	2.52152	1.08820	2.00000	0	4.00000
traveltime	395	1.44810	0.69750	1.00000	1.00000	4.00000
studytime	395	2.03544	0.83924	2.00000	1.00000	4.00000
failures	395	0.33418	0.74365	0	0	3.00000
famrel	395	3.94430	0.89666	4.00000	1.00000	5.00000
freetime	395	3.23544	0.99886	3.00000	1.00000	5.00000
goout	395	3.10886	1.11328	3.00000	1.00000	5.00000
Dalc	395	1.48101	0.89074	1.00000	1.00000	5.00000
Walc	395	2.29114	1.28790	2.00000	1.00000	5.00000
health	395	3.55443	1.39030	4.00000	1.00000	5.00000
absences	395	5.70886	8.00310	4.00000	0	75.00000
G1	395	10.90886	3.31919	11.00000	3.00000	19.00000
G2	395	10.71392	3.76150	11.00000	0	19.00000
G3	395	10.41519	4.58144	11.00000	0	20.00000

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
age	649	16.74422	1.21814	17.00000	15.00000	22.00000
Medu	649	2.51464	1.13455	2.00000	0	4.00000
Fedu	649	2.30663	1.09993	2.00000	0	4.00000
traveltime	649	1.56857	0.74866	1.00000	1.00000	4.00000
studytime	649	1.93066	0.82951	2.00000	1.00000	4.00000
failures	649	0.22188	0.59324	0	0	3.00000
famrel	649	3.93066	0.95572	4.00000	1.00000	5.00000
freetime	649	3.18028	1.05109	3.00000	1.00000	5.00000
goout	649	3.18490	1.17577	3.00000	1.00000	5.00000
Dalc	649	1.50231	0.92483	1.00000	1.00000	5.00000
Walc	649	2.28043	1.28438	2.00000	1.00000	5.00000
health	649	3.53621	1.44626	4.00000	1.00000	5.00000
absences	649	3.65948	4.64076	2.00000	0	32.00000
G1	649	11.39908	2.74527	11.00000	0	19.00000
G2	649	11.57011	2.91364	11.00000	0	19.00000
G3	649	11.90601	3.23066	12.00000	0	19.00000

Figure 3. Simple statistics for Mathematics and Portuguese.

Mathematics

Spearman Correlation Coefficients, N = 395 Prob > |r| under H0: Rho=0																
	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1.00000	-0.16129 0.0013	-0.14960 0.0029	0.10980 0.0291	0.03156 0.5317	0.23646 <.0001	0.03138 0.5340	0.00030 0.9952	0.14013 0.0053	0.09707 0.0539	0.13280 0.0082	-0.07515 0.1360	0.14928 0.0029	-0.05763 0.2532	-0.16762 0.0008	-0.17344 0.0005
Medu	-0.16129 0.0013	1.00000	0.63158 <.0001	-0.14785 0.0032	0.06350 0.2079	-0.24237 <.0001	0.01236 0.8065	0.02849 0.5723	0.06495 0.1977	0.02273 0.6525	-0.04433 0.3795	-0.03569 0.4794	0.09756 0.0527	0.20966 <.0001	0.23635 <.0001	0.22504 <.0001
Fedu	-0.14960 0.0029	0.63158 <.0001	1.00000	-0.15445 0.0021	0.01843 0.7150	-0.23662 <.0001	0.01140 0.8213	-0.01713 0.7343	0.04796 0.3417	0.00399 0.9369	-0.01449 0.7741	0.01811 0.7197	0.00357 0.9437	0.19474 <.0001	0.19484 <.0001	0.17005 0.0007
traveltime	0.10980 0.0291	-0.14785 0.0032	-0.15445 0.0021	1.00000	-0.10597 0.0353	0.07992 0.1128	-0.03866 0.4436	-0.02228 0.6589	-0.00143 0.9774	0.06648 0.1873	0.06365 0.2068	-0.01545 0.7595	-0.02506 0.6195	-0.08550 0.0897	-0.12380 0.0138	-0.12053 0.0165
studytime	0.03156 0.5317	0.06350 0.2079	0.01843 0.7150	-0.10597 0.0353	1.00000	-0.15763 0.0017	0.05814 0.2490	-0.13132 0.0090	-0.06598 0.1907	-0.21790 <.0001	-0.26402 <.0001	-0.09150 0.0693	-0.04618 0.3600	0.16229 0.0012	0.12916 0.0102	0.10517 0.0367
failures	0.23646 <.0001	-0.24237 <.0001	-0.23662 <.0001	0.07992 0.1128	-0.15763 0.0017	1.00000	-0.05139 0.3083	0.08806 0.0805	0.10542 0.0362	0.18749 0.0002	0.12791 0.0109	0.07969 0.1138	0.09603 0.0565	-0.34605 <.0001	-0.36236 <.0001	-0.36122 <.0001
famrel	0.03138 0.5340	0.01236 0.8065	0.01140 0.8213	-0.03866 0.4436	0.05814 0.2490	-0.05139 0.3083	1.00000	0.14314 0.0044	0.06355 0.2076	-0.10634 0.0346	-0.11606 0.0210	0.08534 0.0903	-0.08658 0.0857	0.02643 0.6004	0.00816 0.8715	0.05498 0.2757
freetime	0.00030 0.9952	0.02849 0.5723	-0.01713 0.7343	-0.02228 0.6589	-0.13132 0.0090	0.08806 0.0805	0.14314 0.0044	1.00000	0.28518 <.0001	0.19422 0.0001	0.13025 0.0096	0.08898 0.0774	0.01340 0.7907	0.00697 0.8901	-0.01677 0.7398	-0.00499 0.9212
goout	0.14013 0.0053	0.06495 0.1977	0.04796 0.3417	-0.00143 0.9774	-0.06598 0.1907	0.10542 0.0362	0.06355 0.2076	0.28518 <.0001	1.00000	0.25515 <.0001	0.39333 <.0001	-0.01854 0.7133	0.13328 0.0080	-0.15164 0.0025	-0.16099 0.0013	-0.16612 0.0009
Dalc	0.09707 0.0539	0.02273 0.6525	0.00399 0.9369	0.06648 0.1873	-0.21790 <.0001	0.18749 0.0002	-0.10634 0.0346	0.19422 0.0001	0.25515 <.0001	1.00000	0.63991 <.0001	0.09514 0.0589	0.12965 0.0099	-0.11144 0.0268	-0.11009 0.0287	-0.12094 0.0162
Walc	0.13280 0.0082	-0.04433 0.3795	-0.01449 0.7741	0.06365 0.2068	-0.26402 <.0001	0.12791 0.0109	-0.11606 0.0210	0.13025 0.0096	0.39333 <.0001	0.63991 <.0001	1.00000	0.09362 0.0630	0.20851 <.0001	-0.10837 0.0313	-0.10914 0.0301	-0.10446 0.0380
health	-0.07515 0.1360	-0.03569 0.4794	0.01811 0.7197	-0.01545 0.7595	-0.09150 0.0693	0.07969 0.1138	0.08534 0.0903	0.08898 0.0774	-0.01854 0.7133	0.09514 0.0589	0.09362 0.0630	1.00000	-0.07013 0.1642	-0.05222 0.3005	-0.05090 0.3129	-0.04779 0.3435
absences	0.14928 0.0029	0.09756 0.0527	0.00357 0.9437	-0.02506 0.6195	-0.04618 0.3600	0.09603 0.0565	-0.08658 0.0857	0.01340 0.7907	0.13328 0.0080	0.12965 0.0099	0.20851 <.0001	-0.07013 0.1642	1.00000	0.00448 0.9293	-0.03360 0.5055	0.01773 0.7254
G1	-0.05763 0.2532	0.20966 <.0001	0.19474 <.0001	-0.08550 0.0897	0.16229 0.0012	-0.34605 <.0001	0.02643 0.6004	0.00697 0.8901	-0.15164 0.0025	-0.11144 0.0268	-0.10837 0.0313	-0.05222 0.3005	0.00448 0.9293	1.00000	0.89479 <.0001	0.87800 <.0001
G2	-0.16762 0.0008	0.23635 <.0001	0.19484 <.0001	-0.12380 0.0138	0.12916 0.0102	-0.36236 <.0001	0.00816 0.8715	-0.01677 0.7398	-0.16099 0.0013	-0.11009 0.0287	-0.10914 0.0301	-0.05090 0.3129	-0.03360 0.5055	0.89479 <.0001	1.00000	0.95713 <.0001
G3	-0.17344 0.0005	0.22504 <.0001	0.17005 0.0007	-0.12053 0.0165	0.10517 0.0367	-0.36122 <.0001	0.05498 0.2757	-0.00499 0.9212	-0.16612 0.0009	-0.12094 0.0162	-0.10446 0.0380	-0.04779 0.3435	0.01773 0.7254	0.87800 <.0001	0.95713 <.0001	1.00000

Figure 4. Spearman Correlation for Mathematics.

Portuguese

Spearman Correlation Coefficients, N = 649 Prob > |r| under H0: Rho=0	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1.00000	-0.10229 0.0091	-0.11021 0.0049	0.06712 0.0875	0.01696 0.6663	0.29073 <.0001	-0.01937 0.6222	-0.00987 0.8019	0.13054 0.0009	0.08132 0.0384	0.09434 0.0162	-0.01805 0.6462	0.12426 0.0015	-0.16737 <.0001	-0.10559 0.0071	-0.06628 0.0916
Medu	-0.10229 0.0091	1.00000	0.64719 <.0001	-0.26329 <.0001	0.09842 0.0121	-0.20824 <.0001	0.02509 0.5235	-0.02789 0.4781	0.01021 0.7953	0.00196 0.9602	-0.01823 0.6429	0.01611 0.6820	-0.00601 0.8785	0.27640 <.0001	0.28564 <.0001	0.28393 <.0001
Fedu	-0.11021 0.0049	0.64719 <.0001	1.00000	-0.22203 <.0001	0.06908 0.0787	-0.16131 <.0001	0.02128 0.5883	-0.00015 0.9969	0.02879 0.4641	-0.00490 0.9009	0.02973 0.4497	0.04635 0.2383	0.03202 0.4154	0.23495 <.0001	0.24629 <.0001	0.23463 <.0001
traveltime	0.06712 0.0875	-0.26329 <.0001	-0.22203 <.0001	1.00000	-0.08939 0.0228	0.12361 0.0016	-0.02565 0.5142	-0.00105 0.9787	0.04071 0.3004	0.06846 0.0814	0.03152 0.4228	-0.06384 0.1042	0.02292 0.5599	-0.16623 <.0001	-0.16690 <.0001	-0.14695 0.0002
studytime	0.01696 0.6663	0.09842 0.0121	0.06908 0.0787	-0.08939 0.0228	1.00000	-0.16031 <.0001	0.01937 0.6223	-0.07650 0.0514	-0.08232 0.0360	-0.17131 <.0001	-0.22209 <.0001	-0.07673 0.0507	-0.11695 0.0028	0.27141 <.0001	0.25925 <.0001	0.27471 <.0001
failures	0.29073 <.0001	-0.20824 <.0001	-0.16131 <.0001	0.12361 0.0016	-0.16031 <.0001	1.00000	-0.05872 0.1351	0.10044 0.0105	0.04167 0.2892	0.10886 0.0055	0.06475 0.0994	0.04113 0.2954	0.12091 0.0020	-0.43243 <.0001	-0.43574 <.0001	-0.44836 <.0001
famrel	-0.01937 0.6222	0.02509 0.5235	0.02128 0.5883	-0.02565 0.5142	0.01937 0.6223	-0.05872 0.1351	1.00000	0.14412 0.0002	0.08778 0.0253	-0.09753 0.0129	-0.10203 0.0093	0.09254 0.0184	-0.10391 0.0081	0.02631 0.5034	0.05878 0.1347	0.04776 0.2244
freetime	-0.00987 0.8019	-0.02789 0.4781	-0.00015 0.9969	-0.00105 0.9787	-0.07650 0.0514	0.10044 0.0105	0.14412 0.0002	1.00000	0.35435 <.0001	0.12717 0.0012	0.12015 0.0022	0.09511 0.0154	-0.02848 0.4689	-0.10512 0.0074	-0.12096 0.0020	-0.12837 0.0010
goout	0.13054 0.0009	0.01021 0.7953	0.02879 0.4641	0.04071 0.3004	-0.08232 0.0360	0.04167 0.2892	0.08778 0.0253	0.35435 <.0001	1.00000	0.23398 <.0001	0.37245 <.0001	-0.01212 0.7579	0.10387 0.0081	-0.07822 0.0464	-0.11170 0.0044	-0.10497 0.0074
Dalc	0.08132 0.0384	0.00196 0.9602	-0.00490 0.9009	0.06846 0.0814	-0.17131 <.0001	0.10886 0.0055	-0.09753 0.0129	0.12717 0.0012	0.23398 <.0001	1.00000	0.61306 <.0001	0.08495 0.0305	0.10428 0.0078	-0.19848 <.0001	-0.20059 <.0001	-0.20839 <.0001
Walc	0.09434 0.0162	-0.01823 0.6429	0.02973 0.4497	0.03152 0.4228	-0.22209 <.0001	0.06475 0.0994	-0.10203 0.0093	0.12015 0.0022	0.37245 <.0001	0.61306 <.0001	1.00000	0.11428 0.0036	0.14510 0.0002	-0.15796 <.0001	-0.16999 <.0001	-0.17090 <.0001
health	-0.01805 0.6462	0.01611 0.6820	0.04635 0.2383	-0.06384 0.1042	-0.07673 0.0507	0.04113 0.2954	0.09254 0.0184	0.09511 0.0154	-0.01212 0.7579	0.08495 0.0305	0.11428 0.0036	1.00000	-0.01117 0.7764	-0.06313 0.1081	-0.09915 0.0115	-0.10567 0.0071
absences	0.12426 0.0015	-0.00601 0.8785	0.03202 0.4154	0.02292 0.5599	-0.11695 0.0028	0.12091 0.0020	-0.10391 0.0081	-0.02848 0.4689	0.10387 0.0081	0.10428 0.0078	0.14510 0.0002	-0.01117 0.7764	1.00000	-0.17043 <.0001	-0.16389 <.0001	-0.15851 <.0001
G1	-0.16737 <.0001	0.27640 <.0001	0.23495 <.0001	-0.16623 <.0001	0.27141 <.0001	-0.43243 <.0001	0.02631 0.5034	-0.10512 0.0074	-0.07822 0.0464	-0.19848 <.0001	-0.15796 <.0001	-0.06313 0.1081	-0.17043 <.0001	1.00000	0.89306 <.0001	0.88329 <.0001
G2	-0.10559 0.0071	0.28564 <.0001	0.24629 <.0001	-0.16690 <.0001	0.25925 <.0001	-0.43574 <.0001	0.05878 0.1347	-0.12096 0.0020	-0.11170 0.0044	-0.20059 <.0001	-0.16999 <.0001	-0.09915 0.0115	-0.16389 <.0001	0.89306 <.0001	1.00000	0.94445 <.0001
G3	-0.06628 0.0916	0.28393 <.0001	0.23463 <.0001	-0.14695 0.0002	0.27471 <.0001	-0.44836 <.0001	0.04776 0.2244	-0.12837 0.0010	-0.10497 0.0074	-0.20839 <.0001	-0.17090 <.0001	-0.10567 0.0071	-0.15851 <.0001	0.88329 <.0001	0.94445 <.0001	1.00000

Figure 5. Spearman Correlation for Portuguese

Mathematics

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	studytime	3	2	6.7228	0.0813

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
absences	1	5.8422	0.0156
higher	1	8.9478	0.0028

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-3.3999	0.4940	47.3638	<.0001
Intercept	2	1	-2.2792	0.4776	22.7751	<.0001
Intercept	3	1	-1.5429	0.4723	10.6732	0.0011
Intercept	4	1	-0.4377	0.4669	0.8788	0.3485
absences		1	-0.0314	0.0130	5.8422	0.0156
higher	yes	1	1.4177	0.4739	8.9478	0.0028

Portuguese

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
absences	1	6.3902	0.0115
higher	1	57.1552	<.0001
studytime	1	24.2030	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-4.5687	0.3265	195.7508	<.0001
Intercept	2	1	-3.4298	0.3108	121.8053	<.0001
Intercept	3	1	-2.3174	0.2999	59.6949	<.0001
Intercept	4	1	-0.4987	0.2836	3.0924	0.0787
absences		1	-0.0399	0.0158	6.3902	0.0115
higher	yes	1	1.9617	0.2595	57.1552	<.0001
studytime		1	0.4345	0.0883	24.2030	<.0001

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

<table>
<tr><th>Mathematics</th><th>Portuguese</th></tr>
<tr><td colspan="2">Ho: Reduced Model is Appropriate / Proportional Odds Assumption Appropriate
Ha: Need Flexible Odds per logit/The Proportional Odds Assumption is not Appropriate</td></tr>
<tr><td>Full Model: $\text{logit}[P(Y \le j|x)] = \alpha_j + \beta_j absences + \beta_5 higher$ – Because there are 5 levels of Grade

Reduced Model: $\text{logit}[P(Y \le j|x)] = \alpha_j + \beta_1 abscences + \beta_2 higher$ — Assuming odds ratios are the same across levels of Grade.</td><td>Full Model: $\text{logit}[P(Y \le j|x)] = \alpha_j + \beta_j absences + \beta_5 higher + \beta_6 studytime$ – Because there are 5 levels of Grade

Reduced Model: $\text{logit}[P(Y \le j|x)] = \alpha_j + \beta_1 abscences + \beta_2 higher + \beta_3 studytime$ — Assuming odds ratios are the same across levels of Grade.</td></tr>
<tr><td><table><tr><th colspan="3">Score Test for the Proportional Odds Assumption</th></tr><tr><th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr><tr><td>8.3690</td><td>6</td><td>0.2123</td></tr></table></td><td><table><tr><th colspan="3">Score Test for the Proportional Odds Assumption</th></tr><tr><th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr><tr><td>10.7699</td><td>9</td><td>0.2918</td></tr></table></td></tr>
<tr><td colspan="2">There is not enough evidence to reject the proportional odds assumption for both Mathematics (p-value = 0.2123) and Portuguese (p-value = 0.2918).</td></tr>
<tr><td colspan="2">Figure 7. Score Test for the Proportional Odds Assumption.</td></tr>
</table>

Mathematics

Response Profile

Ordered Value	grade	Total Frequency
1	1	40
2	2	60
3	3	62
4	4	103
5	5	130

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2029.122	1940.131
SC	2047.024	1966.983
-2 Log L	2021.122	1928.131

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-3.3999	0.4940	47.3638	<.0001
Intercept	2	1	-2.2792	0.4776	22.7751	<.0001
Intercept	3	1	-1.5429	0.4723	10.6732	0.0011
Intercept	4	1	-0.4377	0.4669	0.8788	0.3485
absences		1	-0.0314	0.0130	5.8422	0.0156
higher	yes	1	1.4177	0.4739	8.9478	0.0028

Portuguese

Response Profile

Ordered Value	grade	Total Frequency
1	1	82
2	2	112
3	3	154
4	4	201
5	5	100

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1212.810	1199.866
SC	1228.726	1223.739
-2 Log L	1204.810	1187.866

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-4.5687	0.3265	195.7508	<.0001
Intercept	2	1	-3.4298	0.3108	121.8053	<.0001
Intercept	3	1	-2.3174	0.2999	59.6949	<.0001
Intercept	4	1	-0.4987	0.2836	3.0924	0.0787
absences		1	-0.0399	0.0158	6.3902	0.0115
higher	yes	1	1.9617	0.2595	57.1552	<.0001
studytime		1	0.4345	0.0883	24.2030	<.0001

** 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 above any fixed level by .969 ($e^{-.0314} = 0.969$; 95% $C.I.$ [0.945, .994]) 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%($e^{.0314} - 1 = .032$).

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 4.13 times the estimated odds of students without a preference for higher education ($e^{1.4177} = 4.13$); 95% $C.I.$ [1.63, 10.45] 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 ($1 - e^{1.4177} = .758$).

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**
$\hat{P}(Y \leq IV)$ - *Pass*	0	0.3923	0.7271
	4	0.3628	0.7015
$\hat{P}(Y = I)$ - *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 above any fixed level by a factor of .961 ($e^{-.0399} = 0.961; 95\% C.I. [.932, .991]$).

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 7.11 times the estimated odds of students without a preference for higher education ($e^{1.962} = 7.11; 95\%\ C.I. [4.277,\ \ 11.826]$).

Finally, with a given preference for higher education and level of absences, additional study time increases the odds of earning a grade above any fixed level by a factor of 1.55 ($e^{-.4345} = 1.544; 95\% C.I. [1.299, 1.836]$).

Wald's Confidence Limits in Our Models

Mathematics	**Portuguese**

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
absences	0.969	0.945	0.994
higher yes vs no	4.128	1.630	10.450

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
absences	0.961	0.932	0.991
higher yes vs no	7.112	4.277	11.826
studytime	1.544	1.299	1.836

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

$$\hat{P}(Y = I) = \hat{P}(Y \le I) = \frac{e^{-3.399 - .0314(4)}}{1 + e^{-3.399 - .0314(4)}} = 0.0286$$

$$\hat{P}(Y \le II) = \frac{e^{-2.2792 - .0314(4)}}{1 + e^{-2.2792 - .0314(4)}} = .0828$$

$\hat{P}(Y = 2) = .0828 - .0286 = .0542$

$$\hat{P}(Y \le III) = \frac{e^{-1.5429 - .0314(4)}}{1 + e^{-1.5429 - .0314(4)}} = .1586$$

$\hat{P}(Y = III) = .1586 - .0828 = .0758$

$$\hat{P}(Y \le IV) = \frac{e^{-.4377 - .0314(4)}}{1 + e^{-.4377 - .0314(4)}} = .3628$$

$\hat{P}(Y = IV) = .3628 - .1586 = .2042$

$\hat{P}(Y \le V) = 1$

$\hat{P}(Y = V) = 1 - .3628 = .6372$ ← failing grade

4 absences, higher = 1

$$\hat{P}(Y \le IV) = \frac{e^{-.4377 - .0314(4) + 1.4177}}{1 + e^{-.4377 - .0314(4) + 1.477}} = .7015$$

$\hat{P}(Y = V) = 1 - .7015 = .2985$

0 absences, higher = 0

$\hat{P}(Y \leq IV) = \dfrac{e^{-.4377}}{1+e^{-.4377}} = .3923$

$\hat{P}(Y = V) = 1 - .3923 = .6077$

0 absences, higher = 1

$\hat{P}(Y \leq IV) = \dfrac{e^{-.4377+1.4177}}{1+e^{-.4377+1.477}} = .7271$

$\hat{P}(Y = V) = 1 - .7271 = .2729$

$\hat{P}(Y \leq IV)$ - *Pass*	Higher= 0	Higher = 1
0 absences	.3923	.7271
4 absences	.3628	.7015

For Grade I - Excellent

4 absences, higher=0

$\hat{P}(Y = I) = \dfrac{e^{-3.399-.0314(4)}}{1+e^{-3.399-.0314(4)}} = 0.0286$

4 absences, higher = 1

$\hat{P}(Y = I) = \dfrac{e^{-3.399-.0314(4)+1.4177}}{1+e^{-3.399-.0314(4)+1.477}} = .1084$

0 absences, higher = 0

$\hat{P}(Y = I) = \dfrac{e^{-3.399}}{1+e^{-3.399}} = .0323$

0 absences, higher=1

$\hat{P}(Y = I) = \dfrac{e^{-3.399+1.4177}}{1+e^{-3.399+1.4177}} = .1212$

$\hat{P}(Y = I)$ - *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;
```