Trey Sands
Tian Jiang
Tom Verghese
Kelsey Lucas
Spring 2010

There Are 'Major' Problems With This Data Set: Are Reed College Students Really This Predictable?

Introduction

Our assignment is to try to predict what majors Reed students will choose based on their admissions information and first year courses. The data we used ranged from the years 1989 to 2002 and included variables for SAT scores, ethnicity, gender, high school gpa, first year courses at Reed, grades in those first year courses, and other miscellaneous admissions information. With this data we were able to construct a model to predict major choices using the multinomial logit regression technique. We find that we are able to make accurate predictions about major choices with varying levels of certainty.

Data

We were given a very large and rather messy dataset from Reed's Office of Institutional Research. There was a fair amount of missing data, categorical variables, and unusable data. We had to drop numerous cases for various reasons.

First, we had to drop cases in which the students did not graduate from Reed. Since they did not graduate, we cannot accurately say what their major choices would have been.

We also had to drop cases in which people had double majors because of the constraints of our econometric estimation technique. Because of the nature of mlogit, we cannot code these students as belonging to two categories of majors. Since there were

only a few double majors in our dataset, less than 5% of our observations, we do not think they would affect our predictions significantly.

In addition, we dropped 2 cases of students majoring in a department that no longer exists, Medieval Studies. It does not seem appropriate to include majors in our model that no longer exist. Since we are supposed to be making predictions about what future students will major in, including majors that no longer exists will not help us make accurate predictions about major choices.

Finally, we dropped students who were a part of the Young Scholars program. These are students who take classes at Reed while they are still in high school. We were unsure of what constitutes their first year classes at Reed- the classes they took at Reed while in high school or their official freshman year classes once they were a regular student at Reed. Since we were unsure of how their classes were coded, and since there were less than 30 observations, we decided to drop these cases.

Model

Mlogit

To estimate major choice predictions we used the multinomial logit regression technique. This technique is most appropriate because it allows us to estimate the likelihood that a certain case (student) will choose one type of major over another.

We chose to group individual majors into categories of major types to use as our dependent variable. Trying to compare the probability that a student would choose one individual major over another would have problematic for a few reasons. First, if we had

to compare that many dependent variables[1] using mlogit, it is likely that our independent variables would have had less significant coefficients. Also, interpreting the results of an mlogit with so many dependent variables would be extremely difficult.

We have three different categorizations of majors. Note that the numbers in parenthesis following each category description correspond to the value assigned to that category in our mlogit results.

Dependent Variable Categories

The first categorization (our dependent variable maj_1) has four groups. The first group is The Arts (base outcome, 1). This group includes the following majors: art, music, theatre, dance-theatre, and literature-theatre. The second group is Social Sciences (2). This group includes the following majors: anthropology, sociology, psychology, economics, political science, ICPS, math-economics, and linguistics. The third group is Literature and Humanities (3). This group includes the following majors: history, philosophy, American studies, religion, classics, classics-literature, history-literature, Chinese, English, French, German, Russian, Spanish, and general literature. The fourth group is the Science and Math (4). This group includes the following majors: biology, chemistry, math, physics, chemistry-physics, BCMB and math-physics.

The second categorization (maj_2) has 5 groups. The Arts (base outcome, 1) and Science/Math (4) groupings are the same as in the maj_1 specification. The Social Science (2) category includes the following majors anthropology, sociology, psychology, economics, political science, ICPS, and math-economics. The Humanities (3) category

[1] There can only be one dependent variable in a regression. When using mlogit, we assign arbitrary values to the various possible outcomes of dependent variable (given the independent variables) to transform the categories we would like to compare into numerical values that we can estimate.

includes the following majors: history, philosophy, American studies, religion, classics, classics-religion, and history-literature. The final category is Literature and Languages. This group includes Chinese, English, French, German, Russian, Spanish, general literature, and linguistics.

The third categorization (maj_3) is the same as the second categorization (maj_2) except that English is in the Humanities group instead of the Literature and Language group.

We chose to define and test a few different major groupings to determine help determine the best model. We tried to group majors together using our intuition about what majors have enough characteristics in common to attract a similar type of student.

Results

We first needed to determine which of our three "Major" specifications yields the largest pseudo-R-squared. We found, after a multitude of models varying in their regressors, that our first division (1 = Arts, 2 = Behavioral and Social Sciences, 3 = Mathematics and Natural Sciences, 4 = Literature, Languages, and Humanities) yielded the greatest pseudo-R-squared over all of the different models by a difference of almost .02. This is a relatively significant increase in magnitude between the models. Because of this, we decided to continue using solely the first specification for the rest of our regressions.

Our first regression using the first "Major" division consisted of using a multinomial logit model with or maj_1 as the dependent variable and SAT verbal, SAT math, high school GPA, the reader rating, and the gender (1 = Male). We made our base outcome the Arts majors (group 1). Our results are in Table 1 below. We find that SAT verbal was only significant in determining whether the student was in group three or four,

where it features negatively in the former and positively in the latter. SAT math was only significant at a 5% level for group three. Both high school GPA and the reader rating were not significant. Being a male, however, made you more likely to not be anything other than an arts major. One important caveat to our results here, however. The pseudo –R-squared was only .0494.

Table 1	(1)	(2)	(3)	(4)
VARIABLES	1	2	3	4
SAT verbal	0	-0.000375	-0.00308***	0.00373***
	(0)	(0.00113)	(0.00112)	(0.00108)
SAT math	0	0.00225*	0.00900***	-0.00203*
	(0)	(0.00121)	(0.00121)	(0.00118)
High School GPA	0	-0.266	0.399*	-0.270
	(0)	(0.217)	(0.217)	(0.205)
Reader Rating	0	-0.100	-0.140	-0.155
	(0)	(0.164)	(0.162)	(0.157)
Gender (male = 1)	0	0.278*	0.623***	0.495***
	(0)	(0.167)	(0.165)	(0.161)
Constant	0	1.188	-3.733**	1.535
	(0)	(1.504)	(1.500)	(1.434)
Observations	3245	3245	3245	3245

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In the second regression we decided to report, we kept maj_1 as our dependent variable but also included dummy variables for the cumulative GPA of courses taken in respective divisions. We found out that SAT Math is no longer statistically significant, while SAT Verbal loses its significance in group 3 but increases in magnitude for group 4. Being male affects the probability of changing from group 1 to group 4 only at a 5% level. Taking a math or science course your freshman year signals (as shown by the statistically significant and positive values) that you are more quantitatively minded than

is required for the base outcome, the arts. Also, it is interesting to note that the effect of having a high cumulative GPA in the Math and Science division affects the probability of being in group three more than group two, and group two more than group 4, which goes with our intuition. Likewise, if one took any classes in the arts division their freshman year (theatre, art, music, or dance), then the probability of being any major other than an arts major decreases, as is seen by the negative and statistically significant coefficients. This pattern continues with the other divisions and their respective majors. One interesting thing to note is that the natural science variable is not significant. This may result from the relatively few number of observations including them. The pseudo-R-squared in this regression was .2830.

	(1)	(2)	(3)	(4)
VARIABLES	1	2	3	4
SAT Verbal	0	0.00229	0.00118	0.00440***
	(0)	(0.00151)	(0.00159)	(0.00142)
SAT Math	0	-0.000367	0.00164	-0.00292*
	(0)	(0.00159)	(0.00171)	(0.00153)
High School GPA	0	-0.366	-0.241	-0.444
	(0)	(0.282)	(0.315)	(0.273)
Reader Rating	0	-0.0277	0.0854	-0.0546
	(0)	(0.215)	(0.230)	(0.204)
Gender (Male = 1)	0	-0.00525	0.0584	0.441**
	(0)	(0.218)	(0.230)	(0.209)
Math and Science Cum. GPA	0	0.111***	0.387***	0.0640*
	(0)	(0.0361)	(0.0388)	(0.0352)
Arts Division Cum. GPA	0	-0.273***	-0.318***	-0.231***
	(0)	(0.0317)	(0.0363)	(0.0253)
HSS Cum. GPA	0	0.397***	0.123*	0.263***
	(0)	(0.0629)	(0.0710)	(0.0627)
PRPL Division Cum. GPA	0	0.243***	0.0151	0.166***
	(0)	(0.0427)	(0.0495)	(0.0418)
Lit. and Language Cum. GPA	0	0.0213	-0.0391	0.0997***

	(0)	(0.0327)	(0.0359)	(0.0312)
Natural Science Cum. GPA	0	0.0342	-0.442	0.292*
	(0)	(0.189)	(0.341)	(0.170)
Hum 110 Cum. GPA	0	0.0326	-0.127	0.193***
	(0)	(0.0758)	(0.0831)	(0.0746)
Constant	0	0.532	-0.791	0.441
	(0)	(1.953)	(2.098)	(1.858)
Observations	2587	2587	2587	2587

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The last regression we report still uses maj_1 as our dependent variable, but we decided to drop High School GPA and Reader Rating from the regression. We did, however, include each of individual dummy variables from the science department. Our reasoning behind this was that non-majors might be biased towards biology as compared to the more quantitative sciences. Moreover, Biology is the one science course that is not required for the other science majors to take. For instance, Biology majors must take both Biology and Chemistry, while Chemistry majors must take both Chemistry and Physics. By separating the overall dummy into its individual components, we hoped to be able to take this into account. We find that this separation was fruitful. The Biology coefficient was significant and positive across the board, meaning that those students who take biology are more likely to not be arts majors. The coefficient on the Math dummy shows us that if one takes Math, one is more likely to be a behavioral and social science major or a math and natural sciences major. The Physics and the Chemistry dummy are positive and only significant for the third group. One interesting result is the negative and significant coefficient on the Hum 110 dummy for the third group. Having a higher grade in Hum 110 means that you are less likely to be a Math and Natural Science major.

Our pseudo-R-squared in this regression was .3142.

	(1)	(2)	(3)	(4)
VARIABLES	1	2	3	4
SAT Math	0	0.000631	0.00393***	-0.00231*
	(0)	(0.00134)	(0.00148)	(0.00128)
SAT Verbal	0	0.00130	-0.00112	0.00365***
	(0)	(0.00124)	(0.00135)	(0.00118)
Gender (Male = 1)	0	-0.137	0.120	0.335*
	(0)	(0.180)	(0.193)	(0.171)
Hum 110 Cum. GPA	0	0.00688	-0.138**	0.156**
	(0)	(0.0627)	(0.0695)	(0.0615)
Biology Cum. GPA	0	0.173***	0.452***	0.0942**
	(0)	(0.0438)	(0.0466)	(0.0426)
Nat. Sci. Cum. GPA	0	0.0258	-0.147	0.0984*
	(0)	(0.0553)	(0.108)	(0.0503)
Art Division GPA	0	-0.277***	-0.313***	-0.241***
	(0)	(0.0263)	(0.0302)	(0.0215)
HSS Division Cum. GPA	0	0.380***	0.121**	0.232***
	(0)	(0.0490)	(0.0561)	(0.0487)
PRPL Division Cum. GPA	0	0.214***	-0.00678	0.134***
	(0)	(0.0353)	(0.0420)	(0.0342)
Lit. & Lang. Cum. GPA	0	0.0326	-0.0301	0.0983***
	(0)	(0.0279)	(0.0312)	(0.0266)
Math Cum. GPA	0	0.146***	0.397***	0.0504
	(0)	(0.0548)	(0.0566)	(0.0539)
Physics Cum. GPA	0	0.0692	0.312***	0.0489
	(0)	(0.0520)	(0.0527)	(0.0503)
Chem. Cum. GPA	0	0.0761	0.393***	0.0113
	(0)	(0.0593)	(0.0574)	(0.0579)
Constant	0	-0.757	-1.443	-0.852
	(0)	(0.931)	(1.030)	(0.893)
Observations	3306	3306	3306	3306

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This table represents our attempt to test the significance of our results. The percentages listed across the top of the table are various thresholds that we chose. The 50% column shows the percentage of students who had both over a 50% probability of choosing a major in that category relative to arts and actually graduated with that major. The historical column shows the percentage of students who graduated with majors in the given categories average over 5 years (2005-2009). We can compare our estimates to actual outcomes.

Based on our results, we underestimated the number of arts majors consistently, fluctuated around the actual number of social science major, science majors, and humanities and literature majors. This suggests are model makes seemingly accurate predictions with varying levels of certainty.

	50% level	30% level	20% level	10% level	0% level	Historical
1	0%	0%	0.6%	2.6%	4.5%	7.0%
2	45.8%	36.3%	32.8%	29.3%	25.9%	37.6%
3	16.4%	17.9%	21.5%	26.9%	30.1%	26.9%
4	37.8%	45.8%	45.1%	41.1%	39.6%	28.4%

Conclusion

We find that mathematics courses are relatively good predictors of whether a student will choose to be a math or science verses a social science major. Additionally, the course grade in Hum 110 is an important predictor of major choice. On the other hand, admissions data does not predict as well as other independent variables.

Possible Issues

Some possible concerns we have with our model include the following. First, we did have to drop several cases as discussed above in the Data section. Although we have

reasonable justification for dropping the cases we did, the more complete cases we can use, the better our model would be. Also, we predicted the probability of each case being a certain major using the predict command that does not generate results that allows us to test for significance. That is, we generate probabilities, but we have no way to test that the probabilities are significantly different from 0. Since we cannot test this using a common test technique, we compared our results to a few different thresholds including predictions of probabilities that were higher than 10%, 20%, 30%, and 50% respectively to determine how accurate our results are. It would be nice to have a more statistically sound technique to test the accuracy of our predictions.

Suggestions for Further Research

If researchers were interested in continuing this line of inquiry, we suggest the following. First, a more complete data set, over a longer period of time would likely yield more accurate predictions. Obviously, the nature of the data set is largely beyond the control of the econometrician, but institutions could try to do a better job of keeping more informed records in an easily accessible format. Also, we think variables for second year classes and grades, international students (high school outside of the U.S.), A.P./ I.B. test scores from high school, and quality of professor (based on student evaluations and tenure status) might help estimate a more accurate model for predicting major choice.

Second, for the researcher who is feeling ambitious, we think it would be interesting to try to define more categories for the dependent variable by creating more specific major groupings with fewer majors per category or by testing each major individually.

Another interesting test related to our project is to compare what students say they want to major in as freshman to what major they actually end up studying. This variable, freshman intended major, could be useful for the prediction model as well.

Appendix

We pretty much shared the work equally.

Stata output available upon request.