11

An Examination of Factors for Percentage of Students Attending College

Executive Summary

The problem that this report attempts to answer is if there is a reliable way to predict the percentage of students that a given school will send on to college. Increasing the participation percentage for the SAT and increasing the 10GMCAS scores has a significant positive effect on the overall percentage of students attending college after high school, as determined by backwards regression techniques. Similar variables also can help predict 4-yr vs. 2-yr university enrollment and public vs. private enrollment. Finally, this report analyzed significant differences in average SAT, 10GMCAS scores and other variables in schools that sent a large percentage of their students to the military, work or other.

Introduction

An investigation of the factors that increase or decrease the percentage of a school's students that attend college after high school was preformed, since one of the main purposes of high school is to prepare its students for college. For students that were going to college, this report tested to see if there was a proportional difference between the percentage of students that went to a 4-year college versus 2-year. This report also investigated differences between the percentages of students that attended public college versus private. If a difference between those results did exist, this report investigated to see if there were certain factors that affected the percentages of students that went to the respective types of college. For students that were not going to college, this report analyzed if there was a statistical difference between the schools that sent a higher percentage of their students into the various areas versus those schools that sent a lower percentage.

The statistical analysis that was used to determine the factors that effect percentages stated above was backwards step-wise logistic regression. A further discussion of the hypotheses and methods are discussed in the "Statistical Analysis" section.

Proportional hypothesis testing was used to determine if a significant difference existed between the proportions of students going to 4-yr. and 2-yr. school and public vs. private schools.

Hypothesis testing for equal means was used to determine if there was a significant difference between schools that sent higher percentages of students into military, work and other vs. schools that had lower percentages.

Summary Information

The data set consisted of 20 variables. Each variable is described below. There was no missing data.

Variable	Data Type	Data Points	Minimum	Average	Maximum
School	Descriptive	135	NA	NA	NA
Enrollment	Continuous	135	266	1103.4	3945
Cost/Pupil	Continuous	135	$4,675	$7,112	$12,586
AveTeach$	Continuous	135	$32,067	$46,963	$66,654
SATV	Continuous	135	387	512.64	609
SATM	Continuous	135	381	518.25	617
SATPartRate	Continuous	135	45.00%	80.78%	100.00%

10GMCAS Eng	Continuous	135	223	243.42	259
10GMCAS Mth	Continuous	135	221	241.83	257
S/TRatio	Continuous	135	9	13.133	19
S/Counsel Ratio	Continuous	135	113	224.52	389
Dropout Rate	Continuous	135	0	2.048	12.2
%College	Continuous Percentage	135	59.00%	82.90%	100.00%
%2YrPub	Continuous Percentage	135	0.00%	13.40%	40.00%
%4YrPub	Continuous Percentage	135	10.00%	27.10%	56.00%
%2YrPri	Continuous Percentage	135	0.00%	3.30%	36.00%
%4YrPri	Continuous Percentage	135	11.00%	39.10%	75.00%
%Military	Continuous Percentage	135	0.00%	2.10%	10.00%
%Work	Continuous Percentage	135	0.00%	9.20%	30.00%
%Other	Continuous Percentage	135	0.00%	5.80%	33.00%

For the purposes of analyses, other variables were created using the raw data from above. When those variables are used, a discussion of what they are and why they are used will be included.

Statistical Analysis

The underlying assumption to all hypothesis testing in this analysis is that the response variables are normally distributed. The Kolmogorov-Smirnov method, used for all variables in this analysis, compares the distribution of the test variable to a generic normal distribution, and tests for large absolute differences. If no difference is larger than a certain size, then the hypothesis that test variable distribution is not normally distribution is rejected. All variables used in this analysis rejected this hypothesis.

All testing in this analysis used principles of the normal distribution to test whether the observed data could come from a test distribution. For example, say a sample had an average value of 100 +/- 20 units. To test whether that sample could have come from a population with an average of 0 +/- 20 units, a test statistic is created that compares the two. If the test statistic is too high or too low, we reject the hypothesis that the observed values could have come from the test distribution. To quantify that, we use a p-value. The p-value is the percent chance that the observed data (or more extreme data) could have come from the test distribution. The cutoff point for rejection in this analysis is .05 level (or <.05 chance that the hypothesis is true).

For the first analysis, this report examined which factors would be useful in predicting what percentage of a school's students would go to college. First, we need to identify the response factor. The variable we are trying to model is %College. We have a problem modeling this variable in that percentages do not lend themselves very well to linear models since you can get results that are greater than 100% or less than 0%, which do not make sense.

In order to get a variable that we can model, we will use ln (p/(1-p)). The value of this variable has a range of potential outcomes that is the entire number line, making it more useful for analytical purposes. There was one data point, where the percentage was 100%. This data point was altered to 99.9% for this analysis.

Now that we have our response variable, we can perform a backwards step-wise regression. The main function of this type of analysis is to start with a model that includes all potential variables and remove them one by one until there are no more variables that meet a pre-set criteria.

The variables we started with were: Enrollment, Cost/Pupil, AveTeach$, SATV, SATM, SATPartRate, 10GMACENG, 10GMACMth, S/TRatio, S/CounselRatio and DropoutRate.

The results of the full model look like this:

Predictor	Coef	SE Coef	T	P
Constant	-5.411	3.074	-1.76	0.081
Enrollment	0.00005919	0.00009775	0.61	0.546
Cost/Pupil	0.00002123	0.00004423	0.48	0.632
AveTeach$	-0.00001250	0.00001110	-1.13	0.263
SATV	-0.002472	0.003394	-0.73	0.468
SATM	0.001020	0.003277	0.31	0.756
SATPartRate	2.6084	0.5850	4.46	0.000
10GMCAS Eng	-0.00047	0.02096	-0.02	0.982
10GMCAS Mth	0.02449	0.02288	1.07	0.287
S/TRatio	0.02460	0.02624	0.94	0.350
S/Counsel Ratio	0.0000912	0.0008740	0.10	0.917
Dropout Rate	-0.01759	0.02988	-0.59	0.557

S = 0.478435 R-Sq = 49.0% R-Sq(adj) = 44.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	11	27.0604	2.4600	10.75	0.000
Residual Error	123	28.1547	0.2289		
Total	134	55.2151			

From here, we will examine the hypothesis:

H_0: The overall model does not fit the data vs. H_a: The model fits the data.

Since the response variable is normally distributed, properties exist that allow us to use the F-distribution to determine if the observed data from the regression significantly fits the data. Based on a p-value of 0.000, the data rejects the null hypothesis above.

How well it fits the data can be determined by looking at the "R-sq" percentage located above the ANOVA table. If a model perfectly predicts the observed data, R^2 would be 100%. 49.0% is not as strong as some later analyses.

The hypothesis test that we want use to determine which variable to eliminate is one that will test whether the coefficient is H_0: $\beta_i = 0$ vs. H_a: $\beta_i <> 0$

For the example above, the p-values determine that the variable "10GMCAS Eng" is the variable that is least different from zero. As such, it is a good candidate to remove from the model. For this exercise, I repeated this process until the only variables left had p-values <0.05.

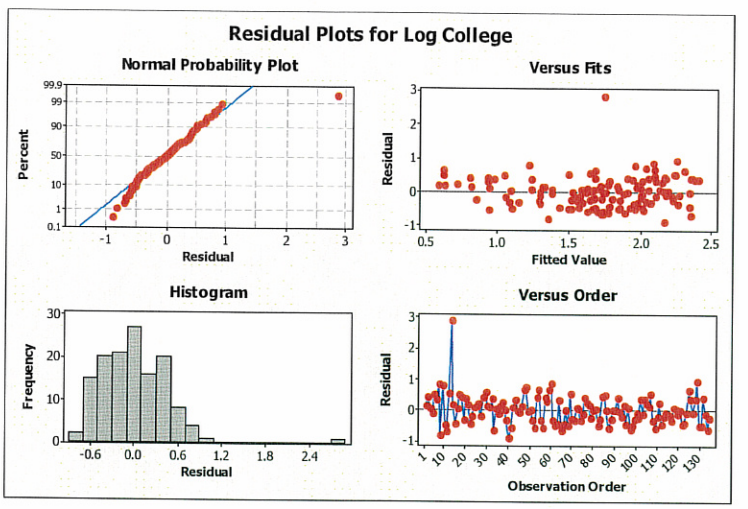
The final outcome looked like this:

Predictor	Coef	SE Coef	T	P
Constant	-4.931	1.970	-2.50	0.014
SATPartRate	2.3668	0.5366	4.41	0.000
10GMCAS Mth	0.019520	0.009562	2.04	0.043

S = 0.467158 R-Sq = 47.8% R-Sq(adj) = 47.0%

Analysis of Variance

Source	DF	SS	MS	F	P



Residuals are the differences between the actual values of the response variable and the predicted values. If the model is good, the upper left chart will have most of the red data points on the blue line. The upper right hand scatterplot and the lower right hand chart should have no discernable patterns to them. The lower left hand histogram should look normally distributed. With the exception of one data point (the 100% school from earlier), these diagnostic plots are all very good.

The results point to an increase in the percentage of students that take the SAT as a strong factor in determining the percentage of students that go into college. What it can not say is if this is a causal relationship.

Because of the translation we did to the percent of students that go to college, we can say that as SAT participation and 10GMCAS scores increase at a school, the proportion of students who attend college from that school tends to increase, but not linearly.

As a follow up, this report wanted to test for any factors that helped determine the SAT participation.

Using the same method as before, the data showed that the total GMCAS score provided a good predictor of SAT participation (r^2: 61.9%). As GMCAS scores go up, the SAT participation went up. This seems to make empirical sense because if students have success at a test, they would be more inclined to take a similar test at a later date.

The results for the 10GMCAS scores regression were three significant variables: Average Teacher Salary, Student/Counselor ratio and dropout rate.

For every point that the dropout rate increased, the average 10GMCAS score went down 4.4 points. For every $2,348.52 extra that teachers were paid in this sample, the average 10GMCAS score went up one point. The data showed that as the ratio of students to counselors increased, the average scores dipped (about 1pt. per 28 extra students on average). The overall r^2 of this regression was 53.5%, but these three variables indicate that there may be some proactive things schools can do if they want to increase the average 10GMCAS scores.

This report tested to see if there were differing factors present to predict when schools would send a higher percentage of their students to 2 or 4 year schools. This report also tested if there was a similar difference between public and private colleges.

First, this report tested to see if the data showed a significant difference between the proportion of college-bound students that went to 4-year schools and 2-year schools. The data shows that 77.6% of the 121,634 students that went to college went to a 4-year school. Therefore, 22.4% went to a 2-year college. Using the techniques described previously, the data says that this is significant. Because the sample size is so large, even the smaller difference between public (51.1%) and private (48.9%) schools is significant.

Using the same techniques, the data showed that for 4-year schools, the same two variables as colleges as a whole were significant. The only major difference between the two was that the r^2 for the 4-year regression was higher (79.2% vs. 61.9%).

For 2-year schools, the SAT participation rate was significant, but the total SAT score was a better predictor than the total 10GMCAS score. Another difference between the 4-year regression and the 2-year regression was that as the total SAT scores increased, the percentage of students going to 2-year colleges decreased. The same was true of SAT participation rate. Since the same variable was so prominent in the 4-year regression, this may be explained by students taking the SAT primarily going to 4-year schools instead of 2-year, thus driving down that percentage.

For public schools, the r^2 was only 40.5%. The 4 variables that finished as significant were enrollment, cost/pupil, average teacher salary and total SAT. For private schools, the same 4 variables tested as significant plus one more: SAT participation rate. The r^2 for the private school regression was 64.6%. Similarly to the results for the 2-year and 4-year schools, the predicting variables have opposite effects on the two response variables. Higher enrollment increases the percentages of students that attend public schools and decrease private. As cost/pupil increases, the percentage of students that attend private schools increases and public decreases. The same effect occurs with teachers salary and average total SAT score.

It would be easy to look at this data and conclude that if you paid your teachers nothing, a very large percentage of students at the school would attend public college. That is clearly not the case. The increase in public schools percentage due to lower teachers salaries is only evident because a decrease in salaries decreases the percentage of students that attend private schools.

Finally, this report analyzed if there were variables that were present in schools that had a "high" percentage of its students that went into the military, work or other vs. those that had a "low" percentage.

For military percentage, 88 schools were above the average and 47 were below. For the schools that were above the average, they had the following attributes compared to those that were below average:

Lower average SATV scores (487.5 vs. 526.1, p-value=.000)
Lower average SATM scores (489.5 vs. 533.6, p-value=.000)
Lower SAT participation rate (72.6% vs. 85.1%, p-value=.000)
Lower average 10GMCAS Eng scores (239.0 vs. 245.8, p-value=.000)
Lower average 10GMCAS Mth scores (237.1 vs. 244.4, p-value=.000)
Higher Dropout Rate (3.23 vs. 1.41, p-value=.000)

For schools that had a higher percentage of their students enter the workforce, the data showed similar traits (77 above the average, 58 below):

Higher enrollment (1236.1 vs. 1003.4, p-value=.019)
Lower cost/pupil ($6,797 vs. $7,349, p-value=.007)
Lower AveTeach$ ($45,895 vs. $47,768, p-value=.015)
Lower average SATV scores (491.6 vs. 528.5, p-value=0.000)
Lower average SATM scores (494.1 vs. 536.4, p-value=0.000)
Lower SAT participation rate (73.2% vs. 86.5%, p-value=0.00)
Lower average 10GMCAS Eng scores (239.8 vs. 246.2, p-value=0.000)
Lower average 10GMCAS Mth scores (237.7 vs. 245.0, p-value=0.000)
Higher dropout rate (2.96 vs. 1.36, p-value=0.000)

The data does not show nearly as big a difference between the schools that had a high percentage of students in the "other" category, but the overall trend is the same (54 above the average, 81 below):

Lower average SATV scores (502.2 vs. 519.6, p-value=.018)
Lower SAT participation rate (76.8% vs. 83.4%, p-value=.004)
Lower average 10GMCAS Eng scores (241.6 vs. 244.6, p-value=.014)
Lower average 10GMCAS Mth scores (239.9 vs. 243.1, p-value=.013)
Higher Student/Teacher Ratio (13.6 vs. 12.8, p-value=.015)

The conclusions that the data point to is that schools that send fewer of their students to college seem to have significant differences on the "bad" side of the statistical spectrum for some key variables. For further research, I would like to investigate further some of these differences in the schools to see if there is some kind of demographic factor that could account for this disparity.

Demographic and economic factors of the area in which the school is located may also help us to understand better some factors outside of the schools control that could be influential. Also, if the high school itself is public or private could be of interest, as is the average experience of the staff. Such factors as participation in sports and clubs by students could shed some light on what schools could encourage if the ultimate goal is to send students to college. Transformations of the independent variables (like enrollment^2, for instance) could be used also.

References

Kleinhaum, Kupper, Muller & Nizam, Applied Regression Analysis and Multivariate Methods, Duxbury Press, Pacific Grove, 1998, pp. 330-332, 395-396.

Riffenburgh, Robert H., Statistics in Medicine, Elsevier Academic Press, San Diego, 2006, pp.480-481.