Multiple Regression Modeling of *U.S. News and World Report* 2014 Liberal Arts Colleges Rankings & the Effect of SAT on Rank Prediction

Abstract

In this study, a multiple regression model was created to predict a school's ranking on *U.S. News*' list of top liberal arts colleges for 2014. An initial model was created using best subset regression. By using an extra sum of squares test, it was shown that it was not necessary to include SAT 75%, which is the 75th percentile SAT score that admitted students of a particular college received, in the regression model. Even though SAT 75% is conceivably an important factor influencing a school's ranking, due to its high correlation with other variables, it did not contribute significantly to the model. This study demonstrates the existence of multicollinearity in multivariable regression. Our findings also contribute to an understanding of how *U.S. News* chooses their explanatory variables to come up with each school's rank.

Research Question

Which variables should be included in the regression model to predict *U.S. News & World Report*'s ranking for liberal arts colleges in 2014? Does adding SAT 75% improve the regression model after other variables have been accounted for?

Background & Significance

As a public magazine, every year *U.S. News* releases a ranking of colleges based on undergraduate academic reputation, retention, faculty resources, student selectivity, financial resources, graduation rate performance, and alumni giving rate (Morse and Flanigan). Even though there are many critics of *U.S. News*' ranking, it has been shown that students actually pay close attention to rankings when choosing which college to attend. In a study done by UCLA, new students were asked which factors influenced them to choose their college, and 16.7% of students reported "Rankings in national magazines" was very important in their choice, (Higher Education Research Institute UCLA). Students also listed good academic reputation as the No. 1 factor in their school choice, and *U.S. News* indeed includes reputation as an important factor in producing the ranking list (Morse and Flanigan). In this study, we will construct a multiple regression model to predict a college's ranking on *U.S. News*' list and determine which variables to include. In particular we will conduct a test to determine the influence of SAT scores in the ranking after taking into account all other variables in the model.

Methods

Data Acquisition & Modification

Data were obtained from the *U.S. News & World Report* website. Additional variables, such as retention rate and graduation rate, were provided by the author's College Institutional Research Office. We were only interested in the *U.S. News* top 180 liberal arts colleges in this study. Some colleges on the list were eliminated in this study because of missing data or special circumstances. For example, schools such as United States Naval Academy that charge no tuition were eliminated from the list.

Statistical Analysis

The *U.S. News*' college ranking was used as the response variable. Fifteen potential explanatory variables were evaluated when developing a reduced model. A full list of variables and their descriptions are provided in the appendix. Using best subsets regression, a reduced model, excluding SAT 75%, with relatively high R^2 and low C_p was chosen as the preliminary model to predict ranking. Normal probability plot and residual plots of the preliminary model were created to check model assumptions as well as test whether interaction terms or variable transformations were needed.

To address our research question, a full model (reduced model with the addition of SAT 75% as an explanatory variable) was created. An extra sum of squares test was conducted to test whether or not including the SAT 75% significantly improved the regression model to predict ranking. Correlation coefficients were calculated to examine whether SAT 75% is highly correlated with other variables in the preliminary model in order to test whether multicollinearity existed in the full model. The final model was developed and compared to the *U.S. News*' ranking methodology.

Results

Best Subset Regression (Reduced Model):

Ranking = 543 - 0.00122 Tuition - 201 Retention Rate - 110 Graduation Rate + 15.3 Acceptance rate - 67.4 % class < 20 students - 0.107 SAT 25%

This model had an R^2 of 91.3%, R^2(adjusted) of 91.0%, C_p of 5.3 (very close to the number of variables in the model, which is desired) and p-values of each coefficient < 0.10.

Checking Model Assumption:

There were several obvious outliers, for which you can find explanations in the Appendix. Excluding the outliers, the residuals appeared to be approximately normally distributed, centered around 0 with equal variance (see Appendix).

Hypothesis Test (Full Model)

By adding SAT 75% to the reduced model and developed a full model. The full regression model predicted ranking as follows:

> Ranking = 558 - 0.00123 Tuition - 202 retention rate - 105 Graduation Rate + 15.2 Acceptance rate - 64.6 % class < 20 - 0.0713 SAT 25% - 0.0446 SAT 75%

We found SAT 75% coefficient p-value = 0.296, R-Sq = 91.4%, R-Sq(adj) = 91.0%. An extra sum of squares test was conducted, which uses an F-test to give an F-statistic that follows an F-distribution, to test whether SAT 75% significantly improves the model:

$$F = \frac{(SSR\ full - SSR\ reduced)/(k-p)}{MSE\ full} = \frac{(426462 - 426196)/(7-6)}{242} = 1.099$$

where k is the number of explanatory variables in full model and p is the number of explanatory variables in reduced model. The resulting p-value of F-statistic equals to 0.29597.

Conclusions

Comparing Full and Reduced Model

R^2 only improved by 0.1%, R^2 (adjusted) did not change after adding SAT 75% into the model. The extra sum of squares F-test determines whether the difference between the sums of squared residuals in the full and reduced models is so large that it is unlikely to occur by chance. Since the p-value of the F-statistic is much larger than 0.05, it was shown that SAT 75% did not significantly improve the model, leading us to conclude that the reduced model should be used as the final model.

Multicollinearity

Even though SAT 75% did not significantly improve the model to predict ranking (p-value = 0.296), that does not mean SAT 75% does not influence ranking. Further investigation shows that there is a very strong correlation between SAT 25% and SAT 75%. It is likely that SAT 75%'s influence on ranking is already contributed by including SAT 25% in the model. The regression equation is: SAT 75% = 342+0.878 SAT 25%, with R^2 = 93.2% and R^2 (adj) = 93.1%.

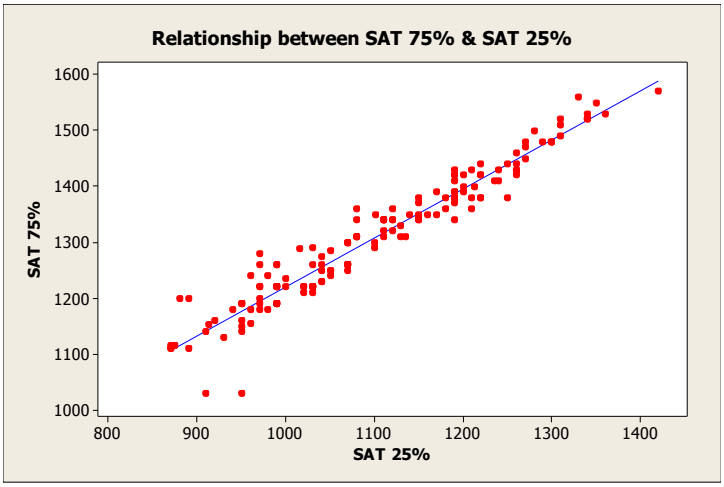

Figure 1. Relationship between SAT 75% and SAT 25%

The reduced model is a good model because it has a high R^2 (R^2 = 91.3%, R^2(adjusted) = 91.0%), which means the percentage of variation explained by the regression line is very high; model assumptions for regression are mostly met and the signs of each coefficient of explanatory variables are sensible.

Due to its high correlation with SAT 25%, we find no evidence that including SAT 75% significantly improves our model, even though it is a factor that is certainly related to the ranking. This result shows that if we know the 1st quartile of the admitted students' SAT score, we do not necessarily need the information of their 3rd quartile SAT score. The study also demonstrates that when we are looking at multivariable regression models, if a variable is not included in the model it does not mean it is not related to the response variable – it may just be highly correlated with other explanatory variables that are already in the model.

Discussion

In *U.S. News*' methodology, they used many more explanatory variables than we did to come up with the rankings. However, our model with only 6 explanatory variables still predicts the response variable well. Our conjectures for this phenomenon are:

1. Some of the variables in their model might not be necessary to include. For example, it might not be necessary to include student to faculty ratio and % of class < 50 students because for liberal arts colleges it's rare to have classes with 50 students or more due to the student population, and student to faculty ratio tends to be small in general.
2. Some variables might be correlated with each other. For example, undergraduate academic reputation is a very holistic variable, which suggests that it might be highly correlated with many other variables. Thus, *U.S. News* probably does not need to include all of the variables in the model. We included tuition in our model but *U.S. News* did not; however, just because *U.S. News* did not list tuition or financial aid in their methodology, it does not mean that those two factors are not influential in ranking predicting. Those two factors might be highly correlated with two variables that *U.S. News* did use - financial resources and the average spending per student.

In the future, we can gather data from individual school to see whether or not admitted students' SAT scores are normally distributed. We can also develop a reduced model without any SAT scores and test whether including SAT scores are important to predict a school's ranking on *U.S. News*' list accurately. Some other variables might be correlated with each other in our model and in *U.S. News*' model. Further study can be done to find correlations among other factors that might affect a college's ranking to better understand why *U.S. News* chose the variables that they used in their methodology and how they determined each variable's weight in the way they did. There are numerous variables that might affect a college's ranking. Our work demonstrates how to choose variables to determine each college's rank, while keeping multicollinearity in mind.

References

1. Forbes. "America's Top Colleges." *Forbes.* Forbes, July 2013. Web 15 Oct. 2013
2. Higher Education Research Institute. "About the CIRP Freshman Survey." *Higher Education Research Institute (HERI).* Higher Education Research Institute, Summer 2013. Web. 02 Nov. 2013
3. Morse, Robert. "Students Say Rankings Aren't Most Important Factor in College Decision." *US News.* US News & World Report, 27 Jan. 2011. Web. 2 Nov. 2013
4. Morse, Robert, and Flanigan, Sam. "How US News Calculated the 2014 Best Colleges Rankings." *US News.* US News & World Report, 9 Sept. 2013. Web. 02 Nov. 2013
5. *U.S. News*. "US News 2008-2013 Ranking Stats." Web. 02 Nov. 2013
6. *U.S. News*. "National Liberal Arts College Rankings." *US News & World Report*. US News, Sept. 2012. Web 2 Nov. 2013