College Scorecard Report

Ziao Lu, Lydia Maher, and Thomas Sun

December 1, 2016

Abstract

In this report we assess the performance of publicly funded schools to determine the allocation of grant money for policymakers. Specifically, we attempt to identify the best schools in the US in terms of their demographic, academic, and socioeconomic characteristics in order to help increase equity and graduation rates for underserved populations. Additionally, we use predictive modeling techniques to find the determinants for graduation rates and post-graduation income. These include ridge regression, lasso regression, principal components regression, and partial least squares regression applied to data on US universities from College Scorecard provided by the US department of education. The best fitting model is then used to make accurate predictions on graduation and income. We find that coming from a low-income bracket family, having federal loans, and being a first-generation student has a strong negative relationship with income. On the other hand, cost of attendance and SAT scores heavily influence completion rates positively.

Introduction

A main concern for policymakers and families alike is the quality of higher education institutions. The demand in the labor market for workers with educational attainment beyond high school has been growing significantly, and those unable to attend college due to lack of income are being further separated from the best opportunities. In an effort to promote equity and level the playing field, the government has set up several programs, such as the Pell Grant, that help enable underserved populations to attend university. Therefore, to determine allocation of these resources, it is important to understand what factors separate effective universities from low value universities.

To help families decide on the best school for their kids, the US Department of Education created a database of all US universities called College Scorecard. It contains highly detailed data various aspects on every university, such as graduation rates, composition of the student body, post-graduation income, etc. This dataset provides useful information that can be used to asses the performance of schools in order to allocate grant money.

Data

To perform our analyses, we use data from College Scorecard in 2012, provided by the US Department of Education. The full College Scorecard dataset contains extensive data on characteristics and indicators for all federal financial aid eligible schools in the US every year for the past several years. For the purpose of our study, we select two variables that we deemed relevant to determining performance of a school. These are four year graduation (completion) rates and post-graduate income ten years after graduation. We selected these two because they reflect the ability of the school to produce degree-holding scholars whose education and skills are valuable in the workforce.

In order to determine what factors influence these variables, we regress these outcomes on several predictors. These are: total number of students, race, income bracket of family, cost of attendance, first generation status, percent of students with federal loans, and average SAT score. For both completion rate and post-graduate income, we include the other one as a predictor as well.

We selected these predictors because we believe that they may be important in determining the outcome of a school. Total number of students may be important because larger student bodies might dilute the ability to focus on individual students and affect their learning. Higher percentage of underrepesented minorities may negatively influence graduation rates and earnings as well. Moreover, students facing high tuition costs and coming from poor income brackets or first generation families may find finishing a degree may be costly and burdensome. We also control for a student's innate ability through SAT scores.

Methods

Exploratory Data Analysis

First, we perform an exploratory analysis on the College Scorecard dataset, We obtain summary statistics for average SAT score, completion rates and threshold earnings for Blacks and as well as other relevant plots.

Data Processing

Before fitting any of the models, we next conduct some data processing. The qualitative variables are categorical and thus need to be transformed into dummy variables, or binary indicators, to be used in the regression functions. We also mean center and standardize the data to remove different measurement scalings and be more comparable. So, each variable has a mean of zero and standard deviation of one.

Model Building

There are five regression models to be fitted to the dataset, ordinary least squares, ridge, lasso, principal components, and partial least squares. The most common model used is ordinary least squares regression (OLS). OLS estimators have the advantage of being unbiased given that the relationship between response and predictors is truly linear.

However, OLS may have high variance and include irrelevant variables. In data with multiple dimensions, like in the College Scorecard dataset, an OLS regression is prone to overfitting, especially when the regressors are highly collinear. We use the other four regressions on the dataset and try to pick the best model to fit the data. That is, find the model that is easy to interpret while having the best predictive capability.

To build the models, we use a training dataset, containing a random sample of 75

Regression Models

In order to find the relationship between graduation data and income, and the predictors to be used for predictive modelling, we assume the relationship between the independent and dependent variables is linear. The relationship is assumed to be the following:

$$Y_i = \beta_0 + \beta_1 X_1 i + \beta_2 X_2 i + \ldots + \beta_{10} X_n i + \epsilon_i \tag{1}$$

Where β_0 is the intercept and β_1, ..., β_{10} are the regression coefficients for their associated predictors X_1, ..., X_n, and ϵ is the error term. Y_i is the dependent variable, either graduation rate or income. We first use an ordinary least squares method to be used as a benchmark for comparison between the other models. OLS estimates the coefficients by minimizing the residual sum of squares (RSS), defined as

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \tag{2}$$

For ridge regression, we minimize the RSS in addition to a shrinkage penalty, defined as

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

Where λ is the tuning parameter. As $\lambda \to \infty$, the shrinkage penalty grows, effectively shrinking the coefficients β_1, ..., β_{10} towards zero.

Lasso regression is another shrinkage method like ridge regression. The quantity we want to minimize is defined as

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

This is similar to the quantity for ridge regression, except the penalty now contains $|\beta_j|$ instead of β_j^2. The tuning parameter λ is the same, except now sufficiently large λ may shrink coefficients to exactly zero.

For principal components regression, the method tries to reduce the dimensions X_1, ..., X_p of the data matrix into principal components Z_1, ..., Z_M and then using the components as the predictors for least squares regression. The M principal compenents will be the tuning parameter, and we use cross validation find which M produces the smallest mean squared error.

Lastly, partial least squares, also a dimension reduction method, tries to find Z_1, ..., Z_M that approximate the original dimensions like PCR, but also tries to find new features related to the response *Balance*. The tuning parameter, the number of M directions, is the same as well. The M that is associated with the smalled mean squared error will be selected for the model.

Once all of the best models are identified, the test set will be used to compute the MSEs of each, and find which model performs best. The best model will finally be used on the full dataset to find official coefficients. The entire process is done twice, once for each outcome of interest.

Analysis

Ordinary Least Squares

First, let's look at our benchmark - OLS regression. We use full set of data that is mean centering and standardizing to fit the OLS model. OLS model will be served as our benchmark for comparison with later four different methods.

Here is more information about OLS regression for four year completion rates and post-graduate income:

Table 1: Summary Table of OLS Regression for Completion Rates

	Estimate	Std. Error	t value	Pr(>|t|)
Total	0.03	0.02	1.69	0.09
White	-0.21	0.06	-3.73	0.00
Black	0.06	0.05	1.20	0.23
Hispanic	0.01	0.03	0.18	0.85
Asian	0.20	0.03	7.56	0.00
AmericanIndian	0.01	0.02	0.51	0.61
NonResAlien	0.00	0.02	0.10	0.92
Completion4yr	-0.05	0.03	-1.62	0.11
LowIncome	-0.34	0.03	-9.90	0.00
Mid1Income	-0.06	0.02	-3.04	0.00
FirstGen	0.27	0.03	8.20	0.00
FedLoans	-0.11	0.02	-4.90	0.00
CostAttendance	0.17	0.03	6.01	0.00
Avg_SAT	0.62	0.03	18.14	0.00

Table 2: Summary Table of OLS Regression for Post-Graduate Income

	Estimate	Std. Error	t value	Pr(>|t|)
Total	0.03	0.02	1.69	0.09
White	-0.21	0.06	-3.73	0.00
Black	0.06	0.05	1.20	0.23
Hispanic	0.01	0.03	0.18	0.85
Asian	0.20	0.03	7.56	0.00
AmericanIndian	0.01	0.02	0.51	0.61
NonResAlien	0.00	0.02	0.10	0.92
Completion4yr	-0.05	0.03	-1.62	0.11
LowIncome	-0.34	0.03	-9.90	0.00
Mid1Income	-0.06	0.02	-3.04	0.00
FirstGen	0.27	0.03	8.20	0.00
FedLoans	-0.11	0.02	-4.90	0.00
CostAttendance	0.17	0.03	6.01	0.00
Avg_SAT	0.62	0.03	18.14	0.00

From the above information, we can conclude that the coefficient for predictors such as family income, federal loans, and SAT score are extremely significant. Meanwhile, proportion of Black and Hispanic population is not significant. This is true for both determinants of completion and income.

Here we find that the statistically significant predictors are similar to those of the completion rate regression. Again, Black and Hispanic populations are shown to have no significant effect on income.

Other methods

Next, we use the other regression methods to possibly find better fits to the data. We load mean centerized and standardized data before the analysis. For ridge and lasso regression, we first check mission value in data for both the train and test sets. We then initialize lambda and use random seeds for cross-validation. Next, we use the train set to conduct 10-fold cross-validation to find out the best tuning parameter.

We find the λ for the best model. Then, we use the test set to compute the test Mean Square Error. Finally we refit the model on the full data set using the best lambda and get official coefficients.

PCR and PLSR follow a similar process As before, we check the mission value in the data for both train and test sets and use random seeds for cross-validation. We then conduct 10-fold cross-validation on the test set to find out the best number of principal components used.

Once finding the number of principal components considered for the best model, we use the test set to compute the test Mean Square Error and finally refit the model on the full data set using the best number of components and get official coefficients. The next section shows the results from these methods.

Results

Completion Rates

From our exploratory analysis we find several interesting relationships between completion rates and several predictors. Below are some scatterplots exhibiting these relationships. From the below figures we see that first generation status is negatively correlated with completion, cost of attendance is positively correlated, and percentage of student body as black seems to have no correlation.

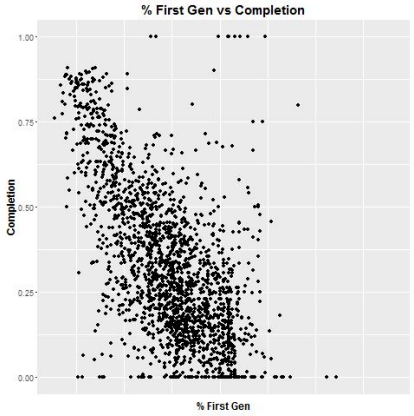

Figure 1: Plot of Completion Rates against First Generation Status

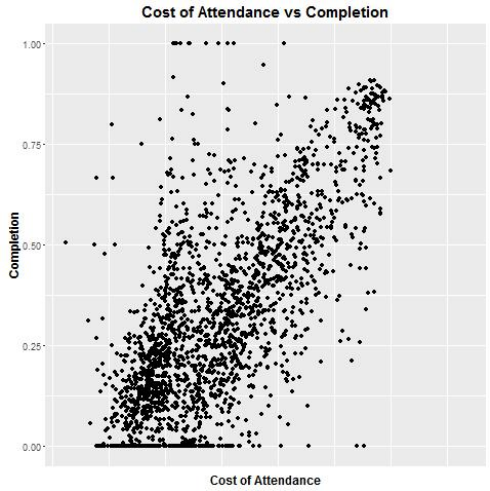

Figure 2: Plot of Completion Rates against Cost of Attendance

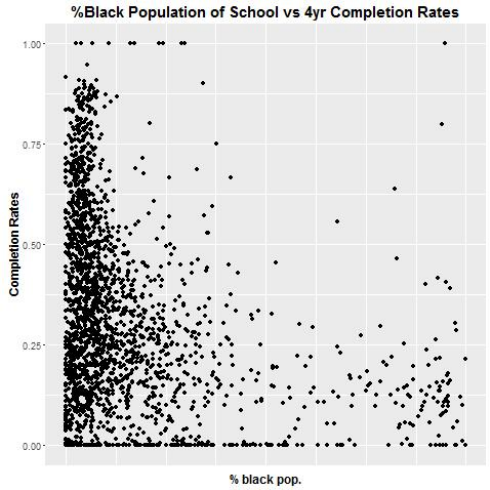

Figure 3: Plot of Completion Rates against Percentage of Black Population

Now let's look at our predictive models. We obtain the test mean squared errors for every predictive model from our analysis. Below is the table of MSEs for each model.

Table 3: Test MSE for All Methods

	Ridge	Lasso	PCR	PLSR
MSE Value	0.277	0.272	0.280	0.280

Table 3 has only one row (Test MSE value) and four columns (one column per regression methods: ridge, lasso, pcr, and plsr).

From Table 3, the result shows that the model with lowest test Mean Square Error is Lasso Regression, which means that lasso regression actually has the best performance when we test the prediction against the true value in testing set. Thus we use the results from lasso regression as our best model and official coefficients for the regression.

The following table contains the regression coefficient results for completion rates. We include the results from all methods for comparison, with lasso regression selected as our official results.

Table 4: Regression Coefficients for All Methods

	OLS	Ridge	Lasso	PCR	PLSR
Total	0.033	0.036	0.023	0.039	0.039
White	-0.213	0.008	0.000	0.020	0.020
Black	0.059	0.092	0.052	0.112	0.112
Hispanic	0.006	0.021	0.000	0.029	0.029
Asian	0.203	-0.002	0.000	0.002	0.002
AmericanIndian	0.009	-0.023	-0.020	-0.020	-0.020
NonResAlien	0.002	-0.036	-0.024	-0.035	-0.035
EarningsAgg	-0.050	-0.028	0.000	-0.039	-0.039
LowIncome	-0.339	-0.251	-0.221	-0.263	-0.263
Mid1Income	-0.060	-0.028	-0.021	-0.027	-0.027
FirstGen	0.273	-0.064	-0.076	-0.049	-0.049
FedLoans	-0.105	-0.064	-0.043	-0.066	-0.066
CostAttendance	0.168	0.452	0.436	0.461	0.461
Avg_SAT	0.616	0.320	0.302	0.335	0.335

From Table 4, the results show that regression coefficients for Ridge, Lasso, PCR, and PLSR are approximately closed to each other's value but are slightly different compared to OLS - our benchmark.

Not surprisingly, we have seen that some coefficients in lasso regression are zero because lasso regression allows coefficients to be zero to minimize the regression penalty.

From our results we see that coming from a low income bracket family, having federal loans, and being a first generation student have a strong negative relationship with income. Interestingly, race does not seem to have a large influence on completion rates after controlling for other factors. If anything, an increased percentage of black population tends to increase completion rates. Cost of attendance and SAT scores heavily influence completion rates positively

as well, possibly because more prestigious schools are more expensive with better students so they have better graduation rates.

Income

From our exploratory analysis we also find some interesting relationships between income and several predictors. Below are some scatterplots exhibiting these relationships. From the below figures we see that, like with completion rates, first generation status is negatively correlated with income 10 years after graduation, while cost of attendance is positively correlated. Also, percentage of student body as black seems to have no correlation with post-graduate income.

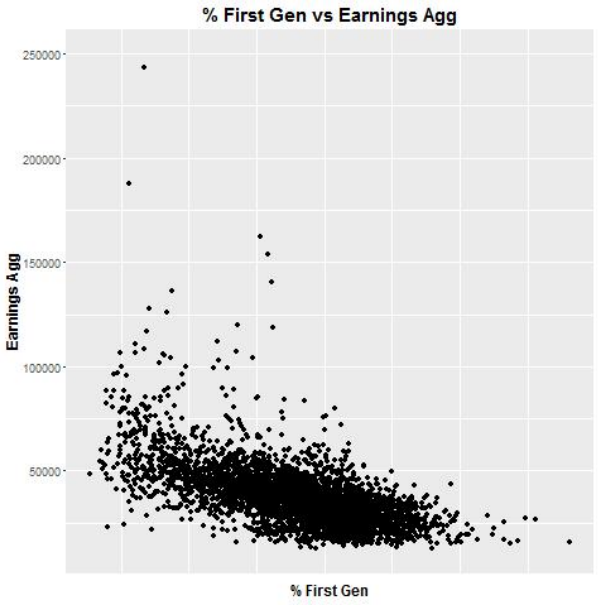

Figure 4: Plot of Earnings against First Generation Status

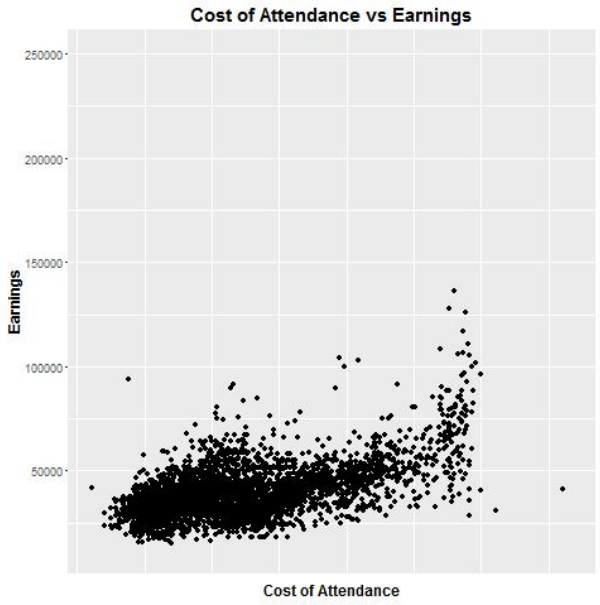

Figure 5: Plot of Earnings against Cost of Attendance

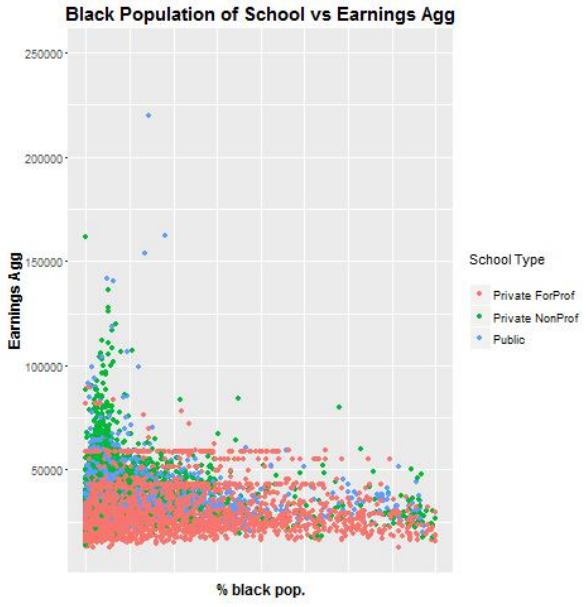

Figure 6: Plot of Earnings against Percentage of Black Population

Now, we look the MSE of each model, this time with post-graduate income as our regressor.

Table 5: Test MSE for All Methods

	Ridge	Lasso	PCR	PLSR
MSE Value	0.322	0.321	0.322	0.322

Table 5 is similar to Table 1, except the MSEs are overall larger than with completion rates as our outcome, meaning that the regression line fits the data somewhat better for completion rates.

From Table 5, the results shows that the model with lowest test Mean Square Error is also Lasso Regression. Thus we use the results from lasso regression as our best model and official coefficients as well.

The following table contains the regression coefficient results. We include the results from all methods for comparison, with lasso regression selected as our official results.

Table 6: Regression Coefficients for All Methods

	OLS	Ridge	Lasso	PCR	PLSR
Total	0.033	0.035	0.022	0.033	0.033
White	-0.213	-0.195	-0.209	-0.164	-0.191
Black	0.059	0.056	0.000	0.101	0.078
Hispanic	0.006	0.011	0.000	0.028	0.016
Asian	0.203	0.208	0.198	0.219	0.208
AmericanIndian	0.009	0.008	0.000	0.014	0.012
NonResAlien	0.002	0.004	0.000	0.011	0.007
Completion4yr	-0.050	-0.036	0.000	-0.052	-0.053
LowIncome	-0.339	-0.312	-0.240	-0.339	-0.338
Mid1Income	-0.060	-0.058	-0.040	-0.059	-0.055
FirstGen	0.273	0.241	0.167	0.274	0.270
FedLoans	-0.105	-0.105	-0.082	-0.105	-0.104
CostAttendance	0.168	0.166	0.133	0.174	0.172
Avg_SAT	0.616	0.584	0.560	0.614	0.618

From these results we learn that coming from a low income bracket and having federal loans are negatively associated with post-graduate income. Black and Hispanic population do not affect income, rather increased percentage of Whites has a negative effect while increased percentage of Asians has a positive effect. From this we learn that underrepresented races in universities do not affect post-graduate income after controlling for other variables. First generation status, cost of attendance, and SAT scores have a positive relationship with income. The last two may be expected for the same reason as the positive effect on completion rate, but the coefficient for first generation status may be suprising.

Conclusions

From the College Scorecard data, we wanted to learn what characteristics of an institution determine its value and performance in order to determine how to allocate grant money to the most effective institutions. Our exploratory data analysis showed that

We fitted an OLS model upon the dataset so that we could get more insight about the information hidden behind the data. And the summary statistics served as our benchmark. Then, after some pre-model data processing, we looked to pick the best model from two shrinkage methods (ridge and lasso) and two dimension reduction methods (PCR and PLSR). The test mean squared error from all 4 different methods were competitive with each other. Lasso regression method achieved the lowest test MSE and is thus considered the "best model" among the four methods.

Using the estimated coefficients from lasso regression, we find that strong negatively associated predictors of completion rates include include coming from a lower income family, first generation status, and federal loans. Meanwhile, black population, school size, cost of attendance and average SAT score are positively related with completion. For post-graduate income, strong negative predictors are white population, low income status, and federal loans. Suprisingly, black and hispanic populations are not relevant in both regressions.

Our findings suggest that it is income, not race, that influence the completion rates and post-graduate income of students. Therefore we recommend that these grants should be given out based on income. in particular, low to middle class families and first generation students who may have to rely on federal loans otherwise. Giving grants loans to these students may encourage them to complete their degree and not become financially burdened by high debt. Institutions with high percentages of these kinds of students should be assessed in particular.

References

College Scorecard Data https://collegescorecard.ed.gov/data/

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013. Print.