Method: Predictors for Institution Graduate Loan Default Rate

Background

What makes a certain college a good investment? Is it producing graduates with less debt? Is it producing graduates who make more money 10 years after finishing? Or is it something less quantifiable, like guiding students into discovering their vocation or completing meaningful coursework?

The **U.S. Department of Education** quantifies college performance on a number of dimensions. **College Scorecard** is an interactive tool designed to help students choose a college based on government data about cost, financial aid, retention, completion and programs. The data, spanning 1996-2013, is publicly available in .csv format.

Summary of Findings

A linear regression model fit to 4 years of observations (2010-2013; n=6080; training set 80%) found higher completion rate predicted lower default rates, while more Pell Grants predicted higher default rates.

Graduating students on time (or close) should create graduates more likely to keep their momentum going in finding a job and paying down loans.

More students with Pell Grants should predict higher rates of loan default since they indicate that students are taking a bigger financial risk (see this report on financial demographics of Pell Grant recipients; 67% were at or below 150% of the poverty line compared to 16.5% of non-recipients).

Data & Method

The model was trained on 4 years of data, 2010-2013, with 20% of the data held out as a validation set. Only four-year public and private non-profit colleges and universities with at least 500 students were included in these figures (1666 schools).

Dependent Variable: Default Rate
While the dataset includes more than 1700 variables, I decided to focus on the **default rate (3 year cohort)** as the best measure of a college's value. It combines debt with post-graduate earnings into one feature, since a student that defaults on their student loans is clearly either buried in debt, not making high enough salary, or both. The median graduate debt dimension in the dataset wasn't viable since it only includes federal loan debt (which caps at $27,000) and the post-graduate earnings dimension was sparsely populated and therefore not ideal.

Default rate had a right skewed distribution, which I remedied by doing a log transform. See item 1 in the appendix for a comparison of distributions.

Independent Variable Selection
I cherry-picked about 30 variables out of the dataset that I thought would have the greatest effect on default rate, including factors related to retention, completion, student debt, financial aid and cost. A large percentage of the variables in the dataset were too specific to have much weight - things like "percent of dependent students who died within 3 years at original institution" and "percent of high-income (above $75,000 in nominal family income) students withdrawn from original institution within 8 years." A lot of variables were missing 90-100% of values and thus had to be thrown out (e.g., "percent of students who submitted a FAFSA to at least one college"). A few variables were excluded after they were found to be highly correlated with other variables (e.g. average family income and Pell Grants; completion and retention).

Variable	Description	Min	Max	Units
Type	Public or Private Non-Profit	n/a	n/a	n/a
Adm.Rate	Admission Rate	5.69	100	0-100
Enrollment	Size of Enrollment	0.501	59.183	in thousands
Avg.Cost	Average Cost	2.2	62.636	in thousands
Inst.Expense	Institutional Expense per Student	0	169.338	in thousands
Avg.Faculty.Pay	Average Faculty Pay	0.788	19.862	in thousands
Pell.Grant.Pct	Percentage of Students with Pell Grants	0	100	0-100
Completion	Completion Rate (6 years)	0	100	0-100
Pct.Fed.Loan	Percentage of Students with Federal Loans	0	100	0-100

Missing Value Imputation

I only selected variables with less than 25% missing values. Because the Default.2Y and Defaulty.3Y variables were missing values in alternate observations (rarely both missing in an obs.), I was able to combine the 2 year and 3 year default rates and impute the rest. I then used the **mice** package (Multivariate Imputations by Chained Equations) to impute the remaining missing values using the pmm or predictive mean matching method. See item 2 in the appendix for a breakdown of missing values by variable.

Model

Model Selection

I fit a model with all variables, which found all but institutional expense to be significant, providing an adjusted R-squared value of 0.59, F = 778.6, p < 2.2e-16. Stepwise selection recommended keeping all variables, but keeping only percentage of students with Pell Grants and completion rate resulted in a loss of only 0.015 from adjusted R-squared (Adj R-sq = 0.575, F = 3296). I decided to select this model for simplicity in interpretation. See item 3 in the appendix for a summary of the model; item 4 for the diagnostic plots and item 5 for the predictons vs. actual observations of the test set.

Diagnostics

The diagnostic plots are mostly normal save for the presence schools with default rates of 0 (29 total, a good chunk of them seminaries or bible schools, but also included California Institute of Technology, Claremont McKenna College, Harvey Mudd College and Vassar College).

Validation

Comparing the model's predicted values to the actual values in the test set produced an R-squared value of 60.50%.

Coefficient Interpretation

Since the dependent variable was log transformed, the coefficients can be interpreted in a slightly different way. Instead of the unit increase in the DV per one unit increase in each variable, the exponentiated (exp) coefficient represents the percent change in the DV per unit chnage in each variable. Those values are shown below, where log.estimate is the coefficient, exp.estimate is exp(coef) and exp.estimate.minus1 is exp(coef) -1, or the percent change.

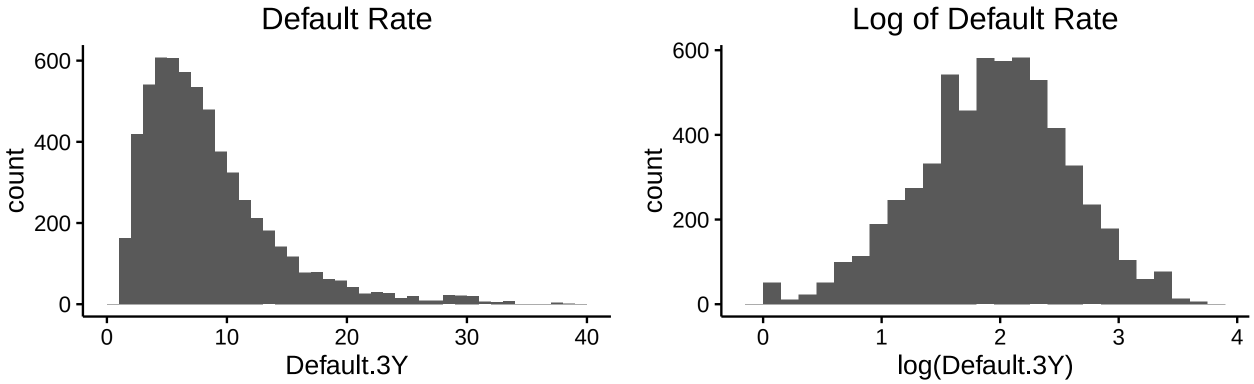
The factor predicting a lower default rate for a school is completion rate (0.0170% decrease in default rate per one percent increase in completion rate).

The factor predicting higher default rates is Pell Grant percentage (0.0113% increase in default rate per one percent increase in Pell Grants).

coef	log.estimate	exp.estimate	exp.estimate.minus1
(Intercept)	2.4554880	11.6521189	10.6521189
Pell.Grant.Pct	0.0113103	1.0113745	0.0113745
Completion	-0.0174821	0.9826698	-0.0173302

Appendix

1. Default rate transformation



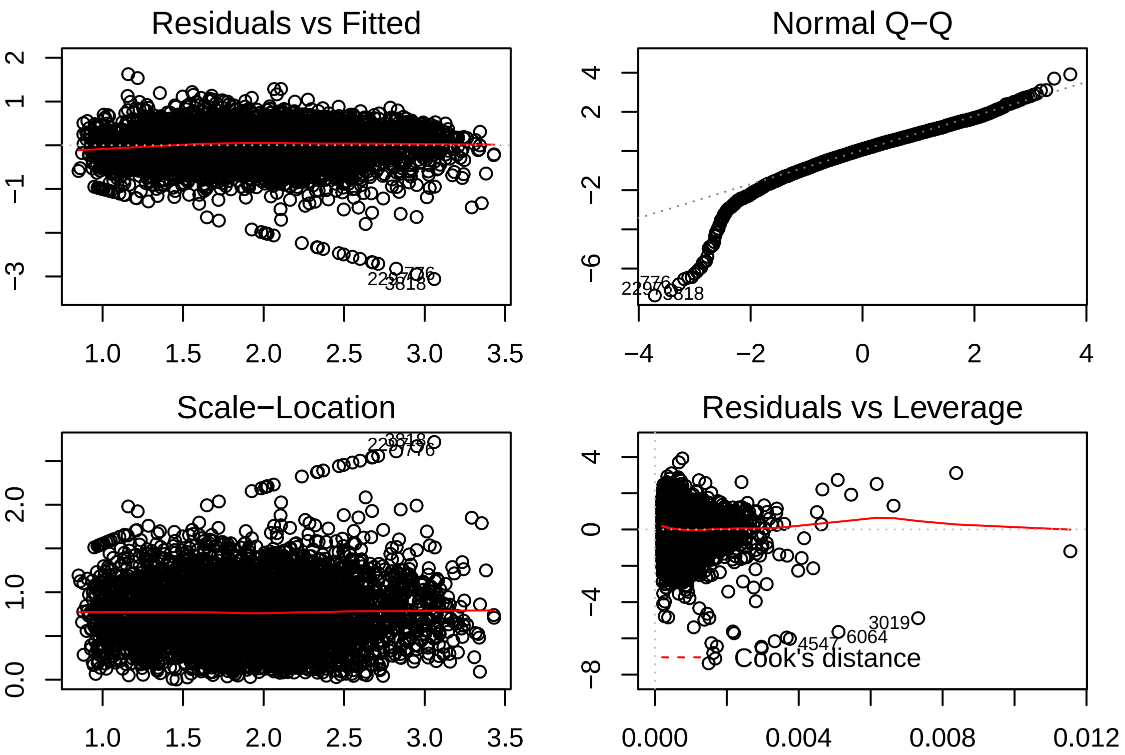
2. Missing values

Variable	Imputed
Admission Rate	529
Completion Rate	165
Median Debt	25
Institutional Expense per Student	31
Avg. Faculty Pay	31
Default Rate	8
Percent of Students with Pell Grants	4

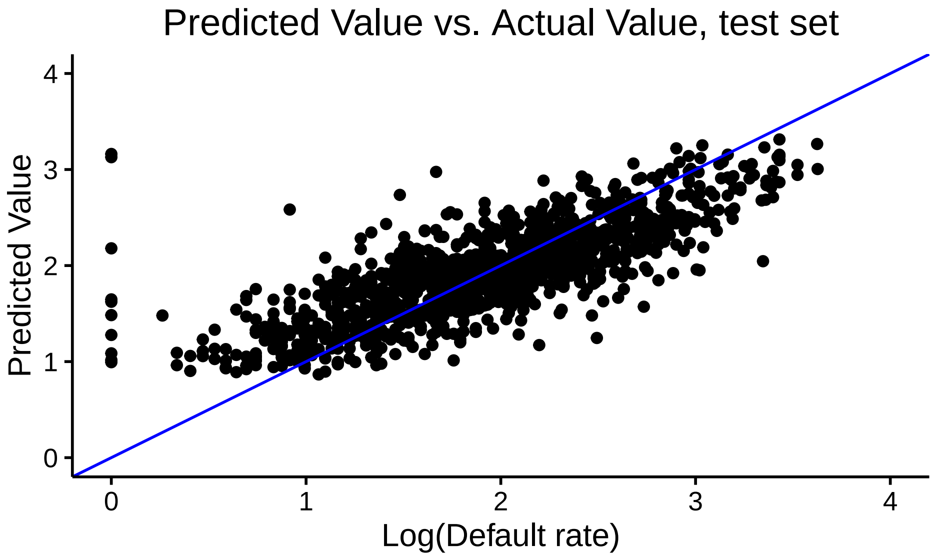
3. Summary of linear model

```
##
## Call:
## lm(formula = log(Default.3Y) ~ Completion + Pell.Grant.Pct, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05960 -0.22105  0.03706  0.26554  1.62560
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.4554880  0.0379167   64.76   <2e-16 ***
## Completion     -0.0174821  0.0004299  -40.66   <2e-16 ***
## Pell.Grant.Pct  0.0113103  0.0004733   23.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4151 on 4861 degrees of freedom
## Multiple R-squared:  0.5756, Adjusted R-squared:  0.5754
## F-statistic:  3296 on 2 and 4861 DF,  p-value: < 2.2e-16
```

4. Linear model diagnostics



5. Linear model validation


Predicted Value vs. Actual Value, test set
Predicted Value
Log(Default rate)