Default Rate Linear Regression Method

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.

Data & Method

Data

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 at least partially by taking its square root. 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within 3 years at original institution" and "percent of high-income (above $75,000 in nominal family income) students withdrawn from original institution within 8 years." Additionally, 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").

Variable	Description
Type	Public or Private Non-Profit
Adm.Rate	Admission Rate
Enrollment	Size of Enrollment
Avg.Cost	Average Cost
Net.Revenue	Net Revenue per Student
Inst.Expense	Institutional Expense per Student
Avg.Faculty.Pay	Average Faculty Pay
Percent.FTE.Faculty	Percent Full Time Faculty
Pell.Grant.Pct	Percentage of Students with Pell Grants
Completion	Completion Rate (6 years)
Retention	Retention Rate
Pct.Fed.Loan	Percentage of Students with Federal Loans
Avg.Fam.Income.Dependent	Avg. Family Income (dependent students)
Md.Debt.n30	Median Debt of Graduates (suppressed for n<30)

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 Selection
I fit a model with all variables, which found all but institutional expense to be significant, providing an adjusted R-squared value of 0.6047, F = 532.4, p < 2.2e-16. Stepwise selection recommended dropping Net Revenue per Student, Average Faculty Pay and Percentage of Students with Federal Loans, which resulted in a model with an increased F statistic but no change in the R squared value of 0.6047, F = 677.4, p < 2.2e-16. 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.

Model

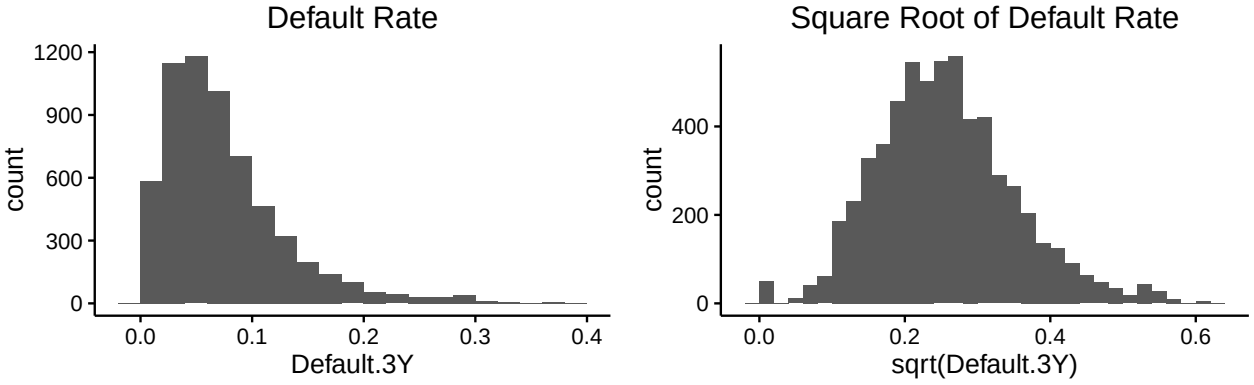
A quick interpretation of the results is that higher retention, higher completion rate, higher institutional spending per student and higher admission rates predict lower default rates, while being public, larger enrollment, higher cost, more full time faculty, more students with Pell Grants and debt predict higher default rates.

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 62.69%.

Appendix

1. Default rate transformation



2. Missing values

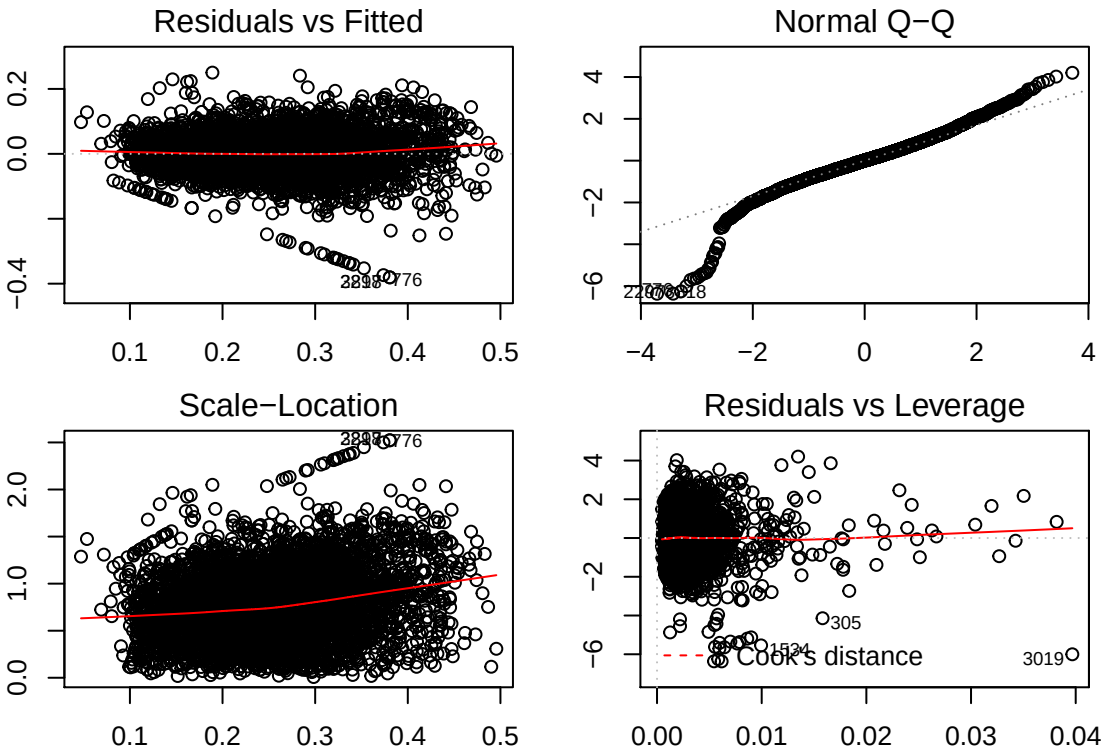
Variable	Imputed
Avg. Family Income (dependents)	908
Admission Rate	529
Completion Rate	165
Retention Rate	141

Variable	Imputed
Median Debt	25
Net Revenue per Student	31
Institutional Expense per Student	31
Avg. Faculty Pay	31
Percent Full Time Faculty	15
Default Rate	8
Percent of Students with Pell Grants	4

3. Summary of linear model

```
##
## Call:
## lm(formula = sqrt(Default.3Y) ~ Type + Adm.Rate + Enrollment +
##     Avg.Cost + Inst.Expense + Percent.FTE.Faculty + Pell.Grant.Pct +
##     Completion + Retention + Avg.Fam.Income.Dependent + Md.Debt.n30,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38061 -0.03429  0.00058  0.03437  0.25023
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.598e-01  1.274e-02  28.254  < 2e-16 ***
## TypePublic                  2.221e-02  3.175e-03   6.995 3.02e-12 ***
## Adm.Rate                   -2.442e-02  5.335e-03  -4.578 4.82e-06 ***
## Enrollment                  5.941e-07  1.563e-07   3.802 0.000145 ***
## Avg.Cost                    1.040e-06  1.561e-07   6.663 2.98e-11 ***
## Inst.Expense               -4.570e-07  1.304e-07  -3.505 0.000461 ***
## Percent.FTE.Faculty         2.296e-02  3.604e-03   6.369 2.08e-10 ***
## Pell.Grant.Pct              1.854e-01  1.077e-02  17.206  < 2e-16 ***
## Completion                 -1.755e-01  1.037e-02 -16.922  < 2e-16 ***
## Retention                  -1.380e-01  1.225e-02 -11.270  < 2e-16 ***
## Avg.Fam.Income.Dependent -4.527e-07  8.878e-08  -5.099 3.54e-07 ***
## Md.Debt.n30                 1.163e-06  2.511e-07   4.633 3.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05996 on 4852 degrees of freedom
## Multiple R-squared:  0.6056, Adjusted R-squared:  0.6047
## F-statistic: 677.4 on 11 and 4852 DF,  p-value: < 2.2e-16
```

4. Linear model diagnostics



5. Linear model validation