Exploring Graduation Rates - How does money matter?

Stats 201b Project Progress Report

Duncan Clark & Leonard Wainstein

February 22, 2018

Data Source

All of the data used in this analysis is publically available and can be downloaded from: https://www.deltacostproject.org/delta-cost-data

Research Question

How do the ways in which public 4-year institutions spend their money affect graduation rates?

Data Exploration:

We reviewed the data dictionary (provided with the data) and identified key variables we would be interested in examining and including in our models. For those variables, we examined the yearly missing rates to identify which ones would be unusable and which ones could be workable with some minor imputation.

Literature Review:

We reviewed two research papers that used this data to examine the impact of Outcomes (or Performance) Based Funding on bachelor's degree attainment in public 4-year institutions to get ideas on how to define our analytical sample, which key variables we should include in our models, and how those who work in higher-education research standardize key variables to control for institution size. The papers are below:

- http://journals.sagepub.com/doi/pdf/10.3102/0162373714560224
- https://aefpweb.org/sites/default/files/webform/42/Playing%20the%20Long%20Game%20-%20Slaughter%20et%20al..pdf

We would like to look more into the literature time permitting, and investigate if there have been other attempts to model such data flexibley.

Data Preparation:

We first found and cleaned data on state and year specific GDP per capita so they could be merged with our final dataset, as the two studies above included them in their models. Because graduation rates are only available starting in academic year 2003, we started by keeping years past then in the data, which left us with 553 institutions. We then thinned our sample out in the following way:

- Kept schools in the 50 states plus DC (544 institutions kept)
- Dropped schools that didn't exist in the entirety of the time range, academic years 2003-2015 (505 institutions kept)
- Dropped schools that were missing bachelor's degrees awarded or graduation rate in any of the years in the data (457 institutions kept)
- Dropped schools that gave out 0 bachelor's degrees in a year or awarded fewer bachelor's degrees than associate degrees in a year (445 schools kept)

We then standardized monetary variables to 2015 dollar values using CPI scalars and standardized variables that would be sensitive to institution size by dividing them by the number of fall full-time equivalent (FTE) students.

The final step for preparation is to determine how to impute missing values for variables with workable missing rates.

We split our data into a training and test set, with stratification on geographic region, since this seems to have a fixed effect on graduatiojn rates, this mitigates the risk of high variance estimates of the fixed effects from the training data, or even no estimate in the case were all of one region were included in the training of test set.

Preliminary Models:

We identified our primary outcome of interest in the data as the percentage of students graduating within 150 percent of normal time. Our initial thoughts on modelling this classically were as follows:

- Include fixed effects for geographic regions.
- Include variables relating to proportion of underrepresented minorities.
- Include variablse to represent overall socio-economic deprevation of student population.
- Include expenditure variables.

In a simple classical linear regression model, we see that each of the following intial variables are significant:

- Pell grants per full time equivalent(fte) student
- Percentage of black students
- Percentage of hispanic students
- Instruction spending
- Student service spending

This gives us confidence that we should be able to find some model that fits the data well with a some skind of positive relation with spending. From which we can hopefully draw an interpretation.

Model Selection

We used lasso to do automatic variable selection. We ran into the problem of missing of data for around 300 of our around 400 covariates. Initially we replaced the missing covariates with mean values, and ran lasso to give a selection of variables. Then ran the same model but restricted to covariates with no missingness. These two approaches essentially yielded different selections of variables. Noting that the top coefficients in the first approach had low levels of missingness i.e. <200 in 4000 observations, we took a hybrid approach, including all covariates with missingness <200.

The Lasso also allowed us to easily find obviously multicolinear variables with our outcome, e.g. bachelor degrees per FTE, and remove them and run the lasso again. We found that glmnet in R does not easily cope with factors, so we needed to do some manual manipulation to allow for this.

We propose to run linear regression models on the lasso selected variables with 2 way clustering for standard errors, and look at the first differences to see if we can draw interpretation from this.

Then we will compare this to our self selected model based on hyptotheses about education spending, and see if we can draw any conclusions from this. We will work on this independently to minimise the influence of the Lasso selection results on the self selection model.

Also we have education spending variables in great granularity, so we could potentially lasso on all of these to see which might be most important then rerun models regression models with the spending variables with the highest predictive power. Then we hope to be able to draw conclusions from this.

Other Ideas

It seems highly doubtful that our data is truely linear, and moreover, Universities may differ significantly in intangible ways not captured by the covariates, e.g. reputational factors. We would like to explore using more flexible modelling techniques e.g. kernalised method, to predict the effect of changing a variable e.g. the effect

of an increase in student service spending. We could do this on both our self and lasso selected variables and compare the differences.

We will us the training data set to train the models, and then the test data set to see if they have out of sample validity. Strong out of sample validity lends credibility to any interpretations we will draw from our models.