Examining the Relationship of Expenditures and Student Outcomes at Public 4-Year Institutions

Duncan Clark and Leonard Wainstein

March 23, 2018

Introduction

Higher education institutions serve a public good; they educate their students with the hope that in the future these graduates will go on and contribute to society as a whole. However, these students also benefit, as they tend to be more successful than their non university educated peers. Thus, since much of higher education is funded by the public purse, it becomes a matter for the public to decide how much to subsidise individuals' educations for the good of the whole, and in particular it becomes crucial that the public money is used efficiently. It is this efficiency that we concerned ourselves with in this project.

In this project we hoped to analyse quantitatively the relationship between student outcomes at public 4-year higher education institutions and their expenditures. Specifically we looked at 6-year graduation rates for first-time full-time bachelor's degree-seeking student cohorts and the number of bachelor's degrees awarded per full time equivalent student (FTE). We decided to look at two outcomes because certain institutions may be more concerned with its bachelor's degrees per FTE than its graduation rate (or vice versa). A small school with mostly full-time student cohorts would likely be more interested in increasing its graduation rate. However, large schools with many part-time students would also want to know if the rate at which they are producing graduates is keeping up with the total number of credits students are taking that institution.

We were interested in using methods developed in this course to predict the two outcomes, but we were also interested in making valid interpretations of our results. With this in mind we first tried only including variables that we believed to be important as in Tanberg and Hillman's 2014 study of Performance/Outcomes Based Funding, "State Higher Education Performance Funding: Data, Outcomes, and Policy Implications" and Research for Action's 2017 study, also on Outcomes Based Funding, "Playing the Long Game". We then tried carrying out automatic variables selection through Lasso regression and using these variables as the basis for or models.

The data we analyzed was obtained from the Delta Cost Project at American Institutes for Research, which gave a dataset for all higher education institutions in the US for 2003 to 2015. The data is publically available here: https://www.deltacostproject.org/

Data Processing

The Delta Cost Project is a panel data set that has institutional characteristics and outcomes for every university in United States and its territories for academic years 1987 to 2015 (academic years are indexed by the year during which conclude - e.g. the 2015 academic year is the 2014-15 school year).

We restricted our sample to 4-year public primarily bachelor's degree-granting institutions in the 50 states (plus DC), and excluded any institutions that did not exist in all years from 2003 to 2015. Years from 2003 onwards generally had a low level of missingness in the outcomes of interest, as well as covariates regarding expenditures, so we chose this as our cutoff for inclusion. We also dropped schools that were missing values in one of the outcomes for any of the years from 2003 to 2015, and schools with missing values for a few other key variables on all years of data. As for examining other types of institutions, this could be interesting though there is liklely to be heterogeneity since the institutions serve different purposes, so for our purposes we simply excluded them. Starting with 553 schools in 2003 to 2015, thinning the sample in the described way went as follows:

- 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)
- Dropped schools that were missing values in all years from 2003 to 2015 for a few key variables (432 schools kept)

Dollar values where scaled to 2015 dollars using CPI, to allow for fair comparison, and monetary totals were scaled by total full time equivalent enrollment to account for the size of the institutions. There are altenate inflation indices, that may have been appropriate, for example HEPI - higher education price inflation, which is specific to the higher education sector, however these were not considered.

We also added state and year specific GDP and unemployment data in line with the approach in . The state GDP data came from the St. Louis Federal Reserve Bank's website for economic research (https://fred.stlouisfed.org/) and the state unemployment data came from the U.S. Department of Labor, Bureau of Labor Statistics website (https://www.bls.gov/).

Training and Testing Split

To validate our models we split our data into a test set and a training set. The test set was utilised only after models had been generated on the training set and did not inform our modeling decisions (e.g., such as dropping of variables)

Since each institution appears in our dataset 13 times (once for each year in 2003-2015), we clustered sampled by institution in splitting our dataset, selecting all years of data into the test or training set for each institution. In addition, we stratified by census division, to assure that each region would be represented in our training and testing set and reduce the variance in our estimators, as Tanberg and Hillman's study identified location as an important predictor of student outcomes. The census divisions were defined as below:

Table X: States in each Census Division

Region	**States**
New England	CT, ME, MA, NH, RI, VT
Middle Atlantic	NJ, NY, PA
East North Central	IN, IL, MI, OH, WI
West North Central	IA, KS, MN, MO, NE, ND, SD
South Atlantic	DE, DC, FL, GA, MD, NC, SC, VA, WV
East South Central	AL, KY, MS, TN
West South Central	AR, LA, OK, TX
Mountain	AZ, CO, ID, NM, MT, UT, NV, WY
Pacific	AK, CA, HI, OR, WA

We decided against including state fixed effects in any of our models, as we worried state fixed effects may account for too much variation, leading to poor prediction. We also decided against stratifying by state when splitting our data into a training and test set, and there were some states with only one institution in them, meaning states would be unrepresented in the training or test set.

After splitting the data, our training sample and test set looked as follows:

Table X: Training and Testing Split

	# of Institutions	**# of Obs.**
Training	347	4511
Testing	85	1105
Total	432	5616

Models Including Variables per Literature

Set-up

In attempt to hone in on the effect that changes in institution spending strategies have on student outcomes, we focused on the nine types of expenditure variables that were available to us in all institutions and years in our sample. These expenditure types are:

Table X: Expenditure Types and Examples

Expenditure Type	**Example(s)**
Instruction	Teacher Salaries
Public Service	Conferences, Community Services
Academic Support	Libraries, Museums, Demonstration Schools
Student Services	Student Activities/Organizations
Institutional Support	General Administration
Operation and Maintenance	Utilities
Depreciation	Cost of Captial Assets
Scholarships and Fellowships	Scholarships, Fellowships
Auxiliary Enterprises	Residence Halls, Student Health Services, Intercollegiate Sports

Along with these expenditure variables, which were converted to 2015 values and divided by FTE enrollment, we also included several institution and state characteristics as per the two research studies mentioned before. These extra covariates were:

Table X: Covariates Included as per Literature

State-level Variables
GDP per Capita (state and year-specific)
Unemployment Rate (state and year-specific)
Institutional Characteristics
Undergraduate Enrollment
% of Total Enrollment that is Under-represented Minority (Black or Hispanic)
Revenue from Pell Grants per FTE
Revenue from State Appropriations per FTE
Net Tuition and Fees Revenue per FTE
Total Revenue per FTE
Categorial Variables
Year Dummies (2003-2015)
Census Division Dummies (9 Divisions)

We performed minor missing imputation on the expenditure variables and above covariates by imputing with the predicted values of a linear regression model including only dummies for institution and year.

Models

With these self-selected variables, we tried seven types of models on each of the two outcome variables:

- Traditional Linear Regression (LM)
- Traditional Linear Regression with Interactions
- Lasso Regression
- Lasso Regression with Interactions
- Ridge Regression
- Ridge Regression with Interactions
- Gaussian Kernel Regression

For the models with interactions, we interacted the nine expenditure variables with each other, as we hypothesized that spending more money in one expenditure area may change the effect of spending more money in another on these outcomes (e.g., spending more money on instruction may have a stronger effect if the institution has already invested heavily in making students feel comfortable on campus, such as high spending for residence halls or student health services).

For the Lasso and Ridge models, we performed ten-fold cross-validation on the training set to estimate the regularization coefficient (λ). For the Gaussian Kernel model, we chose the λ that minimized the leave-one-out cross-validation error on the whole training set.

The testing root mean squared error (RMSE) of the above models on both outcomes is provided in Table X below

Table X: Testing Raw RMSE of Self-Selected Models (standardized RMSE in parentheses)

Model	**6-Year Graduation Rate**	**Bachelor's Degrees per FTE**
LM	0.1008 (0.6385)	0.0348 (0.9194)
LM (with interactions)	0.0983 (0.6230)	0.0359 (0.9461)
Lasso	0.1004 (0.6359)	0.0346 (0.9129)
Lasso (with interactions)	0.0979 (0.6205)	0.0337 (0.8895)
Ridge	0.1005 (0.6366)	0.0344 (0.9074)
Ridge (with interactions)	0.0990 (0.6276)	0.0331 (0.8734)
Gaussian Kernel	0.0959 (0.6076)	0.0274 (0.7235)

For 6-year graduation rate, the gaussian kernel model had the lowest RMSE. However, the kernel model performed only slightly better than the Lasso model with interaction (the model with the second lowest testing RMSE), and in general, the linear models performed quite well on the test data compared to the kernel model. Among the linear models, the models with the interaction terms also performed better than the models without interactions, lending support to our hypothesis that there is some interaction effect between the expenditure variables.

For bachelor's degrees per FTE, the gaussian kernel model again had the lowest RMSE, and in this case performed much better than the linear models. And again, the linear models with interaction effects generally performed better than those without them. However, looking at the standardized errors in the above table

(in parentheses), the models for bachelor's degrees per FTE performed much worse at predicting the outcome than those for 6-year graduation rate.

Interpretation

Because the kernel models performed the best on the testing set in terms of RMSE, we decided to interpret those models. In Table X below are the expenditure average partial derivatives from the kernel model that were significant at the 0.05 level for the 6-year graduation rate model.

Table X: Significant (at the 0.05 level) Average Partial Derivatives of Expenditure Variables for the 6-Year Graduation Rate Kernel Model

Expenditure	Estimate	Standard Error	p-value
Academic Support Expenditures per FTE	4.50e-06	1.20e-06	0.0001495
Student Services Expenditures per FTE	1.62e-05	1.70e-06	0
Institutional Support Expenditures per FTE	-1.25e-05	1.20e-06	0
Depreciation Expenditures per FTE	2.90e-06	7.00e-07	0.0000419
Auxiliary Enterprises Expenditures per FTE	2.07e-05	6.00e-07	0

For 6-year graduation rates, student services and auxiliary enterprises expenditures had the most positive average partial derivates and instruction expenditures did not end up being significant. While we are careful to assign any causality from these models, given that all of the expenditure variables are included in these models as well as is the total dollars per FTE that the school has at its disposal in the year (Total Revenue per FTE), this is interesting finding. It would be irresponsible to suggest that money spent to improve instruction does not have an effect on graduation rates (let alone the quality of education for all those at the school), and that all extra dollars should be funneled into student activities, residence halls, intercollegiate sports, or health services. However, these models do suggest that efforts to improve students' experience and make students feel comfortable and supported on campus should be considered.

Automatic Variable Selection

For this approach the idea was to leave as much information available to the model as possible. Variables that were obviously colinear and not of interest were excluded, e.g. full time retention rate is a strong predictor of graduation rate. However this is not of interest, since it seems obvious that if students remain in full time enrollment, they will graduate within a specified time.

With of around 4000 observations, variables with more that 200 missing observations were dropped, with the remaing missing values imputed with a simple linear regression on region and year.

Since Lasso induces sparsity in the coefficient estimates, i.e it sets many of them to be 0, we used it to automatically select the variable with high predictive power for both our outcomes. We used 10 fold cross validation to select the optimal level for the regularisation parameters, 10 here is arbitrary though doe not seeme unreasonable since it is small in relation to our ~4000 observations.

We selected the covariates with the 10 variables with the largest absolute value coefficients, 10 here is arbitrary, though we suspect our models are robust to this since the size of the coeffiecients does decreased reapidly, due to the sparsity inducing Lasso.

Table 1 and 2 give the selected variables for each outcome as well as the value of their Lasso coefficients.

The full definition of variables is given in the data dictionary publically available at https://www.deltacostproject.org/.

There are three main types of variable that the Lasso model has identified as strong predicitors for both graduation rate and numbers of bachelors degrees per fte:

- part time variables
- expenditure variables
- grant variables

Modeling Selected Variables