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

Duncan Clark and Leonard Wainstein

March 17, 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 in themselves 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 efficiently used. It is the efficiency of the expenditure 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 graduation rates for 4 year degrees within 6 years and the number of bachelor's degrees awarded in per full time equivalent student. We analysed data 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/

We were interested in using the method developed in this course to predict the outcome variables but, we were also interested in making valid interpretations of our results. With this in mind we tried firstly only including variables that we believed to be important as in [1] and [2], as well a carrying out automatic variables selection through Lasso regression and using these variables as the basis for or models.

Data Processing

Firstly we restricted ourselves to institutions primarily granting bachelors degrees, and excluded any institutions not included for 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. 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.

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 GDP and unemployment data for each state in line with the approach in [1]####Need to reference properly here####.

Prediction and Training/Test 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 modelling decisions, regarding the dropping of variables and institutions etc.

Since our observations are not independent since we all year for all institutions, we clustered by institution, selecting all years of data into the test or training set for each institution. In addition we also stratified by region, to reduce the variance in our estimators. We were looking at individual institutions, and wish to predict the outcomes based on their covariates, including state fixed effects may have accounted for too much of this variation and lead to poor prediction.

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.

Table 1: Lasso selected variables for graduation rate

Variable	Coeffiecient	Description
other_full_time	0.3876	Number of full-time undergrads not full-time, first-year
fed_grant_pct	-0.2412	Percentage of first, full-time undergrads with federal grants
total_part_time	-0.2004	Total number of part-time students
ptug_share_of_total_pt_enrl	-0.1639	Share of total part-time that are undergraduates
grant01	-0.1619	Total Pell Grants
auxiliary03	0.1453	Revenue from auxiliary enterprises
total_part_time_postbacc	-0.1141	Total number of part-time postbaccalaureate students
eandg02	0.0898	Total education and general expenditures - salaries and wages
bach_deg_share_of_tot_deg	0.08154	Share of total degrees at bachelor's level
fall_cohort_pct	0.08059	Fall cohort as percentage of total undergraduates

Table 2: Lasso selected variables for bachelors per fte

Variable	Coeffiecient	Description
bach_deg_share_of_tot_deg	0.5253	Share of total degrees at bachelor's level
hbcu2	0.4783	Indicator for historically black college or university status
returning_to_total_undergraduate	0.3518	Share of total undergraduates completed first year
eandr_degree	-0.2552	Education and related expenses per degree
fed_grant_num	-0.159	Number of first, full-time undergrads with federal grants
other_ed_related_cost	0.1345	Instruction, student services, and other educational costs
total_enrollment_black_tot	-0.09466	Total enrollment of black students

Variable	Coeffiecient	Description
inst_grant_num	0.09308	Number of first, full-time undergrads with institutional grants
ptug_share_of_total_pt_enrl	-0.08842	Share of total part-time that are undergraduates
credhoursug	0.08716	Instruction hours based on credit for undergraduates

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