Impact of funding, teachers' salaries and class size on student success

Michael Hauptman, February 2019

Abstract

It is well known that in most countries student performance is closely linked to socioeconomic status, but attention is also paid to factors such as funding, class size and teacher salaries to improve student performance, especially in schools in lower socioeconomic areas. We conducted a study to answer the question: What was the impact of differing funding levels, class size and teacher salaries on 2017 school outcomes, measured by SAT scores and graduation rate, across ~ 250 high schools in Massachusetts, USA, using a Kaggle dataset.

Initial analysis using multilinear regression saw no impact of these factors on the overall school population. However the schools differed widely on situational factors like socio-demographics and urban density. Repeating the regression analysis at a segment level found that these factors had a much higher impact, and that the factors differed according to segment, for instance higher teachers' salaries was important in well-off areas, whereas a large number/variety of classes was important in disadvantaged areas.

The conclusions of the study have a number of qualifications, such as the strong collinearity between key dimensions. For instance, there are very few high needs schools in the sample with low expenditure per student.

Motivation

Governments, charities and society has the best education of its children as one of its primary goals, and they make substantial investments in schools and teachers accordingly. Questions of evidence-based confirmation of the value of these investments on student outcomes, so that scarce money can be best directed, are becoming increasingly pressing. For instance, The Smith Family, a prominent Australian education charity, has as one of its activities:

> *"Research and evaluation helps us to measure the outcomes and assess the effectiveness of our support and programs. Evaluation and regular reporting also drive continual improvement across the organisation."*

One common set of education investments are class size, student funding and teacher salaries. A better, evidence-based understanding of how much these factors affect student outcomes can assist administrators and charities like The Smith Family optimize their investments.

Datasets

The primary dataset used was the “Massachusetts Public Schools” dataset on Kaggle. This set combines data collected by the Massachusetts Department of Education:

- Enrollment by Grade
- Enrollment by Selected Population, including economically disadvantaged and with disabilities
- Enrollment by Race/Gender
- Class Size by Gender and Selected Populations
- Teacher Salaries
- Per Pupil Expenditure
- Graduation Rates
- Graduates Attending Higher Ed
- Advanced Placement Participation
- Advanced Placement Performance
- SAT Performance
- MCAS Achievement Results
- Accountability Report

The dataset covers 1,788 schools in 2017, including 250 higher schools (including grades 9-12) that were the subject of this study.

To help understand situation factors we also got population density for Massachusetts towns from the 2010 census.

Data Preparation and Cleaning (1)

In order to provide structure to the analysis, we grouped the data into 3 categories:

1. **Situational data**	Data that describes the school's town/area, e.g.: • Demographics of the area • Urban density • % economically disadvantaged
2. **Controllable factors**	These are factors that the school authorities have some control over, such as: • Class size • # of classes offered • Funding per student • Average teachers' salaries
3. **Outcomes**	These are factors that describe student success, such as: • Average SAT scores. • Graduation rates.

Data Preparation and Cleaning (2)

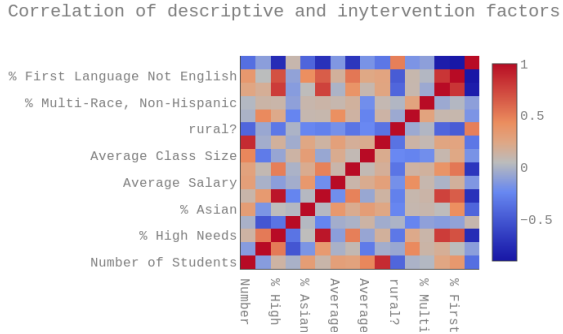
Data preparation and cleaning involved the following:

- Reading the data into a primary dataframe
- Creating a histogram of town density to decide a urban vs rural/suburban cutoff of 1,000 people per square mile, and attaching this as a flag to the main data table
- Identifying the dimensions of interest and dropping the other dimensions from the table
- Filtering on schools that covered grades 9-12, leaving 392 schools
- Dropping schools with missing/NA data, leaving 290 schools
- Slicing the primary dataframe to separate: situation factors (e.g. urban/suburban, %white); controllable factors (funding levels, class size and teacher salaries), and outcomes (average SAT scores, graduation rates)
- Performed a Principle Component Analysis (PCA) on situational factors to attempt to reduce the number of factors in the analysis, though this did not in fact reduce factors
- Used ‘describe’ function to investigate range and distribution of each variable
- Look for correlations and collinearity between data – see next slide

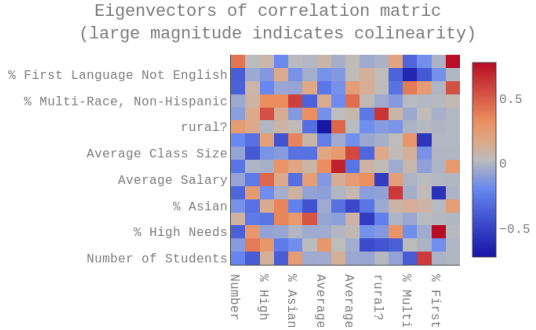
Data Preparation and Cleaning (2)

We discovered some data was strongly co-linear, which qualifies some of our findings

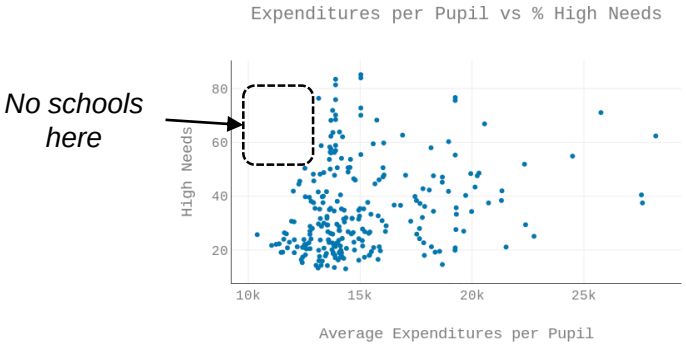
Examining situational and controllable factors together, we discovered several strong correlations



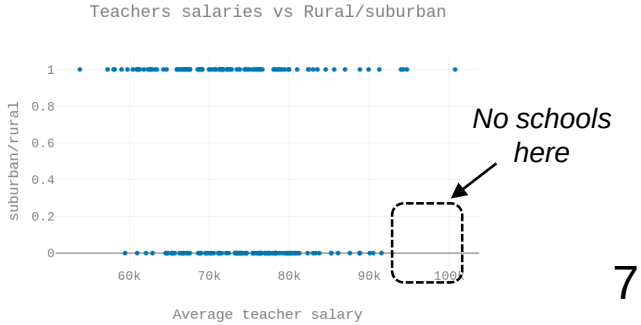
Furthermore, examining eigenvalues and eigenvectors we discovered some co-linearity



For instance, there are very few high needs schools with low expenditure per student



Similarly, there are very few urban schools with highly paid teachers



Research Question

What was the impact of differing funding levels, class size and teacher salaries on 2017 school outcomes, measured by SAT scores and graduation rate, across ~ 300 high schools in Massachusetts, USA?

Methods

Linear Regression

We analysed the relationship between controllable factors (class size, funding per student etc) and student outcomes (SAT scores, graduation rate) using linear regression:

- Scaled investment factors using scikitlearn's 'StandardScaler'
- scikitlearn's LinearRegression package
- 33% test/train split using scikitlearn's train_test_split
- Evaluate strength of regression relationship using scikitlearn's r2_score

Clustering

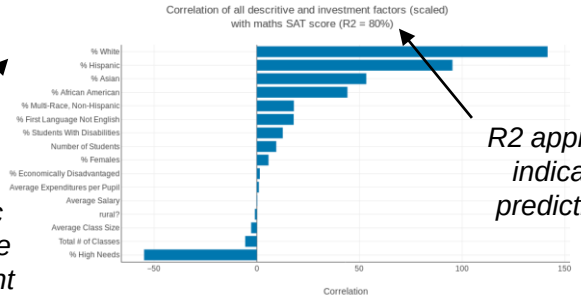
As we did not find strong relationships at the aggregate level we divided the schools into clusters based on situational factors (eg demographics):

- Scaled descriptive factors using scikitlearn's 'StandardScaler'
- scikitlearn's' AgglomerativeClustering to create a dendrite tree to help decide # of clusters
- Split into final clusters using scikitlearn's KMeans

Findings (1)

At the aggregate level, sociodemographic factors are the most important predictors of educational success. Controllable factors like class size appear to have very little impact

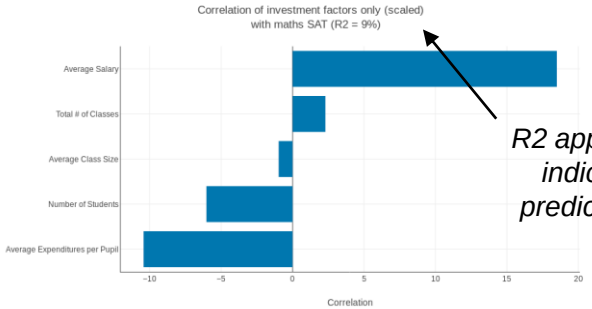
When we combined situational and controllable factors we could describe/predict SAT scores very well



Socio-demographic factors are the most important predictors of educational success

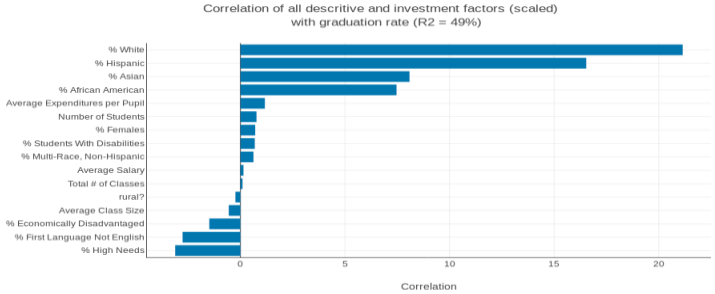
R2 approaching 1 indicates high predictive power

Controllable factors alone had very low descriptive/predictive power for SAT scores …

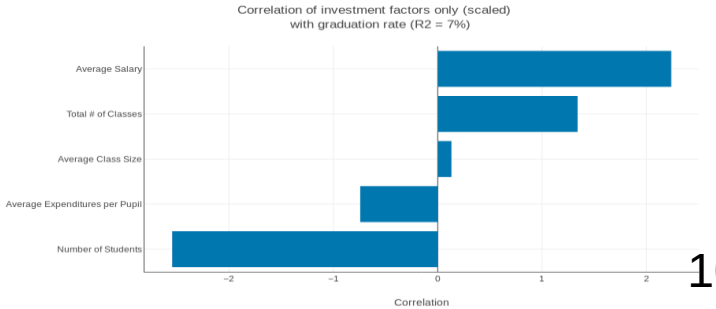


R2 approaching 0 indicates low predictive power

Combining situational and controllable factors predicted graduation rates, but not as strongly as SAT scores



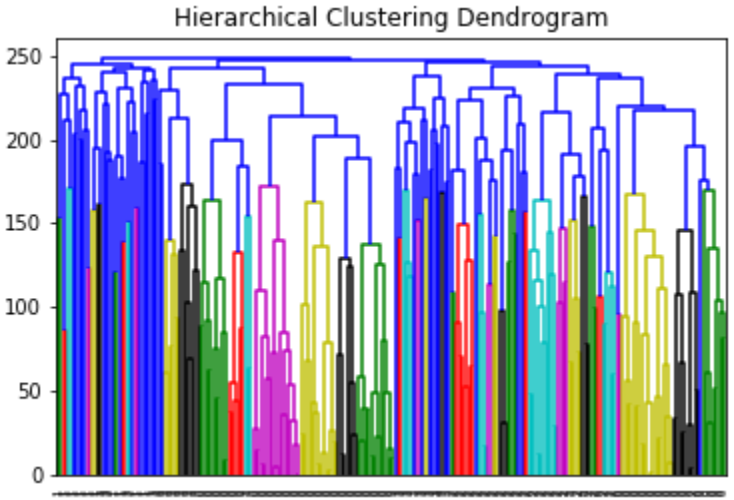
… or for graduation rate



Findings (2)

We then split the schools into 5 segments based on sitiational factors, to see if there were stronger impact of educational investment at the segment level

Agglomorative clustering suggests there were 5 distinct clusters of schools



Good spacing of the branches for 5 clusters/ segments

We then focused on 2 diverse segments:
2 'Urban disadvantaged' and 0 'Suburban privileged'

	0	1	2	3	4
% First Language Not English	3	27	51	7	14
% Students With Disabilities	14	18	17	19	13
% High Needs	23	50	69	43	24
% Economically Disadvantaged	12	36	54	30	11
% African American	2	10	19	4	8
% Asian	3	8	9	2	15
% Hispanic	4	28	51	9	7
% White	89	51	19	82	67
% Multi-Race, Non-Hispanic	2	3	2	3	4
% Females	51	48	48	47	50
rural?	1	0	-0	1	0
prediction	0	1	2	3	4
CLUSTER SIZE	112	27	61	26	24

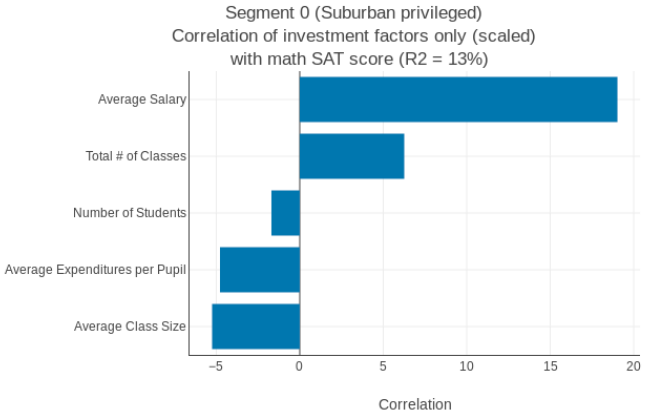
'Urban disadvantaged'

'Suburban privileged'

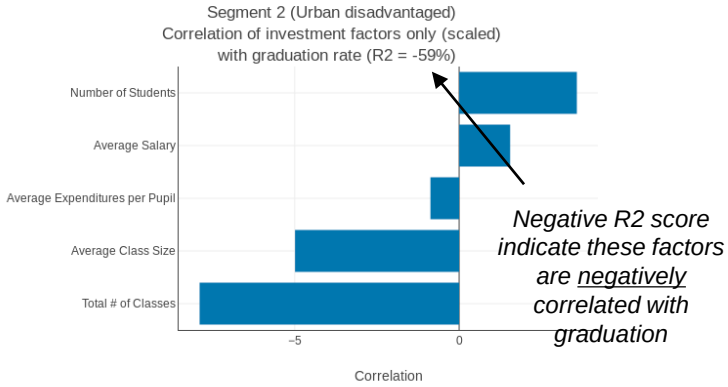
Findings (2)

Repeating the regression analysis at a segment level found these factors had a much higher impact, and the factors differed according to segment

Controllable factors for '**Suburban Privileged**' still have a relatively low impact on success (R2=13%), but much higher than at the aggregate level (R2=9%). The most important factor is average teacher salary



Controllable factors for '**Urban disadvantages**' indicate a high impact on success (R2=59%), and very much higher than at the aggregate level (R2=7%). The most important factor is the number (rvariety?) of classes offered at the school



Limitations

- The data has some strong co-linearity as discussed on page 7, which might mask the real impact of some controllable variables
- There are undoubtably ethical, political and other considerations which may influence decisions on education investment
- The regression analysis at the segment level on page 12 uses small populations, and so is sensitive to which schools are chosen for the training and test populations. The analysis could be repeated with resampling methods

Conclusions

Regarding our question:

> What was the impact of differing funding levels, class size and teacher salaries on 2017 school outcomes, measured by SAT scores and graduation rate, across ~ 300 high schools in Massachusetts, USA?

, we conclude that certain factors do indeed appear to have a significant impact, though these factors differ on the type, and also on the magnitude of impact, depending on the type of school

Acknowledgements

This report was entirely the author’s own work and did not involve any review by third parties.

References

The author would like to acknowledge the kernel/report ‘Exploratory Analysis SAT scores in Public Highschools’ by Luis de Mola for inspiration