Education Economics: Measuring the school effects

Aslan Bakirov

May 2021

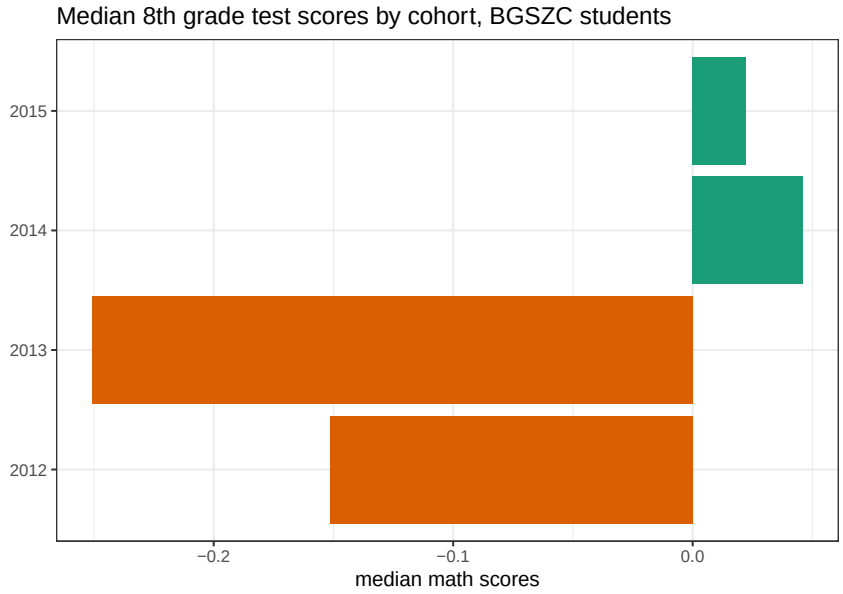
Introduction

In this paper I explore the question whether there are school effects present in one of the high-schools of Budapest, which no longer serves the public.

Data exploration.

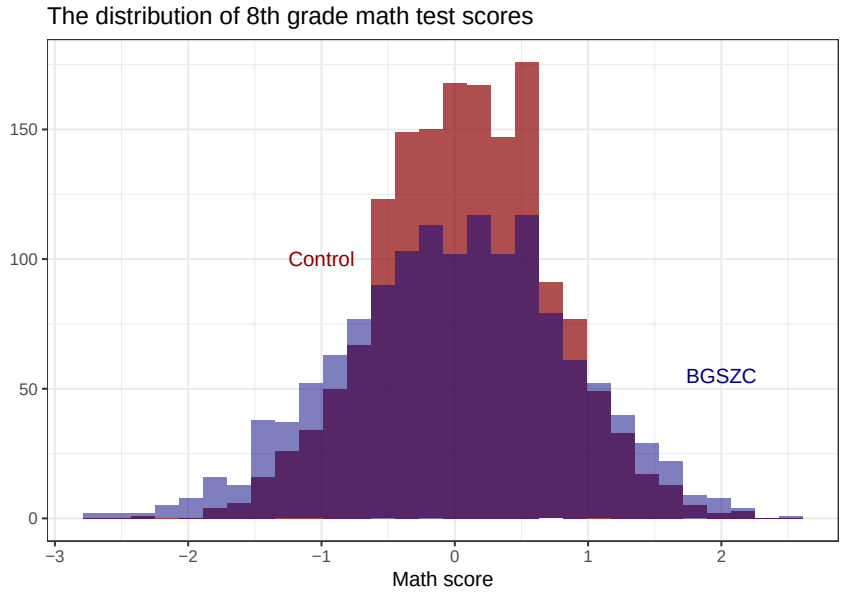
The NABC dataset provides a broad range of variables to explore. One particularly important is the year a student attends the grade. Particularly, there are records of students attending their 6th grade up until 2017. Yet the dataset provides information for the year 2017 the latest, and this fact does not allow us to check the 10th grade performance of those students. To correct for this, I work with a subset of data with students who, by 2010, already attended the 8th grade (*year8>2010*).

The school identifiers were taken for the 10th grade, since the variable of interest is the 10th grade test scores.The school turned to be *203061* with about 1600 students. After getting rid of the *NA* values in the 10th grade test scores 1440 students are left.
Budapest Center of Economic Vocational Training (*Budapesti Gazdasági Szakképzési Centrum*, shortly BGSZC) is a school offering higher-level education, hence only the tenth grade results are present in the dataset. The three types of tracks offered are **four-year academic** (74 students), **mixed** (1273 students), and **vocational training** (93 students). Although according to Oktatás the school was terminated as of 2020, for our purposes it does not matter since we look only at the period before 2017.



Choosing the control group.

To select the control group, I brush up the rest of the schools in the dataset that exactly match the characteristics of the students in the fixed school. First, the test scores of the grade 8 are matched, since the school of interest is offering higher-level tracks. From this I obtain 1600 potential control-group students for math test scores, and then try to ensure as much **common support** as I can by filtering the subset.
Among these 1600 students in the control group, I then take only those who attend the same study tracks as those offered at BGSZC. Furthermore, some of the NA values were also excluded. As a result, I am left with a control group of 1217 students, and 1064 in the treatment. Here I try to explain to what extent the BGSZC affects students' performance in the 10th grade math test.
Another essential moment is to ensure unconfoundedness. Since the selection into the high-school is no longer supporting random pattern, I turn to explain the features that make students select into this particular school. That is, I use **propensity scores** to account for the individual characteristics that better define the probability of attending BGSZC.



Confounders.

The variables I will control for in the following models can be summarized in three groups:

- a) **Mother's education, Free/ discounted school meal**, and **the number of books in the family**.

This group of variables is supposed to serve as a proxy for a student's endogenous characteristic - ability. Ability in the sense that it is inherited, and cultivated in the family. Mother's education is rather straightforward in accounting for the motivation to study further, whereas free/ discounted meal reflects the availability of sibling of the child (since it is common in Budapest to offer discounted meals for pupils whose siblings also attend high-school). The latter can also interfere with the economic background, yet here I assume it is more of a matter how much parental attention each child receives. Next, number of books can make the difference, since parents who are more intelligent tend to buy or own more books. The role of parents is obvious in the former two variables, yet here it is more of a matter of inherited ability. That is, parents who are capable have children who are also capable. The **distance to the school** is also included as to proxy the family income status.

- b) **Year, month of birth,** and **gender**.

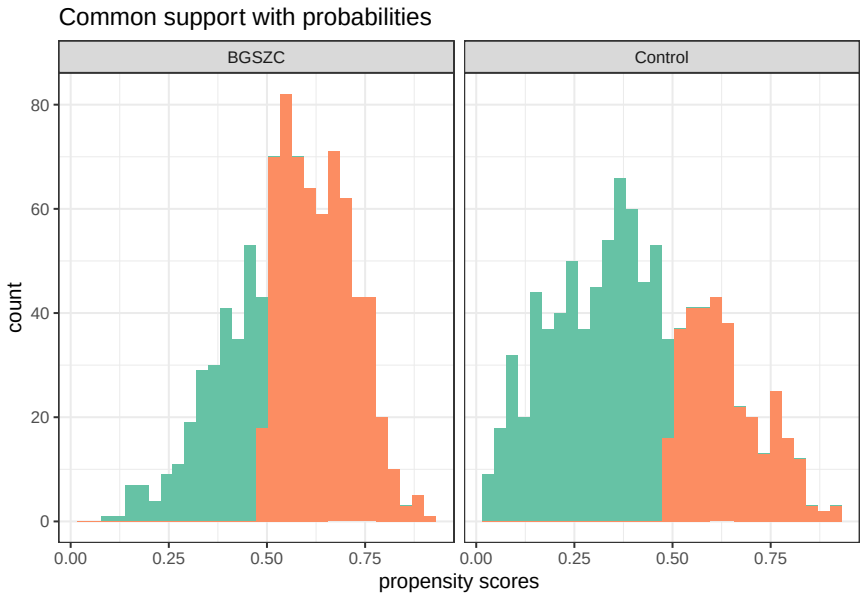
The second group can be referred to as individual exogenous traits. By including month of birth I distinguish the policy effect on the cut-off date when a student can be admitted to school. This is important since if a student does not qualify for the September cohort, because turns the age in November, she can have an (dis)advantage in next year since she would be among the eldest in the class. For teenage students or children this can mean a huge difference in growth.

- c) **Previous scores, track type**.

The last group controls are the academic covariates. Track type of the student is also a variable of interest alongside with the *treatment*.

Methods

To begin with, I estimate the propensity scores to account for the selection bias, and improve unconfoundedness assumption, which essential to move further. I use the logit model with the covariates mentioned above to estimate the probability. The predicted vs actual probabilities are plotted on the histogram below, and suggest that the model leaves out only a modest bit of students who were actually in the treatment.
Since this is also a prediction task, I could use the regularized LASSO model with cross-validation to determine which variables have a major impact on the selection procedure. However, exploring the causality in this case is ambiguous due to the fact that LASSO can assign higher penalty to one of the variables which is correlated with another, leaving crooked importance lists. This is better suited for causal ML methods. Furthermore, the number of covariates is not very large compared to sample size.



Analysis.

First, let's start with a simple regression explaining the math test scores with the school without any confounders, just propensity scores added. It reveals the positive and significant effect of the treatment ($p\text{-}value<0.01$).

Table 1: Treatment and Propensity scores

term	estimate	std.error	statistic	p.value
(Intercept)	0.162	0.049	3.326	0.001
treatment	0.119	0.039	3.043	0.002

term	estimate	std.error	statistic	p.value
preds	-0.444	0.102	-4.368	0.000

In the next regression all the covariates listed prior are included, both for current period (10th grade) and the last period (8th grade), with **inverse propensity scores as weights**. Among interesting findings, there is significant negative effect of the *free school meal*. This is in line with prior expectations, yet to differentiate between the effect of the number of siblings and the income status of the family is hard, especially when the *distance* turns out to be impractical and insignificant. It is possible to argue that families with lots of children prefer to live further from the center, or in wealthier and larger families where parents work hard child-parent quality time is little and thus there's one variable already representing it all.

The greater the number of books at home, the better is the performance on the test. As such, kids who had 300-600 books in 8th grade score *0.2* more than those who had less than 50. In general, owning more books at an early age impacts the scores more than the same amount of books later on.

Girls score *0.15 less* on their math test than boys, vocational track students have the lowest scores, and the mixed track score is *0.21 less* than academic track. Kids born in September fare worse in exams than those born in January and April.

The treatment variable, indicates that the academic track students at BGSZC perform *worse (-0.29)* than the control group, controlling for ability, exogenous individual factors and previous attainment. For instance, the interaction term *treatment:mixed* indicates the coefficient *0.39*, meaning among students of mixed tracks, the conditional average treatment effect is 0.18.

Table 2: All covariates, IPW

term	estimate	std.error	statistic	p.value
discounted meal, class 8	-0.097	0.043	-2.266	0.024
free meal, class 10	-0.186	0.086	-2.148	0.032
50 books owned	0.140	0.053	2.623	0.009
max 150 books owned	0.175	0.052	3.336	0.001
max 300 books owned	0.164	0.057	2.868	0.004
300-600 books owned	0.241	0.061	3.945	0.000
1000+ books at home	0.185	0.082	2.263	0.024
max 150 books owned,class 10	0.089	0.051	1.742	0.082
max 300 books owned,class 10	0.092	0.056	1.656	0.098
300-600 books owned,class 10	0.162	0.060	2.679	0.007
treatment	-0.288	0.077	-3.729	0.000
mixed type	-0.210	0.045	-4.626	0.000
vocational type	-0.617	0.055	-11.133	0.000
math score,class 8	0.602	0.020	30.262	0.000
birth month=April	-0.126	0.067	-1.882	0.060
birth month=September	-0.163	0.064	-2.526	0.012
female	-0.150	0.038	-3.941	0.000
treatment*mixed type	0.388	0.081	4.807	0.000
treatment*vocational type	0.272	0.149	1.823	0.068

Conclusion.

The analysis shows that the treatment group students have lower math scores, controlled for ability, exogenous characteristics, and previous performance. The treatment - BGSZC students, of mixed tracks, however, score higher than other students from the mixed tracks in the control group.

Limitations.

Firstly, this setting is better approached by the clustered methods. Next, to add more variables, the OLS regression is no longer an option, and either kernel regression or causal ML models can be used to exploit high-dimensionality of the data.