Teacher Performance-Based Incentives and Learning Inequality

Deon Filmer
The World Bank

James Habyarimana
Georgetown University

Shwetlena Sabarwal
The World Bank

Online Appendix

Appendix A

Causal Forest Analysis

We use the generalized random forests (grf) algorithm developed by Athey and Wager (2018), Athey, Tibshirani and Wager (2019) to explore heterogeneity in treatment effects.[1] The algorithm is inspired by regression tree ensemble methods used to predict outcomes (Breiman 2001) and adopted to study treatment effects (Athey and Imbens 2016). While regression trees partition explanatory variables to minimize the mean square error in each branch, causal forest algorithms maximize the difference in treatment effects across child nodes. To overcome the tendency of regression tree methods to overfit, the causal forest adopts an honest approach in which half of the sampled data is used to build the tree, while the other half is used estimate the treatment effects in the resulting nodes/leaves.

Data Processing

Below we outline the steps taken to generate the results reported in the paper. Causal Forest takes as its main input a dataset including a single binary treatment variable (in our case we focus on the contrast, in Phase 2, between schools assigned to the repeated teacher incentives treatment and the control group). The main outcome is the average endline score in phase 2 across the three incentivized subjects – Kiswahili, English and Math. Our set of potential predictors of heterogeneity is drawn from baseline characteristics collected at the beginning of Phase 1. The data processing steps to define the set of potential predictors proceed as follows:

- We drop all of the observations assigned to the student incentives and incentives withdrawn arms
- We construct *school level* pre-treatment covariates drawn from school, headteacher and teacher characteristics.
- We drop observations with any missing variables.

Table A1 below shows summary statistics of the set of covariates used in this analysis.

Table A1: School level Covariates for Causal Forest

Variable	Obs	Mean	Std. Dev.	Min	Max
Phase 1 School Baseline Average	408	47.99	6.95	27.86	75.06
Private School	391	0.17	0.38	0.00	1.00
HT Age	386	40.40	9.32	24.00	71.00
HT Female	389	0.15	0.35	0.00	1.00
HT Training	385	0.51	0.50	0.00	1.00
HT Reward	382	0.64	0.48	0.00	1.00
HT Experience	390	13.35	9.99	0.58	57.42
# of neighboring primary schools <= 20kms	389	9.10	11.23	0.00	110.00
# of neighboring secondary schools <= 10kms	387	5.20	8.70	0.00	91.00
Number of classrooms	390	9.56	4.84	2.00	29.00
Electricity	389	0.37	0.48	0.00	1.00
Generator	388	0.20	0.40	0.00	1.00
Uses treated water	388	0.15	0.36	0.00	1.00
With 5km to tarmac	390	0.32	0.47	0.00	1.00
5-20km to tarmac	390	0.18	0.39	0.00	1.00
> 20km to tarmac	390	0.50	0.50	0.00	1.00
Student enrollment	384	355.72	228.35	67.00	1443.00
Students per teacher	382	35.38	28.53	4.00	360.75
Students per classroom	384	40.25	31.37	5.12	512.50
Toilets per student	383	0.04	0.04	0.00	0.47
Log (Form 1 Fees in TZ Shs)	387	9.89	1.80	0.00	14.60
Share teachers absent	386	0.12	0.18	0.00	1.00
Share female teachers	390	0.23	0.19	0.00	1.00

[1] We use the **grf** package: version 0.10.2, published on 11/24/2018

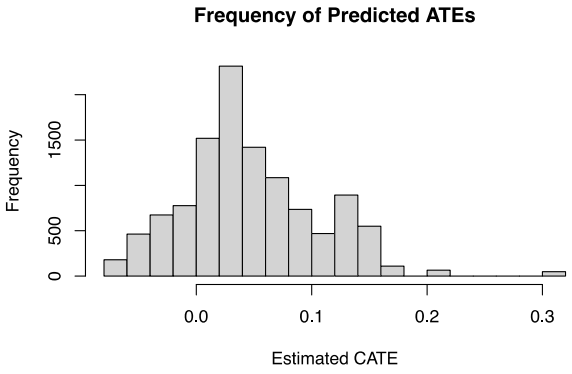
Setting Algorithm parameters

Given that our sample size is sufficiently large, we rely on the algorithm to tune all of the relevant parameters. This involves a selection of 100 initial values for all parameters and a grid search to find optimal parameter values (that minimize error). As we indicate above, we use the honest option in which the generation of trees and the determination of treatment effects in the leaves uses separate samples. Finally, we use the cluster option in which sampling of observations is done within schools.
To increase precision, the algorithm trains a pilot random forest on all of the covariates and then trains the main forest only on the covariates with above median variable importance (a measure of how frequently a variable shows up in the trees, weighted by how high up in the tree it is). The results we present are based on a forest of 10,000 trees.

Results

For each outcome and treatment-control contrast, we use five different seeds to explore robustness of selected predictors of heterogeneity. The panel on the right in Figure A1 below shows the variable importance of selected predictors for each of the five seeds used. School Baseline Average score is the most important variable across all seeds, with a variable importance measure between 19 and 26 percent. Students per teacher (and classroom) are the next most important variables.

Figure A1: Distribution of ATE and Variable Importance

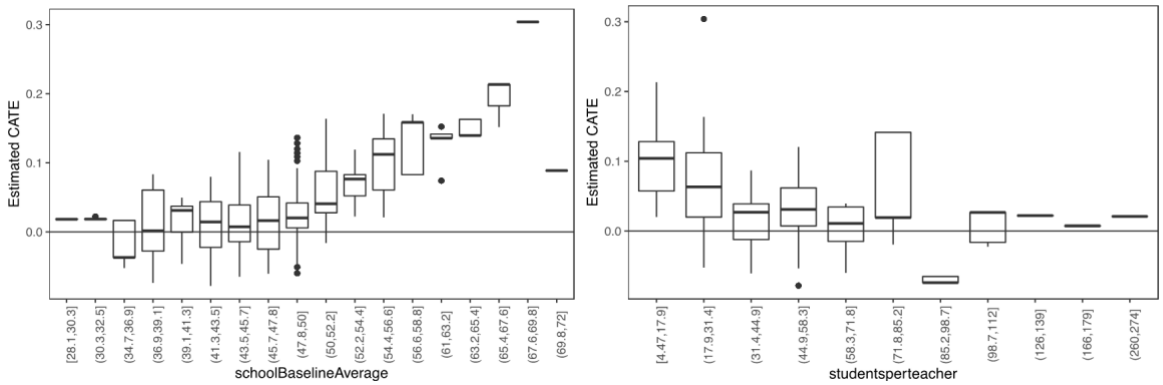


variable	seed1	seed2	seed3	seed4	seed5
Age	10.20	10.67	12.85	12.21	12.11
headteacherexperience	9.67	8.28	10.85	13.64	12.91
logform1fees	10.31	8.65	4.60	10.29	6.77
schoolBaselineAverage	26.16	19.77	22.49	19.19	17.67
sharefemale	NA	7.26	NA	NA	NA
studentperclassroom	11.51	10.39	12.79	14.58	14.03
studentsperteacher	12.50	14.23	17.72	14.62	14.91
teacherabsence	NA	NA	NA	NA	8.16
toiletsperstudent	12.14	8.29	9.44	15.47	13.43

The histogram in Figure A1 above (corresponding to seed 5) shows the distribution of conditional average treatment effects (CATE), across the forest with a modal effect that is small and positive. Importantly, it reveals a relatively long right tail with a few treatment effects above 0.1 SD.[2] We explore how CATE vary with the level of school characteristics in Figure A2 below which shows the pattern for each of the two most important covariates – the school baseline average score and the student teacher ratio.

[2] The median p-value across the five seeds from the test of heterogeneity due to Chernozhukov et al (2018)'s best linear predictor is 0.64 with a range of (0.48, 0.7).

Figure A2: Pattern of Conditional Average Treatment Effects



As the left panel illustrates, the distribution of CATE is centered around zero for schools with a baseline average score less than 50 percent. In fact, up to nearly half of the schools at each level of baseline average score have CATEs that are negative.

For schools scoring above 50 percent, the average treatment effect is positive and rising, reaching a high of 0.3 SD. All other seeds examined produce the same pattern of CATEs across average baseline performance and serves as the basis for the selection of the 50 percent threshold for our heterogeneity analysis.

The panel on the right in Figure 2 shows the relationship between CATEs and the students-teacher ratio. While not as striking as with the baseline school performance, there is suggestive evidence that treatment effects (for the repeated teacher incentives) are larger for schools with lower student to teacher ratios.

References

S. Athey, G. Imbens. 2019. “Recursive partitioning for heterogeneous causal effects”. *Proceedings of the National Academy of Sciences*, 113(27): 7353-60.

S. Athey, J. Tibshirani, and S. Wager. 2019. “Generalized random forests”. *The Annals of Statistics*, 47(2): 1148-78

Athey, Susan, and Stefan Wager. 2019. “Estimating Treatment Effects with Causal Forests: An Application”. *Observational Studies*, 5

Breiman, L. 2001. “Random forests”. *Machine Learning*, 45(1):5–32

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. 2018. “Generic machine learning inference on heterogenous treatment effects in randomized experiments”. Technical report, National Bureau of Economic Research.