Abstract Title Page

Title: Estimating Treatment Effect Variation on Multilevel Data: Learning from ECLS-K data

Authors and Affiliations:

Tianyang Zhang & Bryan Keller
Teachers College, Columbia University

Contact Information: Tianyang Zhang, tz2261@tc.columbia.edu

Abstract Body

Background/Context:

Questions related to the presence and magnitude of causal effects are central to education research and program evaluation. In settings where randomized experiments cannot be carried out, observational study designs may support causal inferences if some assumptions are met. Despite the overwhelming focus on the overall ATE, in most applications it is also interesting to look beyond the average causal effect in order to understand how causal effect vary based on participant background characteristics. The identification of subgroups for which an educational program is highly effective or, on the other hand, has no effect or is possibly harmful, may have important practical implications. For example, understand the heterogeneity of gains due to exposure to special education among children with different family backgrounds and will help policy makers to efficiently allocate resource and prevent potential damage to some families.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is to assess absolute and relative performance of leading methods in estimating heterogeneous treatment effects with the kind of nested data typically encountered in education (i.e. students nested within schools). In recent years, open access to high-quality educational data with natural hierarchical structure and rich variable pools (e.g., ECLS, HSLS, NHES) has made it easier for educational researchers to investigate causal research questions with observational data through computation-intensive methods. These methods adapt and combine machine learning methods with causal inference theory to detect and flexibly estimate heterogeneity along a potentially large number of covariates.

Despite the burgeoning interest in this line of research, there is as yet scant evidence as to the conditions under which the finite sample performance of these methods can be expected to be adequate. Moreover, current literatures mostly generate purely synthetic data with ambiguous recourse to reality. Although a few recent studies have incorporated real data in simulation design (e.g. Wendling et al. 2018; Knaus et al., 2018), results are in healthcare and economics, not education. Indeed, idiosyncratic details common in educational data such as hierarchical nesting (i.e., students within schools, etc.) and strong high-dimensional selection processes may render results from studies grounded in other areas less applicable to education.

Population / Participants / Subjects:

The Early Childhood Longitudinal Study-Kindergarten cohort (ECLS-K; NCES, 2001), is a national longitudinal study focused on child development and early school experiences. The treatment effects of interest are related to the impacts of receiving special education services versus not receiving special education services on mathematics scores measured during fifth grade (year 2004). Following Morgan, Frisco, Farkas, and Hibel (2010), 34 pretreatment covariates were selected based on theory and previous results in the literature that linked them to special education placement. After deleting cases without complete covariate information, 7,362 cases remained, 429 of which were associated with students who had received special education services. The 34 pre-exposure predictors measured at kindergarten are from the following domains: demographic, academic, school composition, family context, health, and parent rating of child. Figure 1 shows the correlation matrix among covariates.

Research Design:

We conduct Monte Carlo simulation studies based on ECLS-K data. For predictor variables, data generation involves simulating from a multivariate normal distribution with:

- marginal means of continuous and binary variables set to match observed sample means
- variance/covariance matrix based on the estimated sample variance/covariance matrix
- binary variables produced by thresholding based on tetrachoric correlations

Let w_{ij} denote the vector of the thirty-four predictor variables for student i in school j. We generate the data from the following model:

$$y_{ij}^{obs} = \alpha_j + \mu(w_{ij}) + Z_{ij}[\tau(w_{ij}) + \gamma_j] + \epsilon_{ij}$$

where μ is a function for generating predicted values based on a neural network model fit to the control arm of the original data, α_j represents a random school effect, γ_j represents school-specific effect heterogeneity, Z_{ij} is the treatment assignment indicator, ϵ_{ij} is idiosyncratic error generated by resampling the residuals from a model fit to the original data, and τ is a treatment effect function of the covariate profile that may be used to induce effect heterogeneity based on particular individual level and group level covariates.

Factors manipulated for the Monte Carlo study include:

- variance of school-level intercept, α_j, which generates different scenarios of intra-class correlation (ICC)
- magnitude of the effect heterogeneity induced through $\tau(w_{ij})$
- cluster size (n = 5, 20, 50)
- number of clusters (m = 50, 100, 200)

The absolute and relative performance of the methods will be compared via root mean square error (RMSE).

Data Collection and Analysis:

Four methods for heterogeneity estimation will be examined: Bayesian additive regression trees (BART), causal forests (Wager & Athey, 2018), causal boosting (Powers et al., 2018), and matching-based methods, i.e. optimal multilevel matching (Pimentel et al., 2018) and optimal full matching with CART summarization (Keller et al., 2019; Chen & Keller, 2019). The combination of propensity scores and these methods will also be examined, e.g. BART/causal boosting with propensity score adjustment. We select these four methods because they have demonstrated good performance for estimation of heterogeneous treatment effects in observational studies (e.g. Wendling et al., 2018; Knaus et al., 2018; Powers et al., 2018; Carvalho et al., 2019) and require minimal tuning. Moreover, these methods, except causal boosting, may be extended in a straightforward manner to nested data, which is of particular interest herein.

Preliminary Findings / Results:

For the original data, kindergarten reading score is an important moderator of the effect of special education services on 5th grade math achievement. In particular, we estimate the average

effect of exposure to special education services to be negative (i.e., harmful) for those students with reading pretest scores lower than 40 and positive (i.e., helpful) for those with reading pretest scores at or above 40.

In Figure 2, preliminary comparison between methods show that, without considering multilevel structure and manipulating (ICC), BART with propensity score adjustment performs better than causal forest, and causal boosting, this is especially true when sample size is large or when propensity score overlaps is sufficient. Nevertheless, more studies in comparing these methods under different scenarios is needed, in order to provide practical guidance to educational researchers on estimating school-level and student-level heterogeneity of treatment effect, and to help policymakers in program evaluation.

Appendices

Appendix A. References

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3-32.

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *arXiv preprint arXiv:1907.07592*.

Chen, J., & Keller, B. (2019). Heterogeneous Subgroup Identification in Observational Studies. *Journal of Research on Educational Effectiveness*, *12*(3), 578-596.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1), 5-86.

Keller, B., Chen, J., & Zhang, T. (2019) Heterogeneous Subgroup Identification with Observational Data: A Case Study Based on the National Study of Learning Mindsets. *Observational Studies,* 5: 93-104.

Knaus, M., Lechner, M., & Strittmatter, A. (2018). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, 43, 236–254.

NCES. (2001). Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual (Tech. Rep.). *Technical Report No. NCES 2001-029*. U.S. Department of Education.

R Core Team. (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from http://www.R-project.org/

Pimentel, S., Page, L., Lenard, M., Keele, L. (2018). Optimal Multilevel Matching Using Network Flows: An Application to a Summer Reading Intervention. *The Annals of Applied Statistics*, 12 (3), 1479-1505.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37 (11), 1767–1787.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37 (23), 3309–3324.

Appendix B. Table and Figures

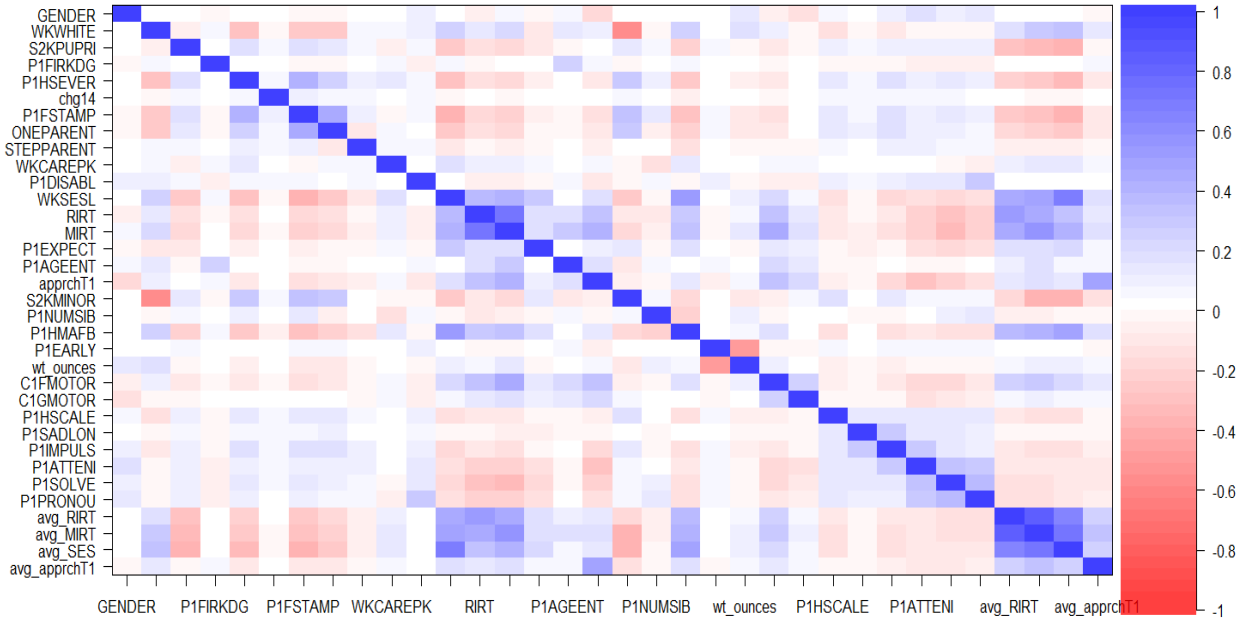

Figure 1: Heatmap of correlation matrix of 34 covariates.

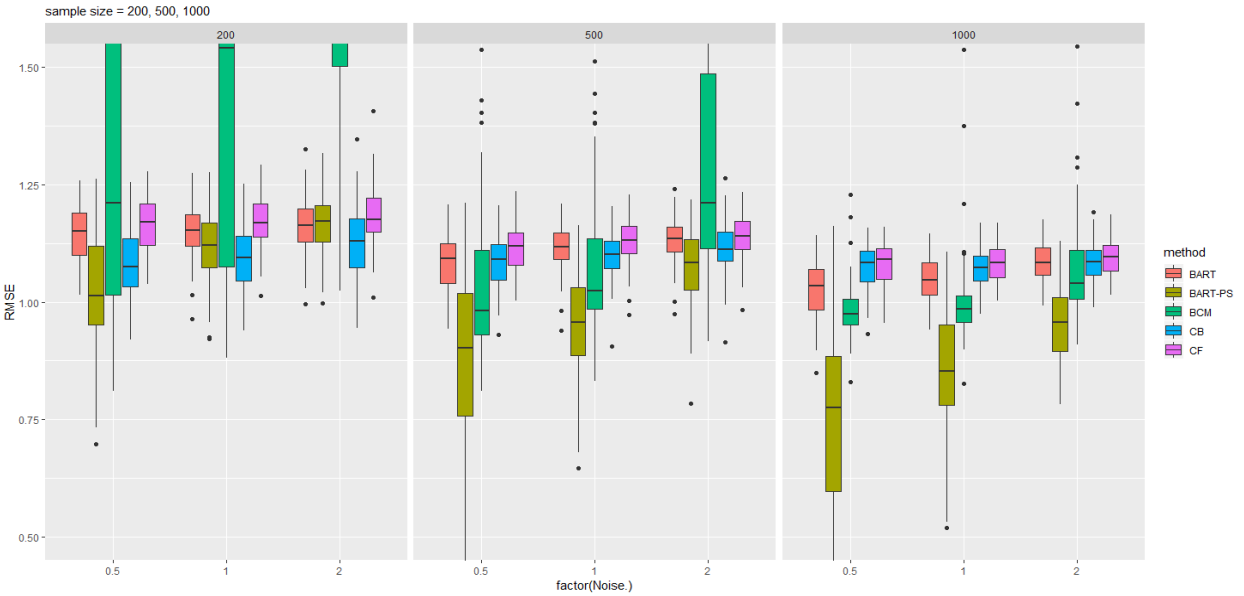

Figure 2: RMSE comparison between five methods under various sample size and noise level.