Detecting Heterogeneous Treatment Effects in an Experimental Study of a Personalized Learning Algorithm

Huan Kuang, Ph.D. Candidate
Walter Leite, Ph.D. & Professor
Jing Zeyuan, Ph.D. Candidate
Corinne Huggins-Manley, Ph.D. & Associate Professor
Research and Evaluation Methodology Program
School of Human Development and Organizational Studies in Education
College of Education, University of Florida

Purpose

Automated personalized learning algorithms have the potential to make a meaningful contribution to student learning by providing learning resources targeting specific needs of students. These algorithms have been implemented within virtual learning environments (VLEs) such as learning management systems and intelligent tutoring systems, which have the capability of implementation of experimental studies within their systems, and quick collection of responses to surveys and psychometric scales, as well as collection of server logs, click streams, time on tasks, and discussion posts (U.S. Department of Education, 2012). Because the access point of VLEs is usually websites or smartphone apps, they can have worldwide reach, giving the opportunity to implement experimental studies using very large and diverse samples. One challenge with these experimental studies is estimating heterogeneous treatment effects (HTE), which is helpful for understanding which subgroups benefited differently from the personalization algorithms. A few methods have been implemented for HTE estimation, with some based on statistical tests (e.g., Ding, Feller, and Miratrix, 2016, 2018; Taddy, Gardner, Chen, and Draper, 2015) and others involving data mining algorithms (e.g., Imai and Ratkovic, 2013). However, the use of any these methods with VLE data for the evaluation of automated personalized learning algorithms has been scarce. The objective of this paper is to demonstrate the application of the methods of structural equation modeling trees (SEMTrees) and causal trees for evaluation of HTE in a large-scale experimental study of a personalized learning algorithm.

Data Source

We used a dataset from an experimental study of a video recommendation algorithm for the Algebra Nation system, which is a VLE that provides algebra tutoring to middle and high school students. The data were provided by the Broward County public school district in Florida. The treatment variable is a binary variable, where the treatment group received personalized algebra video recommendations in the Algebra Nation system during 2019 Spring semester, and the control group received generic recommendations that followed the sequence of algebra topics in the Florida State Standards. There were 14,251 students in this study. Among them, 7,148 were in the control group and 7,103 were in the treatment group. The outcome variable was the students' Algebra I End of Course (EOC) standardized test scores from the 2018~2019 academic year. The EOC is a computer-based and criterion-referenced high-stakes assessment that measures the Florida State Standards for Algebra I. The EOC scores ranged from 425 to 575, with a mean of 500.3 and standard deviation of 31.5 in our sample. Covariates included in this

study were student grade level, race, gender, a dummy first-time test-taker indicator, days of absences, and Florida Standard Assessment (FSA) scores from the standardized math test taken in the previous (i.e., 2017/2018) academic year.

Method

A few statistical approaches have been developed for estimating HTE in experimental studies. To explore the concept of HTE, we investigated two methods. We first explored SEMTrees to identify the heterogeneous subgroups and then estimated the average treatment effect within each node via regression (Foster, Taylor & Ruberg, 2011; Chen & Keller, 2019). This is a two-step approach based on classification and regression trees (Breiman et al. 1984). Foster, Taylor, and Ruberg (2011) used random forests to identify a covariate space for both binary and continuous outcomes. Under the Rubin causal model framework, the authors estimated average outcomes for treatment and control groups using two separated random forests, and then computed the average treatment effect (ATE) as well as estimated average treatment as a function of the covariates (Foster, Taylor, Kaciroti, & Nan, 2016). We argue that the classification/regression trees and random forests could be replaced by SEMTrees. Basically, SEMTrees build trees to separate a data set recursively into subsets based on significantly different parameter estimates in a structural equation model. SEMTrees have been shown to find covariates and covariate interactions that predict differences in structural parameters in observed and latent spaces (Pearl, 1998; Brandmaier, von Oertzen, McArdle & Lindenberger, 2013). *R 3.5.1* (R core team, 2019) and the semtree package (*Brandmaier* & Prindle, 2018) were used to implement this approach.

The second was the causal trees method introduced by Athey and Imbens (2015; 2016). This is a one-step approach. Athey and Imbens (2015; 2016) proposed the causal tree approach to partition the data into subpopulations which differ in the magnitude of their treatment effects. They also constructed confidence intervals for treatment effects without assuming "sparsity", and the change in the variance of treatment effect estimates within each subpopulation was treated as a result of the split. Wager and Athey (2018) extended this approach and developed causal forests. Estimating HTE in a causal forest is a procedure of averaging the HTE obtained from all possible causal trees. *R 3.5.1* (R core team, 2019) and the grf package (Tibshirani, Athey & Wager, 2019) were used for this approach.

Results and Conclusions

With the SEMTrees approach, no heterogeneous subgroup was identified. In other words, the treatment effect was the same across all covariates, and the average treatment effect was 0.005. The causal tree method revealed that the interactions of days of absences and the treatment assignment could be the most important potential indicator for examining HTE. From the total sample, 12% of the students who were absent from school less than two days had an average treatment effect of -0.051 (See Appendix 1) and 88% of the students, who were absent from school two days or more, had an average treatment effect of 0.013. However, the size of the effects was very small, and the package currently does not provide standard errors and significance tests.

The methods investigated are critical for research on VLEs because they have the potential of informing both researchers and practitioners about which subgroups from a large population of students benefit most from personalized learning algorithms. Given that

implementing these algorithms is very costly (Woolf & Cunningham, 1987), it is particularly helpful to find out whether those students benefiting the most are also those in greatest need.

References

Athey, S., & Imbens, G. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, *1050*(5).

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360.

Brandmaier, A. M., & Prindle, J. J. (2018). semtree: Recursive Partitioning for Structural Equation Models. R package version 0.9.13. URL https://CRAN.R-project.org/package=semtree

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological methods*, *18*(1), 71.

Breiman, L. (1984). *Classification and Regression Trees.* New York: Routledge, https://doi.org/10.1201/9781315139470

Chen, J & Keller, B. (2019). *Heterogeneous subgroup identification in observational studies.* Poster presented at the Society of Research Education Effective (SREE) Spring 2019 Conference. Washington, D.C.

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(3), 655-671.

Ding, P., Feller, A., & Miratrix, L. (2018). Decomposing treatment effect variation. *Journal of the American Statistical Association*, 1-14.

Foster, J. C., Nan, B., Shen, L., Kaciroti, N., & Taylor, J. M. (2016). Permutation testing for treatment–Covariate interactions and subgroup identification. *Statistics in biosciences*, *8*(1), 77-98.

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, *30*(24), 2867-2880.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443-470.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, *27*(2), 226-284.

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL https://www.R-project.org/

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*(Feb), 141-158.

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, *34*(4), 661-672.

Tibshirani, J., Athey, S., & Wager, S. (2019). grf: Generalized Random Forests. R package version 0.10.4. https://CRAN.R-project.org/package=grf

U.S. Department of Education Office of Educational Technology. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief.* Washington, D.C.

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.

Woolf, B. P., & Cunningham, P. A. (1987). Multiple knowledge sources in intelligent teaching systems. *IEEE Expert*, 2(2), 41–54. doi:10.1109/ MEX.1987.4307063

Appendix 1 Result from Causal Tree