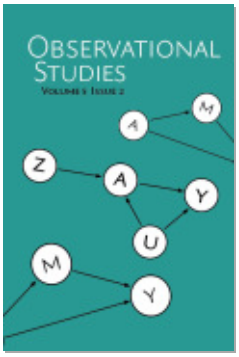
Heterogeneous Subgroup Identification with Observational Data: A Case Study Based on the National Study of Learning Mindsets



Bryan Keller, Jianshen Chen, Tianyang Zhang

Observational Studies, Volume 5, Issue 2, 2019, pp. 93-104 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2019.0010*

Observational Studies 5 (2019) 93-104 Submitted 7/19; Published 8/19

Heterogeneous Subgroup Identification with Observational Data: A Case Study Based on the National Study of Learning Mindsets

Bryan Keller **keller4@tc.columbia.edu**
Department of Human Development
Teachers College, Columbia University
New York, NY 10027, USA

Jianshen Chen **cachen@collegeboard.org**
Learning, Evaluation, and Research
College Board
Yardley, PA 19607, USA

Tianyang Zhang **tz2261@tc.columbia.edu**
Department of Human Development
Teachers College, Columbia University
New York, NY 10027, USA

Abstract

In this paper, we use a two-step approach for heterogeneous subgroup identification with a synthetic data set motivated by the National Study of Learning Mindsets. In the first step, optimal full propensity score matching is used to estimate stratum-specific treatment effects. In the second step, regression trees identify key subgroups based on covariates for which the treatment effect varies. In working with regression trees, we emphasize the role of the cost-complexity tuning parameter, selected through permutation-based Type I error rate studies, in justifying inferential decision-making, which we contrast with graphical and quantitative exploration for future study. Results indicate that the mindset intervention was effective, overall, in improving student achievement. While our exploratory analyses identified XC, $C1$, and $X1$ as potential effect modifiers worthy of further study, we find no statistically significant evidence of effect heterogeneity with the exception of urbanicity category $XC = 3$, but the finding is not robust to propensity score estimation method.

Keywords: Heterogeneous Treatment Effect, Observational Studies, Propensity Score Matching, Regression Trees

1. Methodology and Motivation

1.1 Introduction

Despite the overwhelming focus on the overall average treatment effect (ATE) in the statistics and causal inference literatures, there are many scenarios for which the efficacy of a treatment may vary depending on unit background characteristics. Methods that target *conditional* average treatment effects can explain how pretreatment variables interact with

treatment exposure to cause heterogeneity in treatment efficacy. The identification of such heterogeneity, to the extent that it exists, is of tremendous interest to stakeholders because it can provide insight into which types of participants are likely to be helped the most, helped the least, or even harmed by an intervention. In this paper, we begin with an overview of the synthetic data set generated for the Workshop for Empirical Investigation of Methods for Heterogeneity, a workshop that co-occurred with the 2018 Meeting of the Atlantic Causal Inference Conference in Pittsburgh, PA. We then describe our approach for heterogeneous subgroup identification based on propensity score matching and regression trees. We then discuss data analysis results presented at the workshop, followed by the results of further analyses conducted after the workshop. We conclude with some discussion.

1.2 The Data

The workshop data analyzed herein are synthetic, but were motivated by the National Study of Learning Mindsets, a randomized controlled trial of an intervention designed to encourage a growth mindset in high school students (Mindset Scholars Network, 2018). Approximately 10,000 cases, nested in 76 schools, were simulated to emulate an observational study based on four categorical student-level covariates and six numeric school-level covariates.

The three research questions we were asked to address for the workshop are as follows:

1. Was the mindset intervention effective in improving student achievement?

2. X1 is a measure of the average fixed mindset rating for each school; X2 is a measure of school-level academic achievement; both were measured before the intervention. Researchers suspect either (a) the effect is largest in middle-achieving schools, or (b) the effect is decreasing in school-level achievement. Is there any evidence that X1 and/or X2 moderate the effect of the intervention on student-level academic achievement?

3. Is there evidence that any other covariates moderate the intervention effect?

1.3 Notation

Let Y_i^1 and Y_i^0 be the potential outcomes (Neyman, 1923; Rubin, 1974) under treatment ($Z_i = 1$) and comparison ($Z_i = 0$) conditions, respectively. The average treatment effect, or ATE, is defined as the average of individual treatment effects; that is, ATE $= E[Y_i^1 - Y_i^0]$. A conditional average treatment effect, or CATE, is defined as the average of individual treatment effects, given that a vector of covariates $X_{i1}, X_{i2}, \ldots, X_{ip}$ take on particular values; that is, CATE $= E[Y_i^1 - Y_i^0 | X_{i1} = x_{i1}, X_{i2} = x_{i1}, \ldots, X_{ip} = x_{ip}]$. The propensity score, $e_i(X_i) = pr(Z_i = 1|X_i)$, is the probability that unit i is assigned to (or selects) the treatment group, given the observed covariates. For identification of the ATE and CATE, propensity score analysis, and other conditioning strategies, rely on the strong ignorability assumption (Rosenbaum and Rubin, 1983), which specifies

1. *ignorability*: the potential outcomes are independent of the treatment assignment given observed covariates X; that is, $\{Y^0, Y^1\} \perp\!\!\!\perp Z|X$,

2. *reliable measurement*: observed covariates X have been reliably measured (Steiner et al., 2011), and

3. *positivity*: the propensity score for each unit lies strictly between zero and one; that is, $0 < e_i(X_i) < 1$ for all i.

The observed outcome for unit i, Y_i, is defined via the potential outcomes and the treatment indicator as $Y_i = Z_i Y_i^1 + (1 - Z_i)Y_i^0$.

1.4 Methodology

Our approach to heterogeneous subgroup identification is based on the fact that, under ignorability, $X \perp\!\!\!\perp Z|e(X)$ (Rosenbaum and Rubin, 1983). That is, by conditioning on the propensity score, balance on the observed covariates across treated and comparison groups may be restored to what would have been expected in a randomized experiment; namely, covariate distributions are identical (in the limit) across groups. We use optimal full propensity score matching to stratify units into S strata, each of which contains at least one treated case and at least one comparison case. For each stratum $s \in 1, \ldots, S$, the estimate of the stratum-specific treatment effect is calculated as the difference in sample averages, treated group minus comparison group. That is,

$$\hat{ATE}_s = \frac{1}{n_{T_s}} \sum_{i \in T_s} Y_i - \frac{1}{n_{C_s}} \sum_{i \in C_s} Y_i,$$

where T_s and C_s are, respectively, the sets of indices of the treated and comparison cases in stratum s, and n_{T_s} and n_{C_s} respectively represent the cardinalities of T_s and C_s. Once stratum-specific treatment effect estimates have been calculated, we use those values as estimates of the individual treatment effect for each unit in the stratum. We then regress the vector of individual treatment effects on the set of predictors using a single regression tree. Any predictors identified by the regression tree as important, meaning that the regression tree split on those variables, are interpreted as evidence for effect heterogeneity on the variable or variables involved in the splits.

1.4.1 Regression Trees

A *regression tree* is an algorithmic method invented by Breiman et al. (1984) that models the response surface for an outcome variable, Y, based on predictors, $X_1, \ldots, X_p$, by iteratively splitting units into subgroups based on rectangular regions of predictor values. At each iteration, a split creates two subgroups, called *nodes*, and a node that has not been split is referred to as a *terminal node*. The predicted value based on the regression tree for any unit in a terminal node is simply the mean score on the outcome variable for all units in that node. For unit i in terminal node t, where N_t represents the set of units in t, the tree-predicted value for unit i is simply the mean score on the outcome variable for all units in that node: $\hat{Y}_i = \frac{1}{|N_t|} \sum_{i \in N_t} Y_i$. The deviance for a tree T, $dev(T) = \sum_i \left(\hat{Y}_i - Y_i\right)^2$, is used as a cost function to determine the split point at each iteration. After considering all possible splits on all possible variables, the split that yields the largest decrease in deviance is selected.

If left unchecked, regression trees would continue to split until each terminal node contained only one point. A commonly used approach to prevent this kind of overfitting is based on adding a term to the squared error that penalizes the number of terminal nodes,

$|T|$, in tree T: $C(T)_{cp} = dev(T) + cp|T|$. This approach is referred to as *cost-complexity pruning*, and is implemented in the `rpart` package (Therneau et al., 2015) in R (R Core Team, 2018), which we use to fit regression trees. The tuning parameter, *cp*, is analogous to the smoothing parameter in the lasso or regularized regression, and is typically selected through cross-validation.

2. Workshop Results

In the synthetic workshop data, school sample sizes for the 76 schools ranged from 14 to 529, with a median of 111, and a mean of 136.7. Furthermore, the treatment was non-randomly assigned within schools, such that each school sample contained a proportion of treated cases that ranged from about 17% to about 45%. This design feature allowed us to estimate propensity scores and create matches within schools[1]. As a result of within-school matching, all were matched exactly on the five continuously measured school-level covariates, $X_1, \ldots, X_5$. For the workshop, we used two methods to estimate propensity scores: random forests (RF) and generalized boosted modeling (GBM). Both methods are based on regression trees and, therefore, algorithmically handle interactions and nonlinear relationships.

2.1 Research Question 1

To address the first question, we used standard propensity score methodology and simply took weighted averages of stratum-specific treatment effect estimates. The overall ATE was estimated to be 0.25 or 0.26, based on GBM or RF, respectively, for propensity score estimation. The distribution of estimated individual treatment effects, along with a vertical line denoting the average, is shown for the RF analysis in Figure 1. While the results suggest a positive treatment effect, we did not present standard errors, so we made no claims regarding evidence for an overall effect.

2.2 Research Question 2

We fit regression trees and varied the level of the complexity parameter to search for heterogeneity on X_1 and X_2. With analyses based on propensity scores estimated by RF and GBM, we noted, based on the regression tree output shown in Figure 2, that the treatment effect did appear to vary with X_2 and X_1.

Figure 2 shows the results of a regression tree fit based on random forests with a complexity parameter of 0.0033. Note that at the root node, the overall ATE is estimated to be 0.26 based on 8910 cases. The first split was at $X_2 = -0.71$, which led to conditional ATE estimates of $\hat{\text{CATE}}_{\{X_2<-0.71\}} = 0.12$ and $\hat{\text{CATE}}_{\{X_2\geq-0.71\}} = 0.29$. The next split was also on X_2, thereby modeling a quadratic relationship. In particular, we see that $\hat{\text{CATE}}_{\{X_2\geq0.83\}} = 0.23$ and $\hat{\text{CATE}}_{\{-0.71<X_2\leq0.83\}} = 0.31$. In other words, the estimated average treatment effect for schools with academic achievement scores between -0.71 and 0.83 was 0.31, higher than the estimate of 0.12 for schools with pretest achievement below -0.71, and higher than the estimate of 0.23 for schools with pretest achievement above 0.83.

1. Note, however, that two schools, numbers 11 and 31, were dropped due to insufficient sample sizes of 21 and 14, respectively

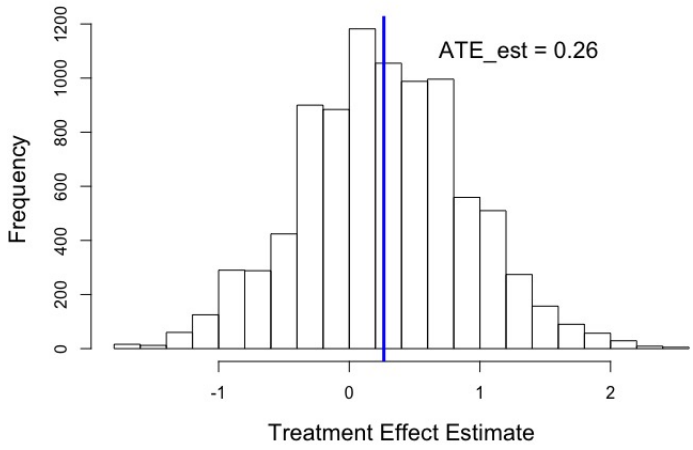


Figure 1: Average Treatment Effect Estimate by the Random Forest (RF) Method

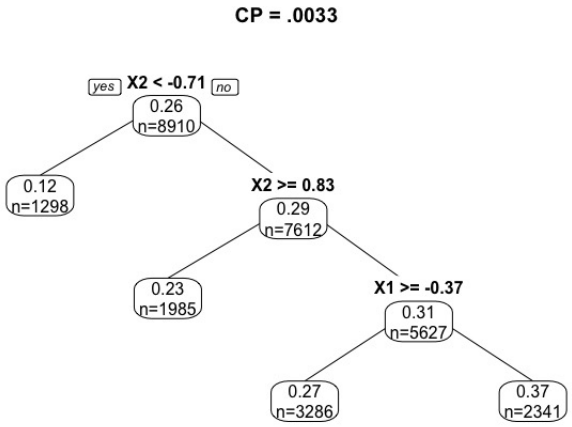


Figure 2: Regression Tree Based on Regressing Individual Treatment Effect Estimates on Observed Covariates; Propensity Scores Estimated by Random Forests

Finally, the last split was on X_1, suggesting that X_1 and X_2 interacted such that, for those schools with X_2 values in the middle range between -0.71 and 0.83, the treatment was more effective for schools with fixed mindset scores lower -0.37 at pretest.

Although we did examine the results of ten-fold cross-validation for *cp* produced by the **rpart** package, we encountered multiple situations in which the cross-validated error rate continued to decrease without bound as the value of the tuning parameter decreased (i.e., favoring more and more complex tree structures; see Figure 3 for an example). The advice given in the **rpart** manual (Therneau et al., 2015) is that "A good choice of *cp* for pruning is often the leftmost value for which the mean lies below the horizontal line."

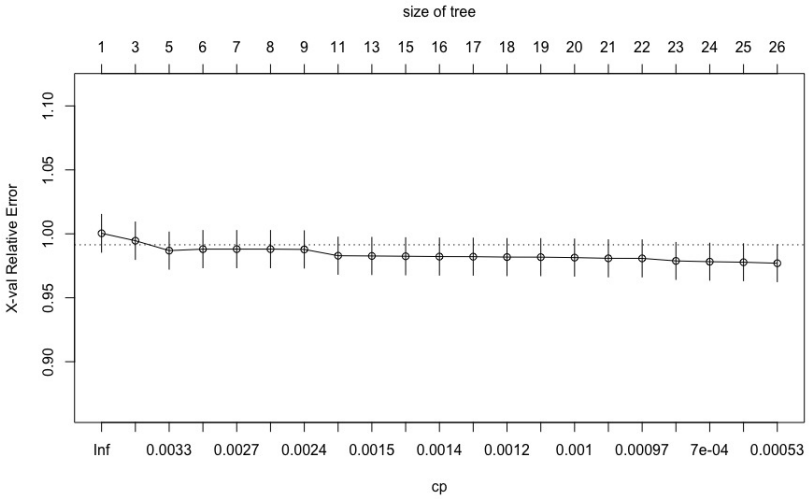


Figure 3: Cross-validated error based on output from package **rpart**; the horizontal line represents one standard error above the minimum value of the cross-validated error curve

where the "horizontal line" represents one standard error above the minimum value of the cross-validated error curve. Despite this rule of thumb, we often encountered tree solutions that were very volatile at the one SE mark. Thus, a limitation of the exploratory approach used for the workshop analyses is a lack of rationale for the selection of the *cp* value, which had the potential to drastically impact results.

2.3 Research Question 3

We noted that student level variables $C1$, a fifteen-category race variable, and XC, a four-category urbanicity variable, were identified in some of the RF and GBM regression tree fits, but did not discuss their roles in detail.

3. Post-Workshop Analysis

For post-workshop analyses, we included main-effects logistic regression (LR) for propensity score estimation, in addition to RF and GBM. Propensity score strata based on optimal full matching were created in each school, as described above. The number of strata per school varied both with the school sample size and the method of propensity score analysis. For propensity scores estimated by GBM, for example, the number of strata per school ranged from 4 for school 13 ($n = 24$) to 161 for school 62 ($n = 529$), with a mean of 36 and median of 27; the numbers of strata based on LR and RF were similar.

Furthermore, we ran a series of Type I error rate studies, using random permutation, to select *cp* values that yielded 5% Type I error rate. Following Chen and Keller (Forthcoming), for each permutation, we shuffled yoked outcome/treatment pairs while leaving covariate values fixed. Under this permutation scheme, the overall average treatment effect and the

covariate marginal distributions and interrelationships remain unperturbed; meanwhile, any dependence between covariates and individual treatment effects is destroyed, which provides recourse to the permutation null hypothesis of no effect (Rubin, 1980; Keller, 2012).

For research questions 2 and 3 for the post-workshop analyses we distinguish between testing and exploration. We test for effect heterogeneity by using cp values that were found, through permutation, to hold the rate of false positives to the nominal 5% level; results based on these cp values are appropriate for inferential decision-making. We explore (a) graphically, by examining graphical depictions of key relationships, and (b) quantitatively, by ranking variable importance ratings from random forest fits. Although these explorations are suitable for hypothesis generation for future study, they are not appropriate for inferential decision-making.

3.1 Research Question 1

We found that the desired nominal Type I error rate of approximately 5% was attained for GBM, RF, and LR, respectively, for *cp* values of 0.006 0.008, and 0.006. The overall ATE estimates were hardly changed when using the *cp* values determined through permutation. For propensity score estimation via GBM, RF, and LR, respectively, the overall ATE estimates, with 95% nonparametric bootstrap confidence intervals (percentile method), were 0.25 (0.22, 0.30), 0.26 (0.22, 0.30), and 0.27 (0.24, 0.28).

3.2 Research Questions 2 & 3

3.2.1 Testing

For propensity scores estimated via LR, and with $cp = 0.006$, one split on variable $XC = 3$ was flagged. No splits were identified using propensity scores estimated via RF with $cp = 0.008$, nor via GBM with $cp = 0.006$. Thus, we found some evidence of effect heterogeneity based on XC, but the finding was not robust to propensity score specification. There was no evidence of significant effect heterogeneity for any other covariates.

3.2.2 Exploration

In Figure 4 we plot nonparametric regression curves to show the relationship between school-level estimates of the average treatment effect on the vertical axis against each school-level covariate. The notion that the intervention was more effective for schools in a "middle range" on $X2$ and with lower values on $X1$ is not inconsistent with the relationships shown in the first two panels of Figure 4.

In Figure 5, because the student-level covariates are categorical, we use conditional boxplots to show how individual treatment effect estimates vary by category across the five student-level covariates. We note what appears to be considerable variability in both median and interquartile range across levels of $C1$, the 15-category race variable. We also note a lower median for category $XC = 3$ as compared with the other categories of XC, a five-category urbanicity variable.

Finally, we fit random forests using the vector of individual treatment effect estimates as outcome and the school- and student-level variables as predictors to calculate variable importance ratings. Because these data constitute a mix of continuously and categorically

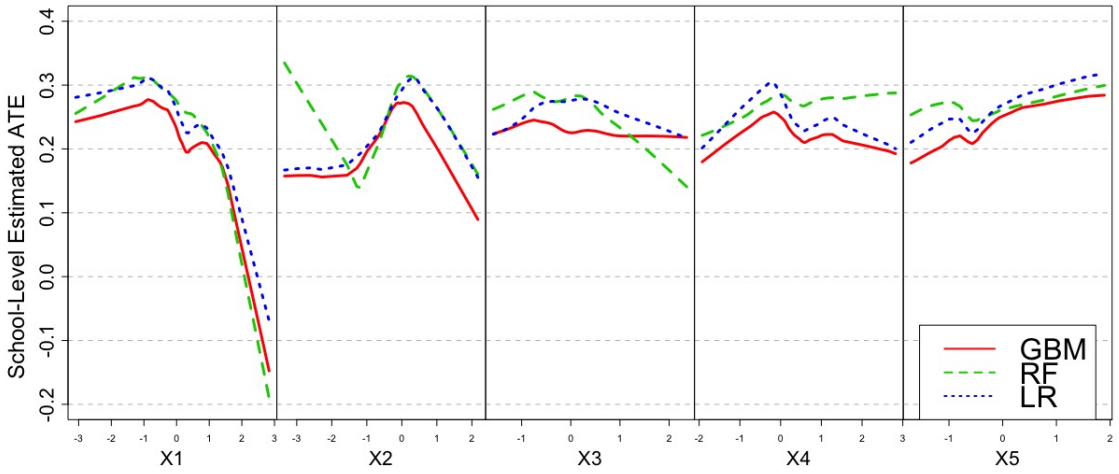


Figure 4: Average School-Level Treatment Effect as a Function of School-Level Covariates

measured predictor variables, and especially because several of the categorical variables have many categories, traditional random forest variable importance (Breiman, 2001) will result in biased importance rankings by unjustly favoring variables with many categories (Strobl et al., 2007). Instead, we report variable importance from random forests based on conditional inference trees, as implemented in R package **party** Hothorn et al. (2006a), which produce unbiased importance values with multi-category predictors.

Conditional inference trees differ from traditional recursive partitioning approaches in that splitting is based on p-values for linear test statistics derived by permutation theory. The p-values are associated with tests of null hypotheses of conditional independence between each predictor and the response, given the tree structure. At each step, these statistics are aggregated to form a global test of the null hypothesis. If the result of the global test is not significant, splitting stops; thus, tree pruning is not needed. If the result of the global test is significant, the p-values for individual predictors are ranked, and the next split occurs on the variable with the smallest p-value. By focusing on p-values, which are not affected by the scales of predictor variables, fair comparisons may be made even for variables on different scales; see Hothorn et al. (2006b) for more details.

After fitting random forests based on conditional inference trees, we find variable XC is ranked as the most important predictor of variability in the individual treatment effect across all three propensity score estimation methods: GBM, RF, and LR. The average importance ranks across the three PS estimation methods identify XC, $X1$, and $C1$ as the three most important predictors, respectively. Notably, $X2$ is among the three least important predictors of effect heterogeneity, according to variable importance rankings.

4. Discussion

We implemented a two-step approach to detect treatment effect heterogeneity characterized by (1) optimal full propensity score matching within schools to estimate individual (stratum-specific) treatment effects, followed by (2) fitting a regression tree of estimated individual treatment effects on covariates. In the analyses prepared for the workshop, we

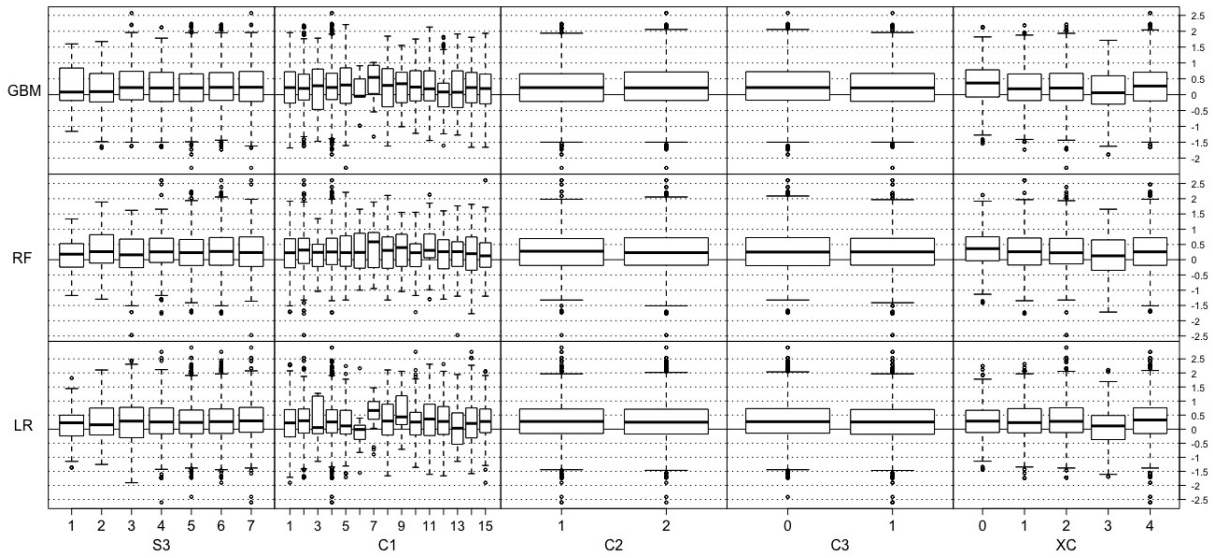


Figure 5: Individual Treatment Effect as a Function of Student-Level Predictors by Propensity Score Estimation Method; GBM = Generalized Boosted Modeling, RF = Random Forests, LR = Logistic Regression

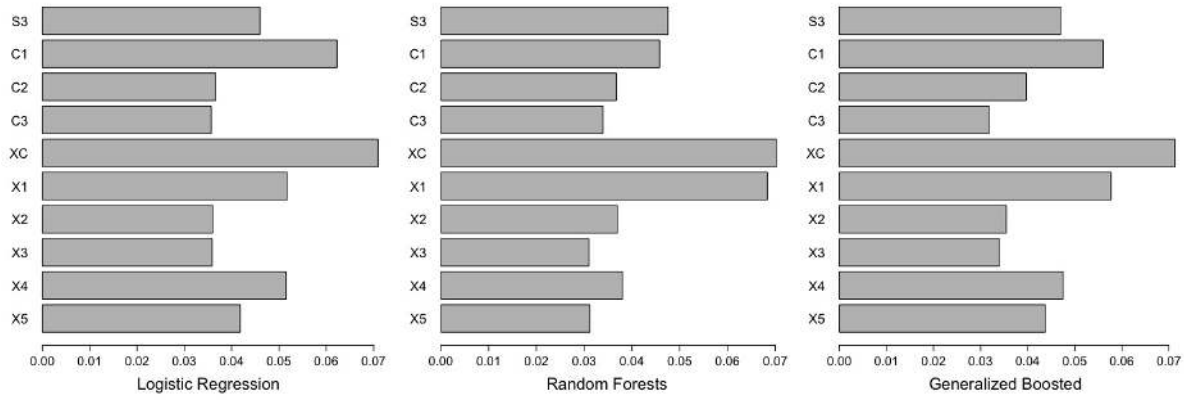


Figure 6: Variable Importance Rankings from Random Forest Runs Regressing the Individual Treatment Effect Estimates on the Ten Predictors of Interest

focused on the second research question by exploring the relationships between $X1$, $X2$, and estimated school-level treatment effects. For the post-workshop analyses, we further demarcated analyses by distinguishing between testing and exploration.

In general, our analyses leaned heavily on the regression tree algorithm, which was used (a) in estimating propensity scores via random forests and boosted modeling, (b) to test for effect heterogeneity through regression tree analysis of individual treatment effect estimates, and (c) for additional exploration through conditional random forest variable importance. With respect to fitting regression trees, we noted that ten-fold cross-validation and the one SE rule of thumb, both methods typically used to select the cost complexity pruning parameter, *cp*, are inconclusive with respect to Type I error rate. Instead, we used a simple permutation approach to select *cp* values that yielded the desired Type I error rate and enabled testing.

For the first research question, we found that the average intervention effect estimates by different methods were all positive, with 95% bootstrap confidence intervals indicating that the mindset intervention was effective in improving student achievement. For the second and third research questions, we found evidence of heterogeneity based on membership in the third category of the urbanicity variable, but the finding was not robust to propensity score estimation method. We found no other significant evidence of treatment effect modification. Based on exploratory analyses, if we were to plan a follow-up study to search for effect modification, we would recommend focusing on the student-level urbanicity variable, XC, the student level race variable, $C1$, and the school-level fixed mindset rating variable, $X1$. We would not recommend prioritizing $X2$, the school-level achievement variable.

As noted by Feller and Holmes (2009), the assumptions required for identification of CATEs are identical to those required for the overall ATE (i.e., strong ignorability, no interference between units, single version of each treatment). We assume these key assumptions are met here. Furthermore, the usual recommended steps for the specification of the propensity score, including iterative respecification to achieve acceptable balance on observed covariates and an examination of overlap are also important, but details are omitted because our focus is on heterogeneous subgroup identification. Finally, resampling approaches such as the jackknife, bootstrap, and boosting may be used to attain error bounds on ATEs and CATEs estimated via our two-step approach; however, care must be taken when using resampling techniques to estimate standard errors for estimators that involve matching (Abadie and Imbens, 2008; Austin and Small, 2014).

Acknowledgments

We thank Carlos Carvalho, Avi Feller, Jennifer Hill, and Jared Murray for organizing the workshop and inviting our submission. Jianshen Chen was employed by Educational Testing Service when this work was carried out.

References

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76:1537–1557.

Austin, P. C. and Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, 33:4306–4319.

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

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, CA.

Chen, J. and Keller, B. (Forthcoming). Heterogeneous subgroup identification in observational studies. *Journal for Research on Educational Effectiveness*.

Feller, A. and Holmes, C. (2009). Beyond toplines: Heterogeneous treatment effects in randomized experiments. Technical report, University of Oxford.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. (2006a). Survival ensembles. *Biostatistics*, 7(3):355–373.

Hothorn, T., Hornik, K., and Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika*, 77:324–338.

Mindset Scholars Network (2018). National study of learning mindsets. Available from: http://mindsetscholarsnetwork.org.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. In *Roczniki Nauk Rolniczych*, volume X, pages 1–51. In Polish, English translation by D. Dabrowska and T. Speed in *Statistical Science* **5**, 465 - 72, 1990.

R Core Team (2018). R: A language and environment for statistical computing. Available from: http://www.R-project.org/.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75:591–593.

Steiner, P. M., Cook, T. D., and Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, pages 213–236.

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, page 8:25.

Therneau, T., Atkinson, B., and Ripley, B. (2015). rpart: Recursive partitioning and regression trees. R package version 4.1-9. Available from: http://CRAN.R-project.org/package=rpart.