Causaltoolbox—Estimator Stability for Heterogeneous Treatment Effects

Sören R. Künzel[*]   srk@berkeley.edu
Department of Statistics
University of California, Berkeley

Simon J. S. Walter[*]   sjswalter@berkeley.edu
Department of Statistics
University of California, Berkeley

Jasjeet S. Sekhon   sekhon@berkeley.edu
Department of Political Science & Department of Statistics
University of California, Berkeley

Abstract

Estimating heterogeneous treatment effects has become extremely important in many fields and often life changing decisions for individuals are based on these estimates, for example choosing a medical treatment for a patient. In the recent years, a variety of techniques for estimating heterogeneous treatment effects, each making subtly different assumptions, have been suggested. Unfortunately, there are no compelling approaches that allow identification of the procedure that has assumptions that hew closest to the process generating the data set under study and researchers often select just one estimator. This approach risks making inferences based on incorrect assumptions and gives the experimenter too much scope for p-hacking. A single estimator will also tend to overlook patterns other estimators would have picked up. We believe that the conclusion of many published papers might change had a different estimator been chosen and we suggest that practitioners should evaluate many estimators and assess their similarity when investigating heterogeneous treatment effects. We demonstrate this by applying 32 different estimation procedures to an emulated observational data set; this analysis shows that different estimation procedures may give starkly different estimates. We also provide an extensible R package which makes it straightforward for practitioners to apply our analysis to their data.

Keywords: Heterogeneous treatment effects, conditional average treatment effect, X-learner, joint estimation.

1. Introduction

Heterogeneous Treatment Effect (HTE) estimation is now a mainstay in many disciplines, including personalized medicine (Henderson et al., 2016; Powers et al., 2018), digital experimentation (Taddy et al., 2016), economics (Athey and Imbens, 2016), political science (Green and Kern, 2012), and statistics (Tian et al., 2014). Its prominence has been driven by the combination of the rise of big data, which permits the estimation of fine-grained heterogeneity, and recognition that many interventions have heterogeneous effects, suggesting

---

[*]. These authors contributed equally to this work.

that much can be gained by targeting only the individuals likely to experience the most positive response. This increase interest of applied statisticians has been accompanied by a burgeoning methodological and theoretical literature and there are now many methods to characterize and estimate heterogeneity; some recent examples include Hill (2011); Athey and Imbens (2015); Künzel et al. (2017); Wager and Athey (2017a); Nie and Wager (2017); Hill (2011). Many of these methods are accompanied by guarantees suggesting they possess desirable properties when specific assumptions are met, however, verifying these assumptions may be impossible in many applications; so practitioners are given little guidance for choosing the best estimator for a particular data set. As an alternative to verifying these assumptions we suggest practitioners construct a large family of HTE estimators and consider their similarities and differences.

Treatment effect estimation contrasts with prediction problems, where researchers can use cross-validation (CV) or a validation set to compare the performance of different estimators or to combine them in an ensemble. This is infeasible for heterogeneous treatment effect estimation because of the fundamental problem of causal inference: we can never observe the treatment effect for any individual unit directly, so we have no source of truth to validate or cross-validate against. Partial progress has been made in addressing this problem; for example, Athey and Imbens (2015) suggest using the transformed outcome as the truth, a quantity equal in expectation to equal the individual treatment effect and Künzel et al. (2017) suggests using matching to impute a quantity similar to the unobserved potential outcome. However, even if there were a reliable procedure for identifying the estimator with the best predictive performance, we maintain that using multiple estimates can still be superior, because the best performing method or ensemble of methods may perform well in some regions of the feature space and badly in others; using many estimates simultaneously may permit identification of this phenomenon. For example, researchers can construct a worst-case estimator that is equal to the most pessimistic point estimate, for each point in the feature space, or they can use the idea of stability (Yu, 2013) to assess whether one can trust estimates for a particular subset of units.

2. Methods

2.1 Study setting

The data set we analyzed was constructed for the Empirical Investigation of Methods for Heterogeneity Workshop at the 2018 Atlantic Causal Inference Conference. The organizers of the workshop: Carlos Carvalho, Jennifer Hill, Jared Murray, and Avi Feller took the National Study of Learning Mindsets, a randomized controlled trial in a probability sample of U.S. public high schools, to simulate an observational study. The organizers did not disclose how the simulated observational data were generated from the experimental data because the workshop was intended to evaluate procedures for analyzing observational studies, where the mechanism of treatment assignment is not known *a priori*.

2.2 Measured variables

The outcome was a measure of student achievement; the treatment was the completion of online exercises designed to foster a learning mindset. Ten covariates were available for

each student: four are specific to the student and describe the self-reported expectations for success in the future, race, gender and whether the student is the first in the family to go to college; the remaining seven variables describe the school the student is attending measuring urbanicity, poverty concentration, racial composition, the number of pupils, average student performance, and the extent to which students at the school had fixed mindsets; an anonymized school id recorded which students went to the same school.

2.3 Notation and estimands

For each student, indexed by i, we observed a continuous outcome, Y_i, a treatment indicator variable, Z_i, that is 1 if the student was in the treatment group and 0 if she was in the control group, and a feature vector X_i. We adopt the notation of the Neyman-Rubin causal model: for each student we assume there exist two potential outcomes: if a student is assigned to treatment we observe the outcome $Y_i = Y_i(1)$ and if the student is assigned to control we observe $Y_i = Y_i(0)$. Our task was to assess whether the treatment was effective and whether the data set exhibits heterogeneity. In particular, we are interested in discerning if there is a subset of units for which the treatment effect is particularly large or small.

To assess whether the treatment is effective, we considered the average treatment effect,

$$\text{ATE} := \mathbb{E}[Y_i(1) - Y_i(0)],$$

and to analyze the heterogeneity of the data, we computed average treatment effects for a selected subgroup S,

$$\mathbb{E}[Y_i(1) - Y_i(0)|X_i \in S],$$

and the Conditional Average Treatment Effect (CATE) function,

$$\tau(x) := \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x].$$

2.4 Estimating average effects

Wherever we computed the ATE or the ATE for some subset, we used four different estimators. Three of which were based on the `CausalGAM` package of Glynn and Quinn (2017). It uses generalized additive models to estimate the expected potential outcomes, $\hat{\mu}_0(x) := \hat{\mathbb{E}}(Y_i(0)|X_i = x)$, and $\hat{\mu}_1(x) := \hat{\mathbb{E}}(Y_i(1)|X_i = x)$, and the propensity score: $\hat{\pi}(x) := \hat{\mathbb{E}}(Z_i|X_i = x)$. With these estimates we computed the Inverse Probability Weighting (IPW) estimator,

$$\hat{\text{ATE}}_{\text{IPW}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i Z_i}{\hat{\pi}_i} - \frac{Y_i(1 - Z_i)}{1 - \hat{\pi}_i}\right),$$

the regression estimator,

$$\hat{\text{ATE}}_{\text{Reg}} := \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right],$$

and the Augmented Inverse Probability Weighted (AIPW) estimator,

$$\hat{\text{ATE}}_{\text{AIPW}} := \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{[Y_i - \hat{\mu}_0(X_i)]Z_i}{\hat{\pi}_i} + \frac{[\hat{\mu}_1(X_i) - Y_i][1 - Z_i]}{1 - \hat{\pi}_i}\right).$$
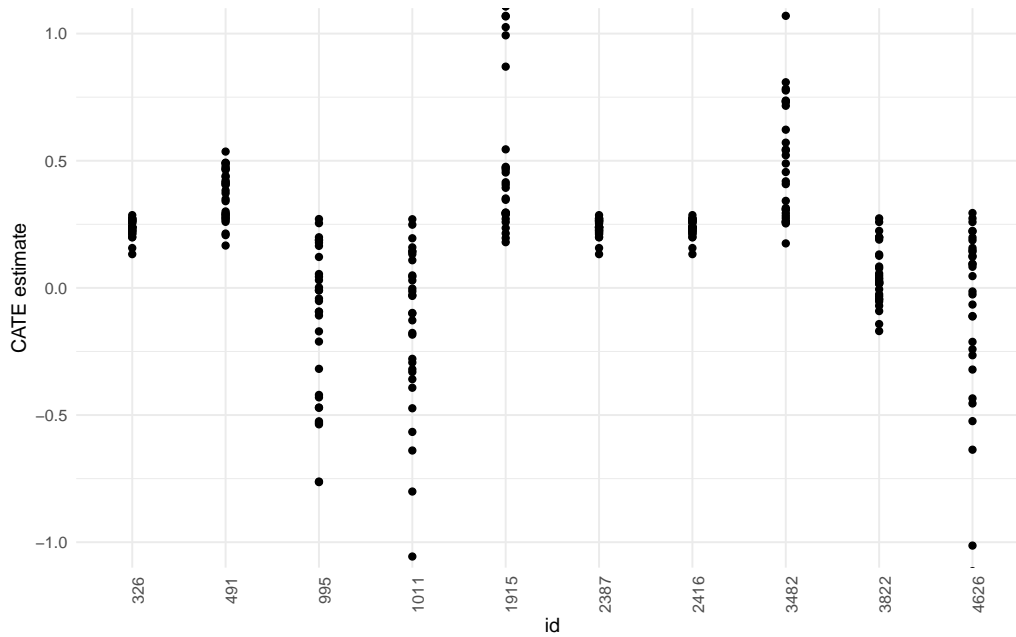
Figure 1: CATE estimation for ten units. For each unit, the CATE is estimated using 32 different estimators.

We also used the `Matching` package of Sekhon (2011) to construct a matching estimator for the ATE. Matches were required to attend the same school as the student to which they were matched and be assigned to the opposite treatment status. Among possible matches satisfying these criteria we selected the student minimizing the Mahalonobis distance on the four student specific features.

2.5 Characterizing heterogeneous treatment effects

In any data set there might be some units where estimators significantly disagree; when this happens, we should not trust any estimate unless we understand why certain estimates are unreasonable for these units. Instead of simply reporting an estimate that is likely wrong, we should acknowledge that a conclusions cannot be drawn and more data or domain specific knowledge is needed. Figure 1 demonstrates this phenomenon arising in practice. It shows the estimated treatment effect for ten subjects corresponding to 32 CATE estimators (these estimates arise from the data analyzed in the remainder of this paper). Some of these estimators may have better generalization error than others. However, a reasonable analyst could have selected any one of them. We can see that for five units the estimators all fall in a tight cluster, but for the remaining units, the estimators disagree markedly. This may be due to those units being in regions with little overlap, where the estimators overcome data scarcity by pooling information in different ways.

In this analysis, our goal was to understand and interpret the heterogeneity that exists in the treatment response. An estimate of the CATE function describes the heterogeneity. However, this estimate is hard to interpret, and drawing statistically significant conclusions

based on it is difficult. Therefore, we sought large subgroups with a markedly different average treatment effects to help characterize the heterogeneity.

Specifically, we split the data into an **exploration set** and an equally sized **validation set**. We used the exploration set to identify subsets, which may have a very different behavior from the rest of the students. To do this, we used all CATE estimators trained on the exploration set and we carefully formulated hypothesis based on plots of all the CATE estimates: For example, based on plots of the CATE estimates we might theorize that students in schools with more than 900 students have a much higher treatment effect than those in schools with less than 300 students. Next, we used the validation set to verify our findings by estimating the ATEs of each of the subgroups.

We randomly split the data into the two sets so that all students from the same school are in the same set. We don't randomize on student level, but on school-level. This is important, because it mirrors the probability sampling approach used to construct the full sample; it also means that we can argue that the estimand captured by evaluating our hypotheses on the validation set is the estimand corresponding to the population from which all schools were drawn.

2.6 CATE estimators

To estimate the CATE we used the default estimators in the `causalToolbox` (Künzel et al., 2018) package. These include four meta-learners: the MO-Learner, the S-Learner, the T-Learner, and the X-Learner (Künzel et al., 2017; Walter et al., 2018). For a specific description and their properties, we refer to the referenced papers. A meta-learner can be combined with any Machine Learning (ML) method, and the idea is that they are combined with ML methods which work well for the given data set. We chose to combine each of them with the random forests algorithm as implemented in `forestry` (**?**) and BART as implemented in `dbarts` (Chipman et al., 2010). We also included `causalForest` (Wager and Athey, 2017b), R-GBM, and R-RF (Nie and Wager, 2017), because they are also tree based CATE estimators and we believe that tree based estimators perform well on mostly discrete and low-dimensional data sets.

However, to protect our analysis from biases caused by using tree-based approaches only, we also included methods based on neural networks. We followed Künzel et al. (2018) and implemented the R, S, T, X and Y-Neural Network methods.

Although we expected that there were school-level effects, and that both the expected performance of each student and the CATE vary from school to school, it was not clear, how to incorporate the school id. The two natural choices are to include a categorical variable recording the school id, or to ignore the school id entirely. The former makes parameters associated with the six school-level features essentially uninterpretable because they cannot be identified separately from the school id; the second may lead to less efficient estimates because we are denying our estimation procedure the use of all data that was available to us. Because we don't want our inference to depend on this decision, we fit each of our estimators twice we fit each of our estimators twice, once including school id as a feature and once excluding it. We considered 16 different CATE estimation procedures; since each procedure was applied twice, a total of 32 estimators were computed.

3. Workshop Results

Our sample consisted of about 10,000 students enrolled at 76 different schools. The intervention was applied to 67% of the students. Pre-treatment features were similar in the treatment and control groups but some statistically significant differences were present. Most importantly a variable capturing self-reported expectations for success in the future had mean 5.22 (95% CI, 5.20-5.25) in the control group and mean 5.36 (5.33-5.40) in the treatment group. This meant students with higher expectations of achievement were more likely to be treated.

We assessed whether overlap held by fitting a propensity score model and we found that propensity score estimate for all students in the study was between 0.15 and 0.46 therefore, the overlap condition is likely to be satisfied.

3.1 Average treatment effects

The IPW, regression, and AIPW estimator yielded estimates identical up to two significant figures: 0.25 with 95% bootstrap confidence interval of $(0.22, 0.27)$. The matching estimator gave a similar ATE estimate of 0.26 with confidence interval $(0.23, 0.28)$.

The similarity of all the estimates we evaluated is reassuring, but we cannot exclude the possibility that the experiment is affected by an unobserved confounder that affects all estimators in a similar away. To address this we characterize the extent of hidden bias required to explain away our conclusion. We conducted a sensitivity analysis for the matching estimator using the `sensetivitymv` package of Rosenbaum (2018). We found that a permutation test for the matching estimator still finds a significant positive treatment effect provided the ratio of the odds of treatment assignment for the treated unit relative to the odds of treatment assignment for the control unit in each pair can be bounded by 0.40 and 2.52. This bound is not very large, and it is plausible that there exists an unobserved confounder that increases the treatment assignment probability for some unit by a factor of more than 2.44. More information about the treatment assignment mechanism would be required to conclude whether this extent of confounding exists.

3.2 Heterogeneous effects

The marginal distribution and partial dependence plots for the 32 CATE estimators as a function of school-level pre-existing mindset norms are shown on the left hand side of Figure **??**. There appears to be substantial heterogeneity present: students at schools with mindset norms lower than 0.15 may have a larger treatment effect than students at schools with higher mindset norms. However the Figure suggests the conclusion is not consistent for all of the 32 estimators. A similar analysis of the feature recording the school achievement level is shown in on the right hand side of this Figure. Again we appear to find the existence of heterogeneity: students with school achievement level near the middle of the range had the most positive response to treatment. On the basis of this figure, we identified thresholds of -0.8 and 1.1 for defining a low achievement level, a middle achievement level, and a high achievement level subgroup.

We then used the validation set to construct ATE estimates for each of the subgroups using the AIPW estimator. We also computed IPW, regression and matching estimators of
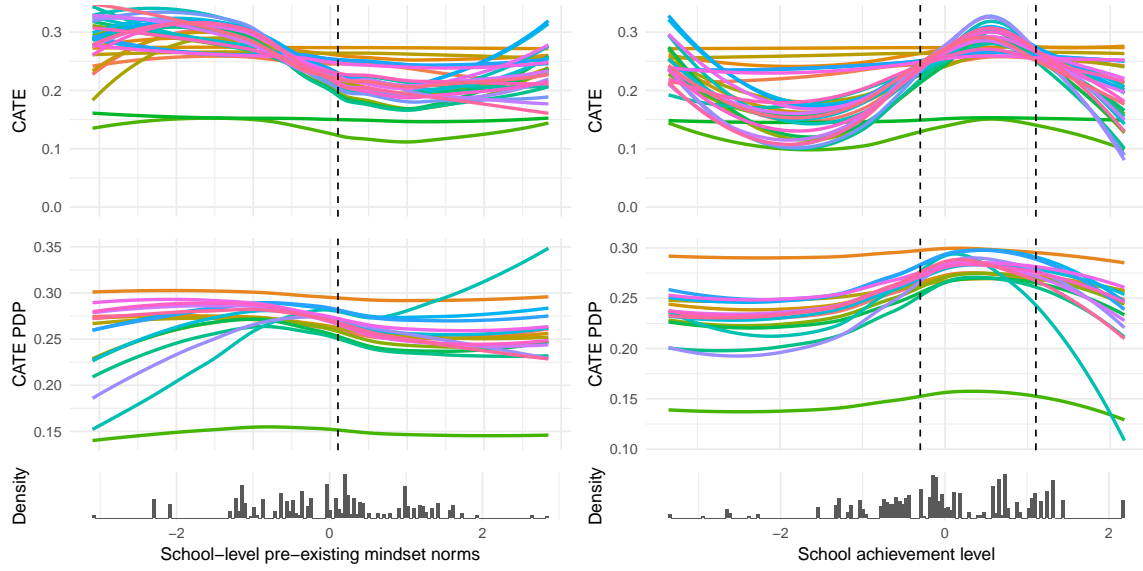
Figure 2: Marginal CATE and Partial Dependence Plot (PDP) of the CATE as a function of School-level pre-existing mindset norms and School achiemvent level.

the ATE but these were practically indistinguishable form the AIPW estimator. We found that students who attended schools where the measure of fixed mindsets was less than 0.15 had a higher treatment effect (0.31, 95% CI 0.26-0.35) than students where the fixed mindset was more pronounced (0.21, 0.17–0.26). Testing for equality of the ATE for these two groups yielded a *p*-value of 0.003. However, when we considered the subsets defined by school achievement level the differences were not so pronounced. Students at the lowest achieving schools had the smallest ATE estimate 0.19 (0.10–0.32); while students at middle and high-achieving schools had similar ATE estimates: 0.28 (0.24–0.32) and 0.24 (0.16-0.31) respectively. However, none of the pairwise difference between the three groups were significant; combining the lowest and highest achievement groups and comparing this combined grouping to the mid-range group did not yield significance at the 0.05 level either.

4. Postworkshop results

For the results presented at the workshop, we focused on measuring whether the treatment was effective on average, and whether there is heterogeneity that can be explained by pre-existing mindset norms or school achievement levels. During the workshop, other contributors found that the variable recording the urbanicity of the schools might explain a lot of the heterogeneity.

The left hand side of Figure 3 shows the CATE as a function of urbanicity, and the right hand side of this figure shows the CATE as a function of the student's self-reported expectation of success. Based on this figure we formulated two hypothesis: students at schools with an urbanicity of 3 seemed to have a lower treatment effect than students at

other schools; students with a self-reported evaluation of 4 might enjoy a higher treatment effect.

These hypotheses were obtained by only using the exploration set; to confirm or refute these hypotheses we used the validation set. The validation set confirmed the hypothesis that students at schools with an urbanicity of 3 had a lower treatment effect (0.16, 0.08–0.24) compared to students at schools with a different urbanicity (0.28, 0.25–0.31); however we could not reject the null hypothesis of no difference for the subsets identified by the self-reported evaluation measure. The urbanicity test yielded a p-values of 0.008 and the self-reported evaluation test yielded a p-value of 0.56.
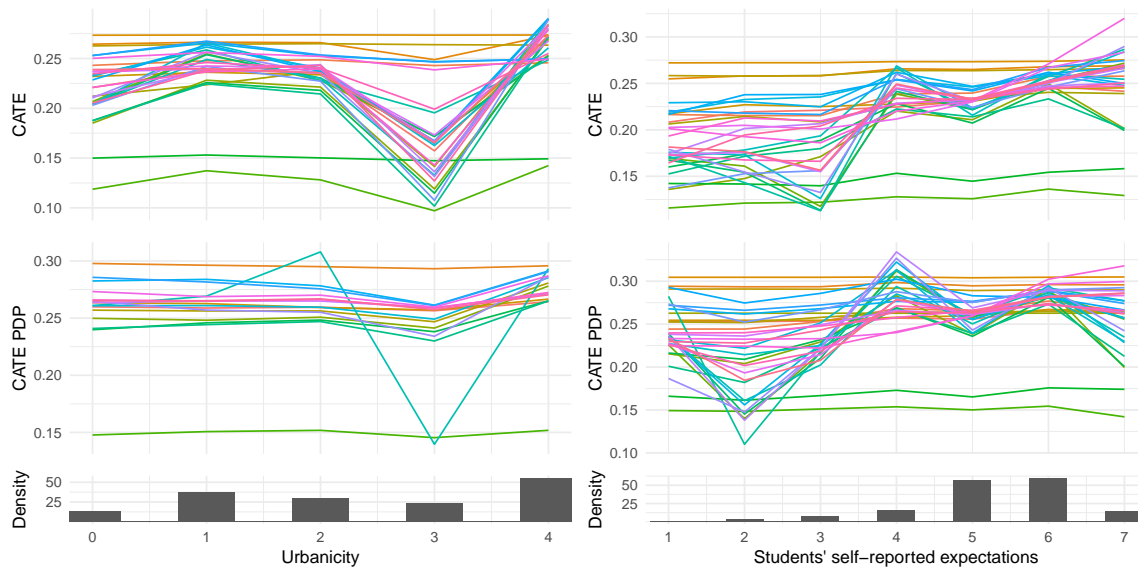

Figure 3: Marginal CATE and Partial Dependence Plot (PDP) of the CATE as a function of Urbanicity and Students' self-reported expectations.

5. Discussion

5.1 The importance of considering multiple estimators

The results of our analysis confirm that point estimates of the CATE can differ markedly depending on subtle modelling choices; so we confirm that an analyst's discretion may be the deciding factor in whether and what kind of heterogeneity is found. As the method-ological literature on heterogeneous treatment effect estimation continues to expand this problem will become more, not less, serious. To facilitate applying many estimation proce-dures we have authored an R package `causalToolbox` that provides a uniform interface for constructing many common heterogeneous treatment estimators. The design of the package makes it straightforward to add new estimators as they are proposed and gain currency.

Differences that arise in our estimation of the CATE function translate directly into suboptimal real world applications of the treatment considered. To see this we propose a

though experiment: suppose we wanted to determine the treatment for a particular student: a natural treatment rule is to allocate her to treatment if her estimated CATE exceeds a small positive threshold or withhold treatment if it is below the threshold. The analyst might select a CATE estimator on the basis of personal preference or prior experience and it is likely that, for some experimental subjects, the choice of estimator will affect the CATE estimated to such an extent that it changes the treatment decision. This is particularly problematic in studies where analysts have a vested interest in a particular result and are working without a pre-analysis plan, as they should not have discretion to select a procedure that pushes the results in the direction they desire. On the other hand, if analysts consider a wide variety of estimators, as we recommend, and if most estimators agree for an individual, we can be confident that our decision for that individual is not a consequence of arbitrary modelling choices. Conversely, if some estimators predict a positive and some a negative response, we should reserve judgment for that unit until more conclusive data is available and admit that we don't know what the best treatment decision is.

5.2 Would we recommend the online exercises?

We find that the overall effect of the treatment is significant and positive. We were not able to identify a subgroup of units that had significant and negative treatment effect and we would therefore recommend the treatment for every student. We are, however, concerned that an unobserved confounder exists. Our sensitivity analysis showed that our findings would still hold if the confounder is not too strong. We cannot, however, exclude the possibility that there is a strong confounder and would have to know more about the assignment mechanism to address this question. This is particularly problematic, because we have seen that students who had higher expectations for success in the future were more likely to be in the treatment group and we must expect that other unobserved confounders exist. Uncovering the heterogeneity in the CATE function proved to be substantially more difficult. We were not able to find subsets with statistically significant different average treatment effects. However, being unable to rejecting a null hypothesis of no heterogeneity, does not verify that the treatment effect is homogeneous, and we generally believe that the data exhibits heterogeneity that could be discovered with more data.

For example, experts believe that the heterogeneity might be moderated by pre-existing mindset norms and school-level achievement. For both covariates, we see that most CATE estimators produce estimates that are consistent with this theory. Domain experts also believe that there could be a "Goldilocks effect" where middle-achieving schools have the largest treatment effect. We are not able to verify this statistically, but we do observe that most CATE estimators describe exactly such an effect. This is notable, since domain experts suspected such an effect without looking at this dataset.

Acknowledgments

References

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5), 2015.

Susan Athey and Guido W Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113 (27):7353–60, 2016. ISSN 1091-6490. doi: 10.1073/pnas.1510489113.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 3 2010. doi: 10.1214/09-AOAS285. URL `http://dx.doi.org/10.1214/09-AOAS285`.

Adam Glynn and Kevin Quinn. *CausalGAM: Estimation of Causal Effects with Generalized Additive Models*, 2017. R package version 0.1-4.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

Nicholas C. Henderson, Thomas A. Louis, Chenguang Wang, and Ravi Varadhan. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services and Outcomes Research Methodology*, 16(4):213–233, Dec 2016. ISSN 1572-9400. doi: 10.1007/s10742-016-0159-3.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Sören Künzel, Jasjeet Sekhon, Peter Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

Sören Künzel, Allen Tang, Ling Xie, Theo Saarinen, Peter Bickel, Bin Yu, and Jasjeet Sekhon. *causalToolbox: Toolbox for Causal Inference with emphasize on Heterogeneous Treatment Effect Estimator*, 2018. R package version 0.0.1.000.

Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*, 2018.

Xinkun Nie and Stefan Wager. Learning objectives for treatment effect estimation. *arXiv preprint arXiv:1712.04912*, 2017.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 2018.

Paul R. Rosenbaum. *sensitivitymv: Sensitivity Analysis in Observational Studies*, 2018. R package version 1.4.3.

Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software*, 42 (7):1–52, 2011.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017a.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017b.

Simon Walter, Jasjeet Sekhon, and Bin Yu. Analyzing the modified outcome for heterogeneous treatment effect estimation. *Unpublished manuscript*, 2018.

Bin Yu. Stabiilty. *Bernoulli*, 19:1484–1500, 2013.