Scientific Research Publishing

Open Journal of Statistics, 2025, 15(6), 445-472
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Services on Academic Achievement: A Robust Estimation Using Bayesian Additive Regression Trees

Yuntian Zuo

School of Public Health, University of Minnesota, Minneapolis, USA
Email: zuo00044@alumni.umn.edu

How to cite this paper: Zuo, Y.T. (2025) Services on Academic Achievement: A Robust Estimation Using Bayesian Additive Regression Trees. *Open Journal of Statistics*, **15**, 445-472.
https://doi.org/10.4236/ojs.2025.156024

Received: October 13, 2025
Accepted: November 24, 2025
Published: November 27, 2025

Abstract

Determining the causal effect of special education is a critical topic when making educational policy that focuses on student achievement. However, current special education research is facing challenges from persistent selection bias and complex confounding. Bayesian Additive Regression Trees (BART) is employed in this study to provide a flexible estimation of the academic performance. Targeted Maximum Likelihood Estimation (TMLE) is also integrated into the BART model, supporting doubly robust estimation of the special education effect. This study extracted survey data from the Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K), to estimate the causal impact of special education status on students' combined mathematics and reading achievement scores. The analysis results of the BART-TMLE model show that children receiving special education services demonstrated approximately 9 points lower scores on average for combined math and reading scores, even adjusting for a considerable number of covariates, compared to their peers who did not receive these services. The estimated negative treatment effect persists after controlling for observed covariates that are closely correlated to the combined test score. The negative effect likely reflects unobserved factors, such as the underlying severity of learning disabilities, parent involvement and other potential traits, which are actual factors that determine the placement of special education status, rather than indicating the ineffectiveness of special education service. The achievement gap in academic performance reflects the current observable status of special education. The estimated effect could be improved by future research incorporating educational domain knowledge, allowing the model to be constructed more accurately.

Keywords

Bayesian Additive Regression Trees, Targeted Maximum Likelihood

Estimation, Unmeasured Confounding, Heterogeneous Treatment Effects, Special Education Causal Impact

1. Introduction

Special education is typically provided for students with disabilities or special needs to accommodate their learning abilities, as they may face difficulties under traditional instruction. The academic performance of children with disabilities may systematically differ from that of their peers, even when holding other relevant student characteristics constant. It is challenging for educators and policymakers that the academic achievement gap persists between students with and without learning difficulties [1] [2]. Previous research has recorded systematic differences in academic achievement between students who receive special education services and their peers. However, the interpretation of learning outcome disparities is contested among researchers, as there is no consensus on the uncertainty of the estimation. Complex variable selection processes and unmeasured confounding factors in observational data are primary challenges for the analysis [3] [4]. Several approaches were utilized in previous studies examining special education effects, including propensity score matching, instrumental variables, and regression discontinuity designs. However, findings from these studies are inconsistent, limiting the reliability of causal effect estimates [5] [6]. Addressing confounding in observational data is one of the challenges in determining the uncertainty of the estimates, where students' characteristics simultaneously influence both special education placement and academic outcomes [7] [8]. Advanced causal inference methods that can flexibly handle complex confounding structures are needed, which is ideal for breaking through the current limitations [9] [10].

The Early Childhood Longitudinal Study (SCLS) program database provides comprehensive information about children's knowledge, skills and development from their birth through elementary school [11]. In order to improve the reliability of the estimated treatment effect of special education services, it is necessary to weight treatment assignment. Typically, the students receiving special education only account for a smaller sample. Therefore, it is necessary to adjust for the imbalance in observations for the treatment and control groups, to get robust results with better sensitivity [12]. BART is a flexible method that can model the outcomes using covariates under different treatment exposure statuses. BART uses posterior simulations to estimate outcomes based on exposure, without the need to meet linearity and additivity assumptions like regression models. This allows for an estimate of the treatment effect without assuming linearity or additivity. This flexibility is very useful when analyzing complex datasets.

Studies in the past have used quasi-experimental methods, like propensity score matching to estimate the treatment effect of special education [13]. Other methods like instrumental variables [14], and regression discontinuity designs [15]

have also been attempted to study this effect. One of the challenges of these methods is that they all require strong parametric assumptions that are difficult to verify. In actual practice, the functional form of the relationship between outcomes and confounders can be very complex, and does not necessarily meet the assumptions. When the assumptions are not met, the estimated treatment effect can be severely biased [10] [12]. The model might fail to adequately adjust for all the effects of the observed confounders, if the actual relationship between the covariates and outcomes is nonlinear, or includes interaction effects. The dependency on these assumptions could be one of the reasons for inconsistent results among the current studies.

This study is an attempt to provide a reliable estimate for overall academic performance using Bayesian Additive Regression Trees (BART). BART is a machine learning model for causal inference that is robust and does not require parametric assumptions [16]. It is a flexible model that works well with nonlinear dependent variables without the requirement of pre-specified functional forms. This trait makes BART models highly robust to misspecification of predicting variables [11]. Using data from the Early Childhood Longitudinal Study (ECLS), students' academic performance in their last year of elementary school is analyzed, adjusting for students' background information, including ethnicity, family context, previous math and reading scores, and health conditions. BART's ability to handle high-dimensional data structures and model heterogeneous treatment effects makes it particularly well-suited for addressing the imbalance of treatment and complex relationships among variables in the special education research [12].

Drawing on nationally representative data from the Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K), this research seeks to produce an accurate estimate of how special education services causally affect academic achievement in math and reading by elementary school graduation. This analysis selects a subgroup of baseline confounders, including demographic variables, family background, earlier academic performance, and school-level characteristics.

The analysis found a consistent and statistically significant negative effect of special education services on combined reading and math scores. On average, children not receiving special education scored approximately 9 points higher than those who did. This negative effect, showing an average treatment effect of approximately −9.1, remained consistent across the full study population, sample, and different subgroups. The 95% credible intervals for these estimates, such as [−14.353, −3.751] for the sample, were entirely negative, providing strong evidence that the observed decrease in scores is a real and robust effect.

This research offers some new insights for special education effectiveness literature and causal inference methodology. The findings of this study indicate a robust average treatment effect suggesting that, on average, children receiving special education services are associated with lower academic achievement. This study contributes to the literature in two key ways. The implemented model demonstrates how BART-TMLE can be used to address confounding in complex educa-

tional settings. The results also present substantive evidence that calls into question the public perception that special education is mostly beneficial. This will allow researchers to examine whether individualized customization is needed in order to improve the overall quality of special education. Last but not least, this study establishes a methodological framework for handling high-dimensional confounding in special education research. Future investigations can use this study as a reference for heterogeneous treatment effect analysis in educational interventions.

The remaining sections will elaborate on the research content sequentially: the data sources will be outlined alongside analytical methodology, followed by the quantitative analysis. The conclusion and discussion of the findings will summarize the results and provide directions for future research.

2. Data and Method

2.1. Data Sources

The study population is retrieved from the Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K), with 7362 observations and 34 variables. This dataset is part of the survey program for national educational database [17], recording the cohort of US children followed from kindergarten through fifth grade. The analytic dataset contains variables recording background information for individual child, family and school to examine academic performance and whether the children receive special education. Demographic and socioeconomic variables are also provided as reference, including gender, race/ethnicity and socioeconomic status. Academic variables capture baseline kindergarten reading and math scores, parental expectations, and whether the child attended the school through head start. School compositions measures (e.g. average peer achievement, socioeconomic mix, and behavioral climate) are aggregated at the school level. Family context variables include household receipt of food stamps, family structure, maternal age at first birth, and number of siblings. Parent-reported child characteristics include health status, attentiveness, verbal skills, and disability status. The exposure variable indicates receipt of special education services, and the primary outcome is the combined final math and reading score at the end of 6^{th} grade. All variables are aligned across surveys for consistency in repeated recording over time. Following a thorough examination of the dataset, no missing values were identified; Therefore, missing data mechanisms will not be applied to the subsequent analysis.

2.2. Bayesian Additive Regression Trees and Causal Inference

Bayesian Additive Regression Trees (BART), Random Forest, and Gradient Boosting Machines (GBM) are popular methods for educational causal inference. Analyzing large datasets with complex demographic and social information using these machine learning methods can be very useful. BART generates full posterior distributions for treatment effects to estimate credible intervals and make infer-

ences based on the estimated effect [10]. Random forests only calculate point predictions without robust estimation of uncertainty. BART is a relatively better choice for robust estimation of the treatment effect compared to Random Forest. BART's Bayesian framework naturally includes regularization using prior distributions, reducing overfitting risks that frequentist methods have in high-dimensional observational data [18]. Additionally, the estimation of treatment effects focusses on the causal effect between different treatment groups. Methods like Gradient Boosting Machines have good predictive performance, but are not the best choice for treatment effect estimation, as they use point estimation that does not measure the error or variability well. BART uses a tree-splitting process to select variables automatically. The model is very interpretable and does not need pre-specified functional form. Policymakers often need the results to be transparent, which makes BART a better alternative, since methods like neural networks do not use "if-then" rules like decision trees to store interpretable decision rationale. The neural network stores the numerical value of the decision weight implicitly [19]. Additionally, BART is very flexible with nonlinear relationships and interaction effects without explicit specification. Complex social and biological background of the students receiving special education makes the flexibility of the BART model very helpful [20]. The regularizing prior structure $\mu_{ij} \sim N\left(0, \sigma_{\mu}^2\right)$ with $\sigma_{\mu}^2 = \left(\frac{k}{2\sqrt{m}}\right)^2$ makes individual trees remains weak learners. Limiting the contribution of each tree prevents any single tree from dominating the overall results. This allows the estimated parameter to be robust, when data include extreme values, or the functional relationship between outcomes and fitted variables are not specified correctly.

BART is a flexible, nonparametric Bayesian approach designed to estimate an unknown regression function: $f(x) = E(Y \mid x)$. BART approximates this function through an ensemble of regression trees:

$$f(x) \approx \sum_{j=1}^{m} g\left(x; T_j, M_j\right)$$

Each component $g\left(x; T_j, M_j\right)$ represents a regression tree characterized by its structure T_j and associated terminal node values M_j. Methods like boosting or random forest could yield similar results [21] [22], but the BART model uses a Bayesian framework. A regularizing prior is placed on the terminal node (leaf parameters) values to make each tree a weak learner. The terminal node values are denoted as $\mu_{jk} \sim N\left(0, \sigma_{\mu}^2\right)$, where $\sigma_{\mu}^2 = \left(\frac{0.5}{\sqrt{m}}\right)^2$ and m denotes the number of trees (set to 200 in this study). Additionally, the model assumes a prior distribution for the error variance: $\sigma^2 \sim \text{Inverse-Gamma}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$ where $\nu = 3$ and λ is calibrated from the data to ensure reasonable scale. These priors promote shrinkage and reduce overfitting of the model [16] [23] demonstrated that

"as m is increased, starting with $m=1$, the predictive performance of BART improves dramatically until at some point it levels off," with $m=200$ representing a "fast and robust option". This model specification improves the balance between computational efficiency and outcome estimation results. While later research suggests that 50 trees are often adequate and sufficient [16] [24], noted that "in applications we have typically used a large number of trees ($m=100$ and $m=200$), as we have found that predictive performance suffers more when too" few trees are used, especially when the data is complex. The MCMC parameters of 1000 burn-in iterations and 4000 post-burn-in samples are used based on the setup from previous studies to ensure the model converges. The convergence of the BART model can be verified using trace plot diagnostics for key parameters including the residual variance σ^2 [16]. The BART model implementation employed 200 trees $m=200$, with MCMC estimation conducted over 5000 total iterations comprising a burn-in period of 1000 iterations followed by 4000 post-burn-in samples for posterior inference. The regularizing prior was set to $\sigma_\mu^2=\left(\frac{0.5}{\sqrt{200}}\right)^2\approx 0.00125$ to ensure each individual tree in the model contributes only a small amount to the final result, which helps prevent overfitting.

Model fitting is carried out using a Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm [16]. In this procedure, each tree is sequentially updated while conditioning on the residuals from the remaining trees: $R_j=Y-\sum_{k\neq j} g\left(x;T_k,M_k\right)$. This iterative Gibbs sampling scheme, based on earlier MCMC-based variable selection methods [25], generates posterior samples of the entire regression function. These samples can then be used to calculate point estimates, credible intervals, and measures of variable importance.

BART is especially useful for causal inference in observational settings, where treatment is not randomly assigned, considering its ability to flexibly model complex nonlinear relationships and interactions, without relying on strong parametric assumptions. Within the potential outcomes framework [26], the observed outcome for unit i is modeled as: $Y_i=D_iY_i(1)+(1-D_i)Y_i(0)$ and estimated using: $Y_i=f\left(x_i,D_i\right)+\epsilon_i$ Using posterior draws, the individual treatment effect (ITE) can be estimated as: $\tau_i=f\left(x_i,1\right)+f\left(x_i,0\right)$ and compute the average treatment effect (ATE) can be estimated as: $\text{ATE}=\frac{1}{n}\sum_{i=1}^{n}\tau_i$. [27] formalized this approach and demonstrated BART's advantages for causal analysis, particularly its capacity to account for uncertainty and flexibly adapt to the data structure.

2.3. Sample-Splitting Framework for Variable Selection in High-Dimensional Bayesian Additive Regression Trees

The validity of causal inference from observational data fundamentally relies on the assumption of no unmeasured confounding $\left(Y^0,Y^1\right)\perp T\mid X$ which requires that all variables jointly affecting treatment assignment and potential outcomes are included in the conditioning set [28]. In practice, this encourages including as

many pre-treatment covariates as possible in the model, as omitting any confounder could introduce omitted variable bias:

$$E\left[\hat{\tau}\right]-\tau=\frac{E\left[\left(X_{omitted}\beta_Y\right)\left(X'_{omitted}\beta_T\right)\right]}{Var\left(T\mid X_{omitted}\right)}$$

where the bias magnitude depends on both the confounder's relationship with the outcome β_Y and its relationship with treatment β_T. The flexible, nonparametric nature of BART makes it particularly compatible for capturing complex confounding relationships without requiring correct specification of functional forms, as the regression trees can automatically model interactions and nonlinearities that might be missed by parametric approaches [29]. In applied settings, it is generally good practice to adjust for a wide range of potential confounders to satisfy the no unmeasured confounding assumption, but practical limitations like computational cost can make it difficult to effectively incorporate high-dimensional covariate sets.

BART's computational demands increase substantially with both the number of covariates and the sample size. With moderate to high-dimensional datasets, it is often impractical to implement the BART model with large burn-in samples and numbers of chains. At each iteration of the Markov Chain Monte Carlo (MCMC) sampler, BART must evaluate potential splits across all p variables, for each terminal node of the m trees. With sample size n, this would result in $O\left(mnp\right)$ per iteration [16]. ECLS-K dataset is consisting of $n=7362$ observations and $p=34$ variables, and requires thousands of MCMC iterations for convergence, the memory requirements exceed typical computational resources as the algorithm must maintain and update m separate tree structures while storing the full $n\times p$ design matrix. Furthermore, the posterior sampling becomes increasingly inefficient in high dimensions due to the dilution of the splitting probability across many covariates, leading to poor mixing of the Markov chain and requiring substantially more iterations for convergence [30]. The effective sample size of the posterior draws decreases as irrelevant variables introduce noise into the tree-growing process, where the probability of selecting an informative variable at each split is only $\frac{q}{p}$ if $q\ll p$ variables are truly relevant [31].

Among the available regularization techniques for high-dimensional variable selection, LASSO (Least Absolute Shrinkage and Selection Operator) was chosen over alternatives such as SCAD (Smoothly Clipped Absolute Deviation) and Elastic Net due to its theoretical properties and practical advantages in the causal inference context. LASSO provides a convex optimization problem with guaranteed global convergence and efficient algorithms [32]. SCAD, needs the model to be tuned using additional hyperparameters. Computational instability is another disadvantage of SCAD method compared to LASSO. Elastic Net combines partitioned sample L1 and L2 penalties on the partitioned sample to optimize correlated predictors. The priority of the double LASSO selection is not to manage mul-

ticollinearity within a single model. It focuses on identifying the minimum set of predictors in the two distinct partitioned samples [33]. Instead of shrinking the values of the parameter using a penalty term, LASSO's exact sparsity property forces the coefficient of predictor to be exactly zero for less relevant variables, providing clean and well-defined variable selection for a for the BART model. [34]. LASSO is well developed for post-selection inference, making the estimated parameters and their confidence interval robust when using established methods [31]. LASSO is computationally efficient, it adapts well to the large-scale administrative datasets, with substantial numbers of potential predictors and observations.

A sample-splitting strategy is used to estimate the treatment effect, to balance valid BART analysis under high-dimensional predictors, and adequate uncertainty for the parameter in the simulation process. Following the principles established by Chernozhukov *et al.* [31] for double/debiased machine learning, data were randomly partitioned $D=\left\{\left(Y_i,T_i,X_i\right)\right\}_{i=1}^{n}$ into two independent subsets: D_1 for variable selection and D_2 for BART inference, where $D_1\cap D_2=\varnothing$.

The 1:1 sample split was used as a conservative strategy to optimize the outcome controlling for both bias and variance in the sample-splitting procedure. This allows the treatment effect to be estimated accurately, with less uncertainty. Using more observations for variable selection (e.g., 7:3 ratio) would provide larger sample sizes for the dual-LASSO procedure. An increased number of observations might identify confounders more accurately. However, this would cause reduced statistical power of the BART analysis for treatment effect on D_2. Alternatively, allocating more samples to the analysis sample (e.g., 3:7 ratio) would maximize power for treatment effect estimation but risk inadequate variable selection due to insufficient data in D_1 that might not meet the assumption of conditional independence $Y\left(0\right),Y\left(1\right)\perp T\mid X_S$ [31]. The equal 1:1 split allows each group to have adequate sample sizes for analysis and variable selection. This is critical, as students receiving special education only account for a small proportion of study population (approximately 6% in the ECLS-K sample). The evenly split samples also facilitate robustness checks through sample re-splitting. Specific composition of D_1 and D_2 will have manageable impact on the analysis if there is no preference given to either of the partitioned samples. Previous research suggests that an equal split of the study sample can be asymptotically optimal under mild regularity conditions theoretically, especially when estimation of the parameter of interest and covariate selection are equally important [35]. Although the partitioned data is not able to use the full sample to estimate the treatment effect of special education, thus sacrificing the efficiency of the estimated parameter, the model is able to model the estimation uncertainty at this cost, accounting for loss of parameters in the selection process.

Variables associated with treatment assignment T and outcome Y is selected from the dual-LASSO procedure, focusing on confounding by solving two separate regularized regression problems:

$\widehat{\beta^T} = \arg\min_\beta \frac{1}{|D_1|}\sum_{i\in D_1} \ell\left(T_i, X_i'\beta\right) + \lambda_T \left\|\beta_1\right\|$ for treatment prediction and another for prediction of the outcome. Penalty parameters λ_T and λ_Y are chosen using cross-validation [36]. The selected variable set $S = \left\{j : \widehat{\beta_j^T} \neq 0\right\} \cup \left\{j : \widehat{\beta_j^Y} \neq 0\right\}$ captures potential confounders satisfying the conditional independence assumption $\left(Y^0, Y^1\right) \perp T \mid X_S$ necessary for causal identification [28]. Critically, because D_2 is independent of the selection event $\{S = s\}$, the fitted BART model on D_2 produces posterior intervals for the average treatment effect $\tau = E\left[Y^1 - Y^0\right]$ with nominal frequentist coverage properties. The posterior distribution $p\left(\tau \mid D_2, S\right)$ is computed on data that played no role in determining S [37] [38]. This method avoids post-selection inference bias that would occur when identical data gets used for both variable selection and effect estimation, where posterior credible intervals would fail to account for the extra uncertainty from the selection procedure, causing under-coverage [20]. The sample-splitting framework therefore offers a practical solution balancing computational complexity with model validity, ensuring causal estimates captures adequate uncertainty with reduced dimensionality (Figure 1) (Table 1).

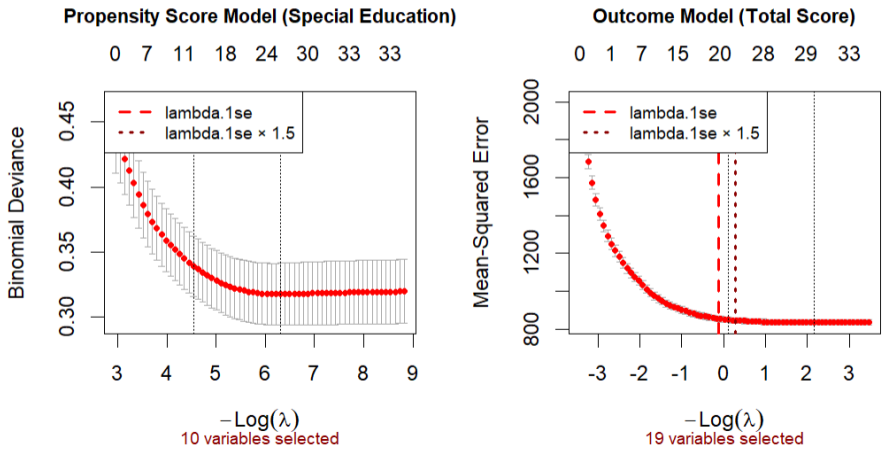

Figure 1. LASSO regularization paths for doubly robust variable selection in special education causal inference (Variables selected from both models: Kindergarten Reading Score, Kindergarten Math Score, Public School Attendance, First-Time Kindergartener Status, Approaches to Learning).

Table 1. Shrinkage coefficient of variables selected by both treatment and outcome variable.

Shrinkage Coefficient From LASSO Selection Process			
Variable	Type	Treatment Coefficient[1]	Outcome Coefficient
RIRT	Confounder	−0.004	0.204
MIRT	Confounder	−0.02	2
S2KPUPRI	Confounder	0.065	0.605

Continued

P1FIRDGE	Confounder	−0.543	13.819
apprchT1	Confounder	−0.268	5.322
C1FMOTOR	Confounder	−0.089	2.195
P1SOLVE	Confounder	0.163	−2.736
P1PRONOU	Confounder	0.491	−2.521

[1]8 variables selected.

The regularization path graph shows that the full model with 34 variables is reduced to a smaller variable set after the LASSO procedure. The coefficients of the treatment model (F5SPECS: special education condition) shrinks with higher lambda values. Most coefficients reach zero by $\log(\lambda)$ around 4, selecting 10 variables at $\lambda = 1.5\lambda_{1st}$. The outcome model (all score: total score of math and reading) shows a similar pattern selecting 19 variables at the specified penalty. The dual-selection procedure identified 8 confounding variables (RIRT: Kindergarten Reading Score; MIRT: Kindergarten Math Score; S2KPUPRI: If the child attended public school; P1FIRKDG: if the child is first time kindergartener; apprchT1: Approaches to Learning Rating; C1FMOTOR: Fine Motor Skills; P1SOLVE: Problem Solving; P1PRONOU: Verbal Communication) The 8 selected variables are related to the treatment group the child receives and the combined academic score, meeting the conditional independence assumption for unbiased estimation of the treatment effect. Instead of using the union of the selected variables from the treatment and outcome model, the intersection of the two models is used. The actual confounding variables are related to the estimated outcome, as well as the treatment group assigned. Using intersecting variables that related to both special education placement and combined test score is more likely to capture the true confounders requiring adjustments. By focusing on the intersection, the model reduces complexity by avoiding variables that may introduce unquantifiable uncertainties and reduced precision on the estimation. Indeed, using the intersection could lead to potential problem for quantifying uncertainty in the model. Less variability of the data is captured by the model when using the intersection. However, the trade-off is carefully considered between bias and variance. Using intersection risks omitting more predictors and potentially sacrificed precision, it prioritizes the inclusion of variables that are vital for confounding control. Under the assumption the true confounders C satisfy $C \subseteq S_y \bigcap S_A$ (confounder effect both treatment and outcome), the mean squared error $MSE(\hat{\tau}) = Bias(\hat{\tau})^2 + Var(\hat{\tau})$ is minimized by the intersection. Since LASSO procedures independently sort out the intersecting variables, the risk of omitting key confounders is reduced. The role of intersecting variables in both treatment mechanism $P(A \mid X)$ and outcome mechanism $E[Y \mid A, X]$. For a variable, $X_j \in S_y \bigcap S_A$, both $\frac{\partial P(A=1 \mid X)}{\partial X_j} \neq 0$ and $\frac{\partial E[Y \mid A, X]}{\partial X_j} \neq 0$, consisting with the definition of con-

founding. Selecting variables that predict the outcome well but are not useful for propensity score estimation can introduce post-treatment bias to the estimation. Post-treatment variables refer to the relationship, where the treatment assignment does not directly affect the outcome, but operates through a mediator, such as student's learning habits. Post-treatment bias could fluctuate the total effect, reducing the efficiency of the estimator. LASSO selection does not explicitly distinguish mediators from confounders, as it emphasizes prediction accuracy.

3. Quantitative Analysis

3.1. BART Framework Estimating Treatment Effect

BART framework estimates two response surfaces: $\mu_0(X_i) = E[Y_i \mid T_i = 0, X_i]$ for the control potential outcome and $\mu_1(X_i) = E[Y_i \mid T_i = 1, X_i]$ for the treated potential outcome. Each response surface is approximated by a sum of m regression trees, shrinkage factor is applied to the result to avoid model overfitting [28]. From these fitted models, three causal estimands are derived using posterior simulation. The Population Average Treatment Effect (PATE) estimates the expected causal effect for the entire sample by integrating over the population's covariate distribution: $\tau_{PATE} = E_X\left[\mu_1(X) - \mu_0(X)\right]$. The Sample Average Treatment Effect (SATE) is representing the average effect specifically for sample observations by averaging individual treatment effect:

$\tau_{SATE} = \frac{1}{n}\sum_{i=1}^{n} E_X\left[\mu_1(X) - \mu_0(X)\right]$. The Conditional Average Treatment Effect (CATE) estimates the heterogeneous treatment effect by estimating the effect for a specific set of variables used in the model: $\tau(X_i) = \mu_1(X_i) - \mu_0(X_i)$. The credible intervals record the variability of the three estimands, showing the uncertainty from the study sample and predictors used to specify the model, with posterior samples. Each tree of the BART model recorded the local pattern of the response surface. The additive combination models global trends. CATE has the advantage of estimating average treatment effect without requiring too much information on the structure of the sample, or the relationship between the covariates and outcome. Compared to PATE and SATE, CATE is a better choice for estimating the effect while adjusting for used variable used, which makes the results more interpretable [39].

3.2. Targeted Maximum Likelihood Estimation

Combining Targeted Maximum Likelihood Estimation (TMLE) and BART helps to produce doubly robust results that offsets potential biases in models that solely rely on outcome, using BART's flexible nonparametric property [40]. TMLE uses a two-stage procedure: first, initial estimates of the outcome regression $\overline{Q_0}(A,W) = E[Y \mid A,W]$ and propensity score $g_0(W) = P(A=1 \mid W)$ are obtained, allowing the outcome model to reflect nonlinear relationships and interactions without relying on pre-defined parametric assumptions. Following the first step, TMLE performs a targeted bias-correction step that refines the original out-

come predictions using information from the propensity score model through a cleverly constructed covariate $H(A,W)=\frac{A}{g(W)}-\frac{1-A}{1-g(W)}$. The clever covariate up-weights observations that are under-represented in their treatment group [41]. This fluctuation step involves fitting a parametric model (typically logistic regression) with the clever covariate as the sole predictor and the initial predictions as offset, yielding updated predictions $\overline{Q_1}$ that solve the efficient influence function equation, thereby achieving asymptotic efficiency under correct specification of either model.

The BART model generates flexible, non-parametric initial estimations $\mu_0(W), \mu_0(W)$ and $g_0(W)$ through the ensemble of regression trees. TMLE utilize the flexible estimates generated by BART and refine the results using the targeting step. TMLE updates the initially predicted outcome $\overline{Q_0}$ to $\hat{Q}^*$ by fitting the fluctuation parameter ε in the model $\operatorname{logit}\overline{Q_1}=\operatorname{logit}\overline{Q_0}+\varepsilon H(A,W)$, where the clever covariate $H(A,W)$ leverages information from propensity scores produced in the BART model. By integrating TMLE's targeting procedure, BART model provides a flexible machine learning estimate that accommodate complex data patterns. The targeting procedure of TMLE correct the remaining bias from the BART model by optimally utilizing information from both the outcome and propensity score.

The Clever-covariate allows observations to receive higher weights in the data when they have extreme propensity scores (students who are either very unlikely to receive special education services but do receive the service; ($A=1$, $g(W)\approx 0$), or very likely to receive them but do not receive the service ($A=0$, $g(W)\approx 1$)). The property of double robustness is necessary since the clever covariate corrects for misspecification in either the outcome model or the propensity score model. Double robustness allows the clever covariate to reduce bias, even when the outcome model is not correctly specified. Similarly, the fluctuation step has minimal impact on whether the propensity score is correctly estimated, if the outcome model is correctly specified [42]. By focusing on the observations with extreme propensity scores the clever covariate helps to refine estimates $\hat{Q}^*$ in the fluctuation step, such that the efficient influence function condition is satisfied; $\mathbb{P}_n D^*\left(O;\hat{Q}^*;g\right)=0$, where $D^*\left(O;\hat{Q}^*;g\right)$ represents the canonical gradient for the target parameter [43]. This procedure allows that treatment effect estimates to achieve the semiparametric efficiency bound if one of the outcome or propensity score models is correctly specified. This provides a layer of protection against misspecified model bias. Furthermore, under certain conditions, the variance of the TMLE estimator captures uncertainty from both the outcome and propensity score models. Under certain conditions, TMLE reaches the lowest possible long-run variance allowed by theory, which is a trait not achieved by any other model [1].

Although BART-TMLE framework flexibly controls for observed confounders,

it is difficult to verify the assumption of no unmeasured confounding, which is important for valid inference. Many aspects of the students' characteristics can be challenging to quantify, which could be actual confounders in the context of special education. These factors that correlate with the treatment and outcome could bring unintended bias to the estimated treatment effect, under the current understanding of instruction methodology that can be observed. Future educational studies that provide additional routes to measure learning quality or individualized classification for academic difficulty could calibrate the estimation of special education efficacy.

The consistent estimate of the average treatment effect without correctly specifying both models makes doubly robust estimation a good characteristic of TMLE [44]. As BART does not assume a fixed form, it is likely that the model accurately captures at least one part of the data generation process. Additionally, TMLE mostly concentrates on reducing bias in treatment effect estimation, rather than on prediction accuracy. As the project's primary interest is the causal effect of special education, TMLE's focus on treatment effects makes it very compatible with the main objective.

The double LASSO variable selection methodology was selected based on its advantages for addressing the fundamental identification problem in causal inference under the no unmeasured confounding assumption. It allows the model to balance the tradeoff between high dimensional data overfitting and reducing the bias of insufficient variables [31] [34]. This dual optimization is closely related to semiparametric efficiency theory. The efficient influence function for the average treatment effect depends on both the propensity score $\pi(x) = P(T = 1 \mid X)$ and the outcome regression functions $\mu_0(X) = E[Y(0) \mid X]$ and $\mu_1(X) = E[Y(1) \mid X]$ [45] [46]. One major disadvantage of the univariate selection method is that it only optimizes a single loss function. Dual LASSO select covariate set S satisfies the approximate sparsity conditions necessary for $\sqrt{n}$ consistent estimation of treatment effects in high-dimensional settings: $\left\|\theta_0 - \hat{\theta}\right\|_2 = O_p\left(\sqrt{\frac{s_0 \log(p)}{n}}\right)$ where s_0 represents the effective sparsity parameter [47]. The method's theoretical advantage over alternatives such as Boruta stems from its explicit targeting of the nuisance parameters essential for unbiased causal estimation, rather than general predictive accuracy which may include noise variables that inflate variance without improving identification [48]. Additionally, using a sample-splitting strategy helps to the results valid by satisfying the Neyman orthogonality condition, eliminating the bias that typically results from using the same data for both variable selection and estimation [31]. The regularization parameter selection via cross-validation provides an adaptive procedure that achieves the optimal convergence rate $\min\left\{\sqrt{\frac{\log(p)}{n}}, \sqrt{\frac{s_0 \log(p)}{n}}\right\}$ for the treatment effect estimator, where p represents the ambient dimension and s_0 de-

notes the true sparsity level [49].

A BART model is fitted in conjunction with TMLE, with the 8 variables selected from the dual-LASSO selection process; The outcome model is specified as $Y_i(A)=\mu_A(X_i)+\varepsilon_i$, $A\in\{0,1\}$. $Y_i(1)$ and $Y_i(0)$ denotes the combined test score under treatment and control for student i, respectively, for student $i=1,\cdots,n$. The conditional mean function for the combined test score under treatment level A can be expressed as $\mu_A(X_i)=\sum_{j=1}^{m} g\left(X_i;T_j^A;M_j^A\right)$, $\varepsilon_i \mid X_i,A_i \sim N\left(0,\sigma^2\right)$, where X_i represents the vector of confounders, $g\left(X_i;T_j^A;M_j^A\right)$ is the regression tree function, where T_j^A represents the tree structure that determines splitting rules, $M_j^A=\left\{\mu_{1j}^A,\cdots,\mu_{bj}^A\right\}$ denotes the terminal node parameters for tree j. In this study, 200 trees are used to fit the BART model. Tree structure prior for each T_j^A is specified as $P\left(depth\left(T_j^A\right)=d\right)=\alpha(1+d)^{-\beta}$, $\alpha=0.95$, $\beta=2$. β controls the tree depth penalization. $\beta=2$ is used to moderately control complexity of the model to prevent overfitting, with a good trade-off between flexibility and regularization. Terminal node parameter prior is specified as $\mu_{ij}^A \mid T_j^A \sim N\left(\mu_s,\sigma_\mu^2\right)$, where $\mu_s=\frac{\bar{y}}{m}$ and $\sigma_\mu=\frac{y_{\max}-y_{\min}}{2k\sqrt{m}}$ with $k=2$. k controls the prior variance of the terminal node, similar to the choice of β, $k=2$ is selected to ensure moderate shrinkage and regularization toward the prior mean. Residual variance prior is specified as $\sigma^2 \sim \text{InverseGamma}\left(\frac{v}{2},\frac{v\lambda}{2}\right)$, $v=3$. This set the prior weakly informative, allowing the data dominate the posterior and provide mild regularization to prevent the variance towards zero.

Similarly, Treatment assignment model is also specified, represented by the following expression: $A_i \mid X_i \sim \text{Bernoulli}\left(\pi\left(X_i\right)\right)$, $\pi\left(X_i\right)=\Phi\left(h\left(X_i\right)\right)$, $h\left(X_i\right)=\sum_{j=1}^{m} g\left(X_i;T_j^\pi;M_j^\pi\right)$. $\Phi\left(h\left(X_i\right)\right)$ is the standard normal cumulative distribution, where $h\left(X_i\right)$ is the latent propensity score, and $g\left(X_i;T_j^\pi;M_j^\pi\right)$ represents tree function structures similar to that of the outcome model. The prior of the treatment model is specified as $P\left(depth\left(T_j^\pi\right)=d\right)=\alpha(1+d)^{-\beta}$, $\alpha=0.95$, $\beta=2$, with terminal node prior $\mu_{ij}^\pi \mid T_j^\pi \sim N\left(\mu_s,\sigma_{\mu,\pi}^2\right)$. The average treatment effect of the fitted model is estimated as $\tau=E\left[Y_i(1)-Y_i(0)\right]=E\left[\mu_1(X_i)-\mu_{01}(X_i)\right]=E\left[\left\{g\left(X_i;T_j^1;M_j^1\right)-g\left(X_i;T_j^0;M_j^0\right)\right\}\right]$, with 1000 burn-in iterations and 4000 posterior samples, assuming unconfoundedness. The model is fitted assisted with *bartCause* package version 1.0-9 in R [10].

It is critical to check for the common support before conducting analysis. Common support is a critical assumption to make for causal inference in observational studies for valid results. Graphical diagnosis is typically used to check this assumption. The graph for propensity score density, also known as the positivity or overlap condition, among treatment and control group is a common choice for such

diagnosis. The positivity assumption states that for every set of observed covariates W, the probability of receiving either treatment is greater than zero and less than one: $0 < P(A=1|W) < 1$. The assumption of common support is satisfied if propensity score density plot for treatment and control overlaps well. The propensity score, $(A=1|W)$ is a scalar summary of the covariate vector W. By plotting the density of these scores, a high-dimensional problem is reduced to a one-dimensional one, making the common support region readily apparent. When the common support is strong, an individual in control group can be matched to similar treated ones to estimate the outcome for the other group. If there's a significant overlap between the treated and control groups, the curve overlap indicates sufficient number of observations from both groups. Sufficient common support is an important condition to check, that for every treated individual, a credible effect can be estimated with or without treatment. If the density curves do not overlap well, this suggests that the common support assumption is not satisfied. In these cases, the model is not able to find a comparable treated and control individual since there's no observation with similar propensity score. Causal effects cannot be empirically estimated in the region with no common support, and any inference would rely on extrapolation [26]. Any inferences made without common support can result in biased estimation of the treatment effect with high uncertainty. In order to ensure the research results can be applied broadly and hold true under different conditions, it is important to show sufficient overlap between the treatment and control groups definitively, which helps to confirm the findings is reliable and relevant for analysis beyond the current sample [6] (**Figure 2**).

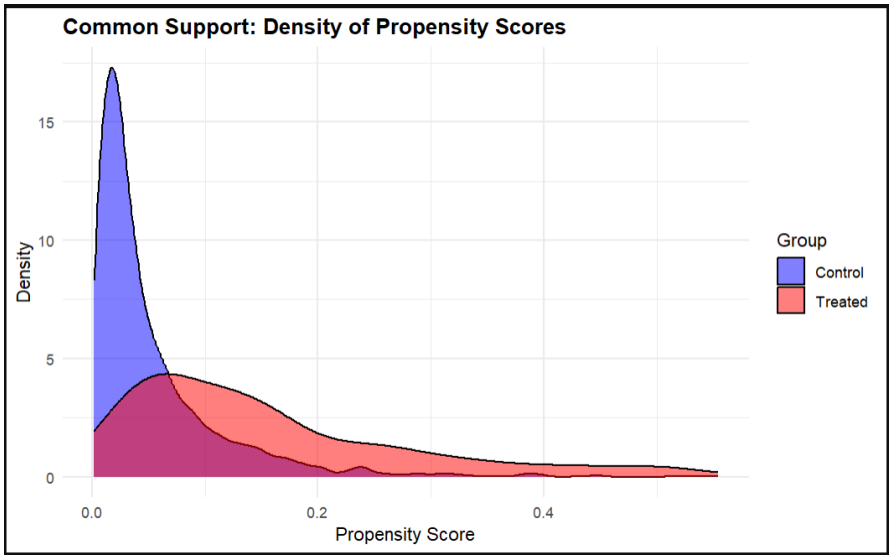

Figure 2. Density of propensity score for control and treated groups.

The propensity density scores show the propensity score distribution for control (blue) and treated (red) groups. The significant overlap between these two density curves in the middle of the distribution indicates a robust common sup-

port region. The plots show some potential issues with propensity scores that is more extreme, where the two groups do not overlap well. The control group has a dense peak at very low propensity scores (near 0), while the treated group shows a long tail extending to higher propensity scores (up to 0.5) with minimal density overlap. When the area of common support is not significant, the model has to rely on extrapolation. If a student's score cannot be matched by a comparable observation, the estimation of the effect only accounts for a small proportion of the sample. This can make the results biased and unreliable (Figure 3).

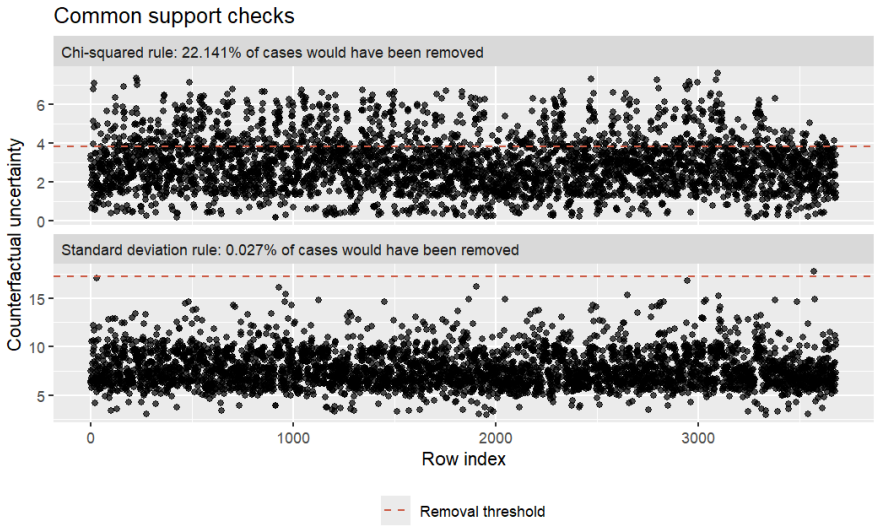

Figure 3. Common support diagnosis under different observation removal rule.

The current model shows a high posterior uncertainty if no observation is excluded from the fit. Based on a common support diagnosis, 22.14% of cases would have been removed based on chi-squared rule, whereas only 0.027% of cases would have been removed under the standard deviation rule. Therefore, using the Standard deviation rule allows the sample to be retained, excluding fewer observations to avoid extrapolation and meet the common support assumption. The rule identifies observations for removal if their predicted counterfactual standard deviation $s_{i,f(1-z)}$ is excessively large compared to the observed predictions' standard deviations, which can be expressed as: $s_{i,f(1-z)} > m_z + a \times sd\left(s_{j,f(z)}\right)$. To test the sensitivity of the model and check the potential issue of standard deviation rule being too lenient, a chi-square common support criterion is also used. The chi-square rule uses a more conservative exclusion compared to the standard deviation rule. Under this alternative approach, observations get excluded from the inferential sample if $\left(\frac{s_{i,f(1-z)}}{s_{i,f(z)}}\right)^2 > q_\alpha$, where $s_{i,f(1-z)}$ and $s_{i,f(z)}$ represent the predicted counterfactual and observed standard deviations respectively, and, and q_α denotes the upper α percentile of a χ^2 distribution with one degree of freedom [16]. With $\alpha = 0.05$, this criterion tests the null hypothesis of equal var-

iance between observed and counterfactual predictions, offering a stricter assessment of extrapolation risk than the standard deviation rule. The chi-square method removed 22.14% of observations, keeping 2866 observations out of 3,681. The observation exclusion rate hikes from 0.027% with the standard deviation rule to 22.14% using the chi-square criterion. Choosing different exclusion strategies shows the key tradeoff between higher statistical power and reliable treatment effect by reducing extrapolation. Comparing causal estimates from both approaches helps to show if the model relies too much on certain observations. This will help to identify if the common support assumption is met by the fitted model, and still reduce potential bias from extrapolation.

3.3. Results for BART-TMLE Model

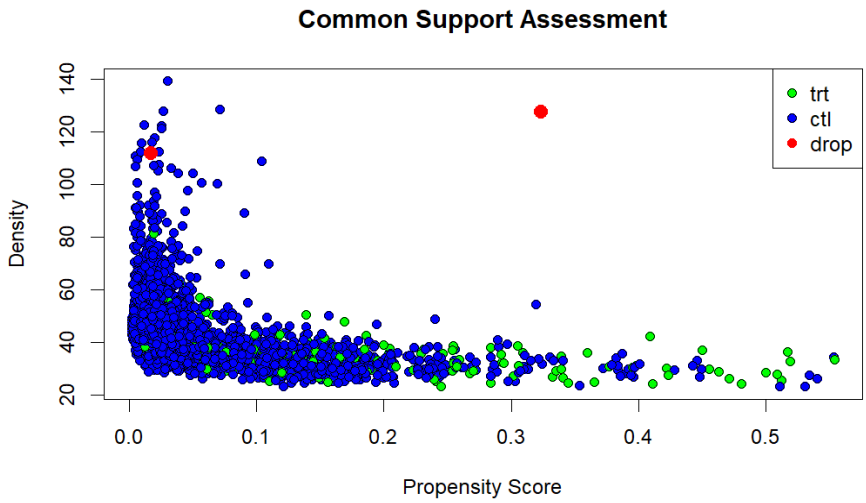

Figure 4. Propensity score density after standard deviation removal rule is adapted.

Observation removal based on standard deviation rule is used to improve the common support for posterior samples. Before conducting formal analysis, a diagnosis is completed in order to ensure for any individual in the treatment group, a comparable individual with similar characteristics exists in the control group. This will allow the model to give relatively robust estimates, as observations with low common support are not used to extrapolate the interpretation of the treatment effect (Figure 4).

The graph for common support shows treatment and control groups overlap well across the propensity score distribution. This makes the results from the analysis reliable without excessive extrapolation. Treatment (green) and control (blue) observations throughout the propensity score range from 0.05 to 0.50 mix well in the trace plot. This suggests children with similar characteristics are represented in both treated and controlled groups. Most children in the controlled group appear at lower propensity scores. This is realistic, since these children did not receive special education services. The treated children are located in a wider range of propensity score from low to high, representing their diverse background of

these students. The graph only identified one excluded observation (red), using the standard deviation common support rule. This shows that almost all observations in the sample contributed to the analysis. The model is able to have a good statistical power while still using the maximum number of observations it can utilize for treatment effect estimation, allowing results to have minimal bias resulting from case exclusion (Figure 5).

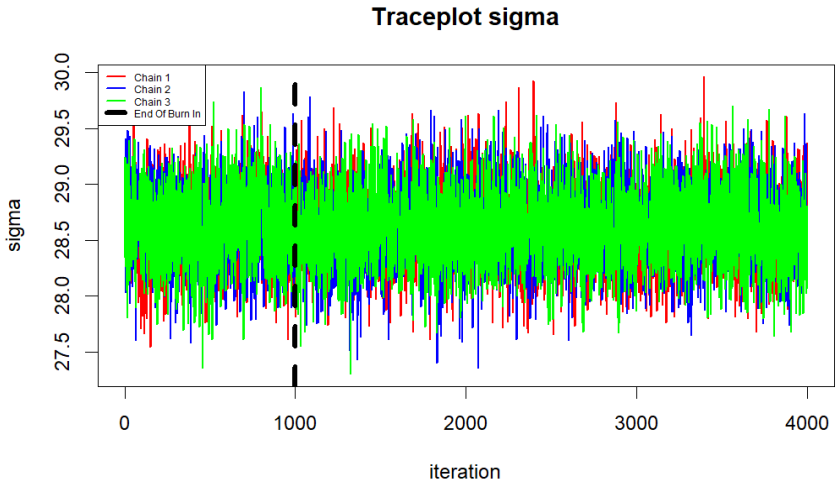

Figure 5. Convergence plot of fitted BART model (Standard deviation removal rule).

The trace plot for the error variance parameter (σ) shows MCMC convergence for three independent chains. All three chains reached a stationary distribution, with σ ranging from 27.5 to 29.5, and a mean around 28.5 after the 1000-iteration burn-in period. The extensive mixing of the post-burin in samples for the three chains (red, green, and blue lines) shows that the three independent chains have successfully explored the posterior distribution and the variance of the distribution without showing systematic difference. The stationary post-burn-in samples and absence of the chain-specific patterns, provide strong evidence that the MCMC algorithm has achieved convergence for this parameter. This will allow the BART model to draw reliable inference based on the stable samples after 5000 total iterations is completed (Figure 6).

The trace plot for average treatment effect gives additional evidence for the convergence of treatment effect for all three chains of the posterior samples. The treatment effect (ATE) of the posterior samples of the three chains (red, blue and green) is consistently well-mixed around −8 to −10 points for post-burn-in samples. This shows stationary negative treatment effect for the post-burn-in samples. The three distinct chains show similar distribution for the post-burin-in samples without significant discrepancies and major divergence. The spread of post-burn-in samples is fairly concentrated around the central value, showing the uncertainty of the post-burn-in samples is stable. The stable. The convergence of both error variance and estimated treatment effect shows a stable patten, which indicates that joint posterior distribution was successfully captured by MCMC sampling pro-

cess, making the estimands robust.

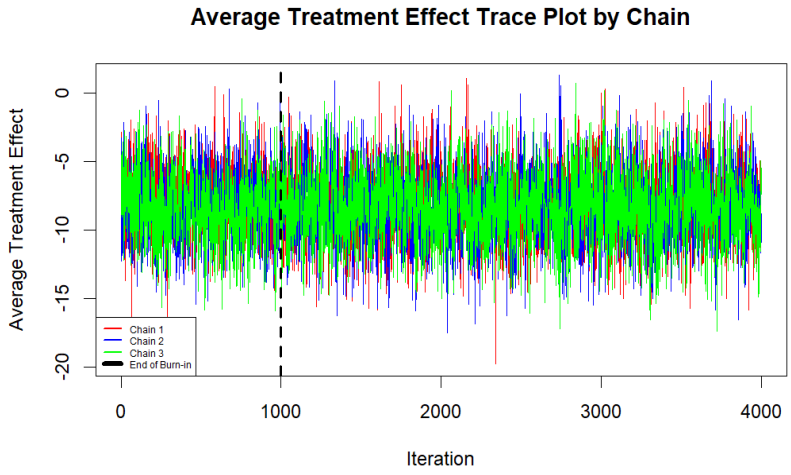

Figure 6. Average treatment effect estimates of 3 different chains of BART model (BART model based on standard deviation removal rule).

The analysis shows a negative effect of special education on combined reading and math scores. Children who did not receive special education scored about 9 points higher on average than those who did. The average score for children did not receive special education and children who received special education were 282.872 and 273.820, respectively. The estimated negative effect is consistent, for different subgroups and samples of the record population. The 95% credible intervals for these estimates [−14.330, −3.753] for the sample average treatment effect, do not include zero, indicating a high probability that the effect is real. The conditional average treatment effect also suggests that the treatment effect is different for distinct student groups.

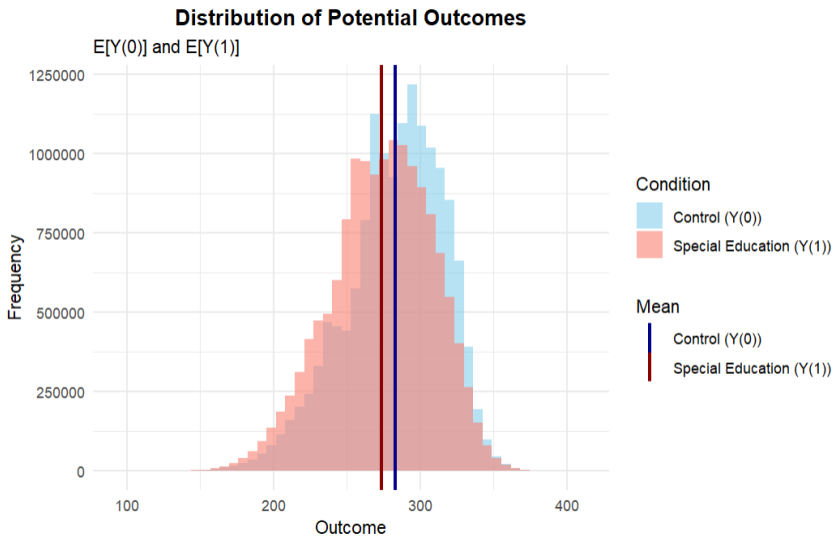

Figure 7. Academic performance difference between special education cohort and control: (BART model based on standard deviation removal rule).

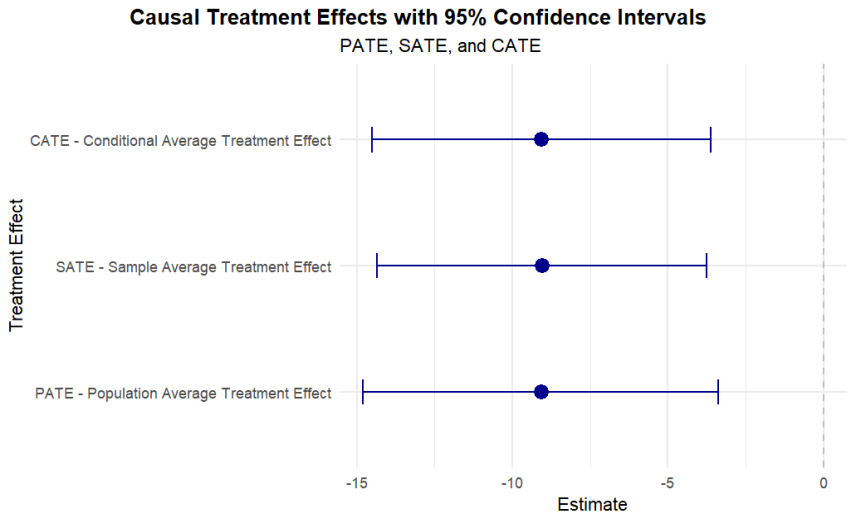

Figure 8. 95% confidence interval of treatment effect: (BART model based on standard deviation removal rule).

The estimated treatment effects for the population (PATE), sample (SATE), and subgroups (CATE) are all approximately −9.1. The 95% credible intervals for these estimates are entirely in the negative range, indicating a high posterior probability that the treatment effect is entirely below zero, which shows that the treatment effect is negative using different methods. This provides strong evidence that receiving special education is associated with a decrease in combined reading and math scores, both on average and across subgroups (**Figure 7**, **Figure 8**).

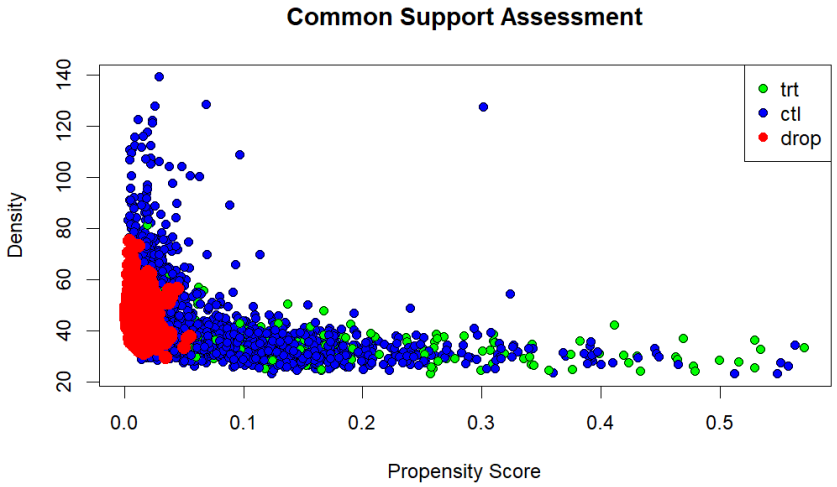

Figure 9. Propensity score density after chi-square removal rule is adapted.

In order to check the sensitivity of the model under different exclusion strategies, the chi-square rule is also used to fit the model under using fewer observations. Using the strict exclusion criteria leads to similar but different results, compared to the standard deviation exclusion method. The common support plot heighted a dense concentration of excluded observations (red) in the low propen-

sity score region (0.0 - 0.1). The chi-square criterion identifies numerous control group observations with high counterfactual uncertainty that were retained under the more lenient standard deviation rule (Figure 9).

Obviously, the chi-square rule trims the data in a significantly aggressive way. The cluster of excluded observations in the lower-left region of the common support plot suggests children with low probability of receiving special education are marked as having unreliable counterfactuals estimates under a conservative strategy. The chi-square method tends to use observations with higher propensity score, at the cost of losing more samples, with possible risk of extrapolation for estimated treatment effect.

An aggressive observation removal rule creates a more homogeneous sample with higher common support in the 0.1 - 0.4 propensity score range, at the cost of substantially reduced external validity. The remaining common support region shows the treatment and control group are better matched, but a considerable number of observations are lost in the process, which could be a source for bias if the treatment effect is estimated from this sample.

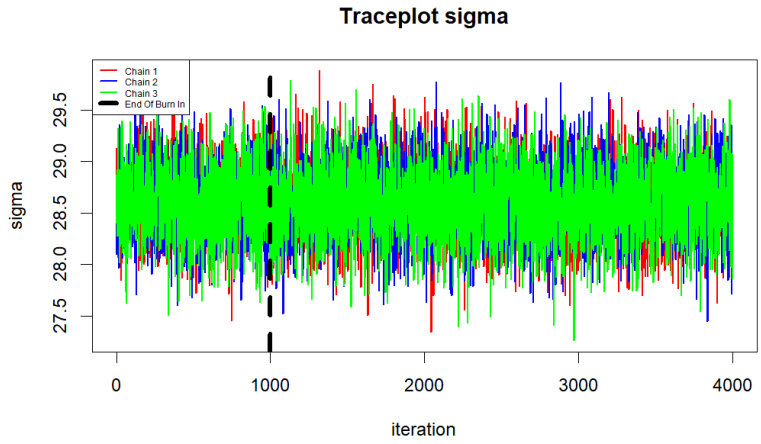

Figure 10. Convergence plot of fitted BART model (Chi-square removal rule).

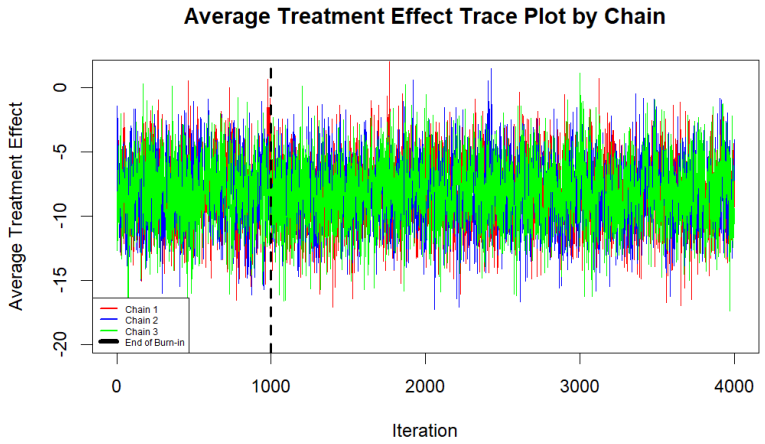

Figure 11. Average treatment effect estimates of 3 different chains of BART model (BART model based on chi-square removal rule).

Similarly, the trace plot of the error variance and treatment effect shows good convergence of the three chains for the post-burn-in samples, even though a significant number of samples are removed by the chi-square exclusion strategy. There is no significant shift of a specific chain compared to the others, showing no systematic difference between chains. The well-mixed chain shows an average error variance around 28.5 - 29.0. This shows the stationary convergence of the post-burn-in samples regardless of the exclusion strategy used (Figure 10, Figure 11).

The trace plot for average treatment effect trace plot shows consistent treatment effect around −10 to −12 points, compared to the estimated effect using the sample under standard effect sample, the estimated negative treatment effect using less samples is slightly greater (−8 to −10 points using standard deviation rule). The greater negative effect might be masked by the observations excluded by stricter chi-square rule. However, this is a hypothesis that need to be validated by statistically significant results.

The results from the BART model show a consistent negative treatment effect on the students' combined score across all estimands. Children receiving special education services are associated with approximately 8 - 9 points reductions in combined academic scores, compared to their peers. The estimated overall score suggests that students would score an average of 282.872 points without special education services compared to 273.820 points with special education services, yielding a raw difference of −9.052 points.

The three causal estimands, PATE, SATE and CATE give consistent point estimates for the treatment effect: the Population Average Treatment Effect (PATE) at −8.395, Sample Average Treatment Effect (SATE) at −8.453, and Conditional Average Treatment Effect (CATE) at −8.395. Although the estimands rely on different assumptions and target different and target different groups, the similarity of the point estimates suggest that the observed effect is mostly consistent regardless of the estimation method (Table 2).

Table 2. Comparison of treatment effect estimates under different exclusion criteria.

BART Causal Inference Results Comparison						
Standard Deviation vs. Chi-Square Common Support Rules						
Metric	Standard Deviation Rule			Chi-Square Rule		
	Estimate	95% CI Lower	95% CI Upper	Estimate	95% CI Lower	95% CI Upper
E[Y(0)]: Control Potential Outcome	282.872	208.874	336.233	282.875	208.773	336.205
E[Y(1)]: Treated Potential Outcome	273.82	200.596	334.498	273.758	200.649	334.035
Difference (Raw)	−9.052	–	–	−9.117	–	–
Mean Propensity Score	0.059	–	–	0.06	–	–

Continued

[1]PATE: Population Average Treatment Effect	−9.051	−14.607	−3.407	−8.395	−13.346	−3.43
[1]SATE: Sample Average Treatment Effect	**−9.041**	**−14.33**	**−3.753**	**−8.453**	**−12.953**	**−3.954**
[1]CATE: Conditional Average Treatment Effect	−9.051	−14.497	−3.606	−8.395	−13.11	−3.679

[1]The SATE and PATE involve calculating predicted response values under different treatment conditions. E[Y(Ai)] shows read and math scores for the most recent school year, 1 = special education.

The 95% credible interval provides additional information on the magnitude and uncertainty of the estimated effect. The SATE interval [−12.953, −3.954] indicates the true effect is very likely negative, considering the entire interval is less than zero. The PATE interval [−13.346, −3.430] is relatively wider, reflecting additional uncertainty when applied to the general population outside of the current sample. All three intervals of the estimands are entirely negative, providing strong Bayesian evidence against the null hypothesis of no treatment effect.

The mean propensity score of 0.060 confirms that children receive special education are a relatively small population in the sample, accounting for approximately 6% of students. The low prevalence rate of special education and substantial common support between the treatment and control groups support the reliability of the estimated results. Notably, the estimated treatment effect applies to the specific subset of students with characteristics that make them eligible for special education services.

4. Conclusions

The conclusion of the analysis is formed based on the sample produced after the standard deviation rule is applied to exclude observations lacking common support. (approximately 99.97% of the original data). In order to check the sensitivity of the BART model, the chi-square rule was also used, removing 22.14% of observations. The estimated treatment effect yielded consistent results, with slightly more negative point estimates in the same direction. The similarity of estimated results suggests robustness of the negative treatment effect using different common support distribution and exclusion methods. The observed negative effects may reflect the stigmatization of students receiving special education and reduced academic expectations. The concept of special education could be significantly effective, but a lack of individualization can diminish the effort of the educator. There are also several drawbacks and limitations to this study. Treatment assignment could be significantly influenced by potential confounding and measurement error. Research in the future should focus on differentiating heterogeneous treatment effects based on individualized characteristics, such as disability type

and family background. Investigating underlying reasons for the academic performance discrepancy will benefit the children receiving special education in the long term with nuanced customization of instruction.

Although implementing the model with BART and TMLE helps to address potential confounding factors that might introduce bias to the results on the overall academic score, the exclusion strategies used could also make the methodologies used in this study not applicable to other educational studies. Although the BART sample used almost all observations under the standard deviation rule, the conservative exclusion method still shows high uncertainty for out-of-sample estimation, as a fairly large portion of the sample is discarded. If similar studies use an exclusion strategy that eliminates a substantial number of observations, the remaining part might produce biased results with high uncertainty. The negative average treatment effect could be a reference for further investigation, but the researchers should consider the context of their study cohort and modify the model specification based on their needs.

5. Discussion

There are also several limitations to acknowledge based on the data source and structure. One of the potential issues of the study is using special education services as a single binary exposure variable. In reality, special education has a set of highly heterogeneous interventions that could be considerably different based on the duration and method of delivery. The resources allocated to each student could be a wide range of selections. For example, a student can simply be granted extended test time as a form of supported instruction or a fully customized instruction with one-on-one interactions. Different types of disabilities may also impact the student differently. For example, a disorder of intellectual development may stall a student's learning ability, while a student with hearing loss might not have reduced learning ability, but a limited pathway to learn. Collapsing the complex dimension of disability to a single binary factor of whether students receive special education could omit heterogeneity in treatment effect across different special education service types. The negative estimated average treatment effect from this study could represent a weighted average across different types of interventions that are substantially different. This aggregated result could obscure certain positive effects for some type of special education service type, with ambiguous boundaries compared to the other service type. Research in the future should emphasize distinguishing the special education service by disability category, intensity, instructional setting and other specific intervention types to better understand the efficacy of each unique category.

The math and reading scores only reflect the student's performance over a short period. The educational potential of the students can also be expressed by other performances in their life, such as their ability to socialize with their peers, aptitude or special talent in a specific knowledge domain or the ability to adapt to the environment. Educators might want to explore other aspects of children's lives to

fully determine the efficacy of the special education program. The negative treatment effect of the traditional test score might not reflect special education's potential for improving educational outcomes in other ways. Instead of abolishing the service, the policymaker might want to consider how to individualize the instruction, to make the program beneficial for traditional test-based education.

It is possible that there is potential unmeasured confounding that led to a negative treatment effect. Factors such as physical development delay, severity of learning disability are not recorded in the dataset, but could influence both special education placement and academic performance substantially. The estimated effect of the current model could reflect pre-existing conditions related to physical disadvantages of the children rather than the actual causal impact of special education services. For example, special education emphasizing basic skills could also limit the student's potential in grade-based classes, where adequate support is scarce. [50]. Children receiving special education also have limited access to learn with their peers who do not receive special education, giving them fewer opportunities to learn from classmates with higher academic performance on average [51]. If a more comprehensive dataset is available, a broader set of observed covariates could be considered, mitigating the issue of unmeasured confounding that was not adjusted in the current model, but has an impact on both treatment assignment and outcomes.

Future Research could dive deeper into several aspects. A full analysis of the CATE could reveal the generality of the estimated treatment effect for different schools and student subgroups, or only students with specific characteristics (e.g. students with a certain type of learning disability or family background). If the treatment effect differs for distinct subgroups, checking the discrepancies based on student characteristics could reveal the key covariates that drive heterogeneity, allowing investigation of potential effect modifiers, making the estimated effect of the special education service more accurate.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Blackorby, J., Wagner, M., Cameto, R., Davies, E., Levine, P., Newman, L., Marder, C., Cardoso, D. and Garza, N. (2005) Engagement, Academics, Social Adjustment, and Independence: The Achievements of Elementary and Middle School Students with Disabilities. National Center for Special Education Research. https://www.seels.net/designdocs/engagement/All_SEELS_outcomes_10-04-05.pdf

[2] Wagner, M., Newman, L., Cameto, R., Garza, N. and Levine, P. (2005) After High School: A First Look at the Postschool Experiences of Youth with Disabilities. National Center for Special Education Research. https://files.eric.ed.gov/fulltext/ED494935.pdf

[3] Hanushek, E.A., Kain, J.F. and Rivkin, S.G. (2002) Inferring Program Effects for Special Populations: Does Special Education Raise Achievement for Students with Disabilities? *Review of Economics and Statistics*, **84**, 584-599.

https://doi.org/10.1162/003465302760556431

[4] Shifrer, D., Muller, C. and Callahan, R. (2011) Disproportionality and Learning Disabilities: Parsing Apart Race, Socioeconomic Status, and Language. *Journal of Learning Disabilities*, **44**, 246-257. https://doi.org/10.1177/0022219410374236

[5] Feng, Y. (2024) Assessing the Effectiveness of Special Education Services on Fifth Grade Math Scores: Using Traditional and Machine Learning Methods with ECLS-K Data. *Applied and Computational Engineering*, **45**, 7-16. https://doi.org/10.54254/2755-2721/45/20241019

[6] Setoguchi, S., Schneeweiss, S., Brookhart, M.A., Glynn, R.J. and Cook, E.F. (2008) Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study. *Pharmacoepidemiology and Drug Safety*, **17**, 546-555. https://doi.org/10.1002/pds.1555

[7] 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. https://doi.org/10.1093/biomet/70.1.41

[8] Stuart, E.A. (2010) Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, **25**, 1-21. https://doi.org/10.1214/09-sts313

[9] Austin, P.C. (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, **46**, 399-424. https://doi.org/10.1080/00273171.2011.568786

[10] Hill, J.L. (2011) Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, **20**, 217-240. https://doi.org/10.1198/jcgs.2010.08162

[11] Tourangeau, K., Nord, C., Lê, T., Pollack, J.M. and Atkins-Burnett, S. (2006) Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Com-Bined User's Manual for the ECLS-K Fifth-Grade Data Files and Electronic Code-Books (NCES Publication No. 2006-032). U.S. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubs2006/2006032_2.pdf

[12] Imai, K. and Ratkovic, M. (2014) Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **76**, 243-263. https://doi.org/10.1111/rssb.12027

[13] Morgan, P.L., Frisco, M.L., Farkas, G. and Hibel, J. (2010) A Propensity Score Matching Analysis of the Effects of Special Education Services. *The Journal of Special Education*, **43**, 236-254. https://doi.org/10.1177/0022466908323007

[14] Imberman, S.A. (2014) The Effect of Special Education Services on Student Achievement. *Journal of Public Economics*, **118**, 98-110. https://doi.org/10.2139/ssrn.1031693

[15] Sullivan, A.L. and Field, S. (2013) Does Special Education Improve Preschoolers' Academic Skills? CASTL Research Brief. Curry School of Education, University of Virginia. https://files.eric.ed.gov/fulltext/ED544029.pdf

[16] Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4**, 266-298. https://doi.org/10.1214/09-aoas285

[17] Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G. and Najarian, M. (2009) Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Com-Bined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks (NCES Publication No. 2009-004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/ecls/data/eclsk_k8_manual_part1.pdf

[18] Kapelner, A. and Bleich, J. (2016) BartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software*, **70**, 1-40. https://doi.org/10.18637/jss.v070.i04

[19] Dorie, V., Hill, J., Shalit, U., Scott, M. and Cervone, D. (2019) Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, **34**, 43-68. https://doi.org/10.1214/18-sts667

[20] Green, D.P. and Kern, H.L. (2012) Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly*, **76**, 491-511. https://doi.org/10.1093/poq/nfs036

[21] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[22] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[23] Chipman, H.A., George, E.I. and McCulloch, R.E. (1998) Bayesian CART Model Search. *Journal of the American Statistical Association*, **93**, 935-948. https://doi.org/10.1080/01621459.1998.10473750

[24] Tan, Y.V. and Roy, J. (2019) Bayesian Additive Regression Trees and the General BART Model. *Statistics in Medicine*, **38**, 5048-5069. https://doi.org/10.1002/sim.8347

[25] George, E.I. and McCulloch, R.E. (1993) Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881-889. https://doi.org/10.1080/01621459.1993.10476353

[26] Rubin, D.B. (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688-701. https://doi.org/10.1037/h0037350

[27] Hahn, P.R., Murray, J.S. and Carvalho, C.M. (2017) Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3048177

[28] Pearl, J. (2009) Causality. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

[29] Linero, A.R. (2018) Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, **113**, 626-636. https://doi.org/10.1080/01621459.2016.1264957

[30] Ročková, V. and van der Pas, S. (2020) Posterior Concentration for Bayesian Regression Trees and Forests. *The Annals of Statistics*, **48**, 2108-2131. https://doi.org/10.1214/19-aos1879

[31] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., *et al.* (2018) Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, **21**, C1-C68. https://doi.org/10.1111/ectj.12097

[32] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[33] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[34] Belloni, A., Chernozhukov, V. and Hansen, C. (2014) Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, **81**, 608-650. https://doi.org/10.1093/restud/rdt044

[35] Wasserman, L. and Roeder, K. (2009) High-Dimensional Variable Selection. *The Annals of Statistics*, **37**, 2178-2201. https://doi.org/10.1214/08-aos646

[36] Fithian, W., Sun, D. and Taylor, J. (2014) Optimal Inference after Model Selection. arXiv:1410.2597. https://doi.org/10.48550/arXiv.1410.2597

[37] Rinaldo, A., Wasserman, L. and G'Sell, M. (2019) Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference. *The Annals of Statistics*, **47**, 3438-3469. https://doi.org/10.1214/18-aos1784

[38] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid Post-Selection Inference. *The Annals of Statistics*, **41**, 802-837. https://doi.org/10.1214/12-aos1077

[39] Gruber, S. and van der Laan, M.J. (2012) tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software*, **51**, 1-35. https://doi.org/10.18637/jss.v051.i13

[40] Van der Laan, M.J. and Rose, S. (2011) Targeted Learning: Causal Inference for Observational and Experimental Data (pp. 89-135, 165-210). Springer. https://doi.org/10.1007/978-1-4419-9782-1

[41] Porter, K.E., Gruber, S., van der Laan, M.J. and Sekhon, J.S. (2011) The Relative Performance of Targeted Maximum Likelihood Estimators. *The International Journal of Biostatistics*, **7**, 1-34. https://doi.org/10.2202/1557-4679.1308

[42] Schuler, M.S. and Rose, S. (2017) Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American Journal of Epidemiology*, **185**, 65-73. https://doi.org/10.1093/aje/kww165

[43] Zheng, W. and van der Laan, M.J. (2011) Cross-Validated Targeted Minimum-Loss-Based Estimation. In: van der Laan, M.J. and Rose, S., Eds., *Targeted Learning*, Springer, 459-474. https://doi.org/10.1007/978-1-4419-9782-1_27

[44] Imbens, G.W. and Rubin, D.B. (2015) Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press. https://doi.org/10.1017/cbo9781139025751

[45] Hahn, J. (1998) On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, **66**, 315.-331 https://doi.org/10.2307/2998560

[46] Hirano, K., Imbens, G.W. and Ridder, G. (2003) Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, **71**, 1161-1189. https://doi.org/10.1111/1468-0262.00442

[47] Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C. and Kato, K. (2017) High-Dimensional Econometrics and Regularized GMM. arXiv:1806.01888. https://doi.org/10.48550/arXiv.1806.01888

[48] Athey, S. and Imbens, G.W. (2019) Machine Learning Methods That Economists Should Know about. *Annual Review of Economics*, **11**, 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

[49] Wainwright, M.J. (2019) High-Dimensional Statistics. Cambridge University Press. https://doi.org/10.1017/9781108627771

[50] Kurz, A., Elliott, S.N., Wehby, J.H. and Smithson, J.L. (2010) Alignment of the Intended, Planned, and Enacted Curriculum in General and Special Education and Its Relation to Student Achievement. *The Journal of Special Education*, **44**, 131-145. https://doi.org/10.1177/0022466909341196

[51] Burke, M.A. and Sass, T.R. (2013) Classroom Peer Effects and Student Achievement. *Journal of Labor Economics*, **31**, 51-82. https://doi.org/10.1086/666653