Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241019

Assessing the effectiveness of special education services on fifth grade math scores: Using traditional and machine learning methods with ECLS-K data

Yuxiang Feng

School of Science, China University of Mining and Technology (Beijing), Beijing, 100083, China

fyx20020921@163.com

Abstract. The Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) is a well-known research endeavor in the field of child development. In this research, some special education services are offered to those students who need supplementary support in some aspects. In this paper, our study aims to estimate the average treatment effect on students' fifth grade math scores and assesses the effectiveness of these special education services based on the ECLS-K dataset, through both machine learning methods and traditional methods. We introduce Donald Rubin's causal model and Propensity Score Analysis in the part of traditional methods, and Ordinary Least Squares (OLS), Targeted Maximum Likelihood Estimation (TMLE), Bayesian Additive Regression Trees (BART), Generalized Random Forests (GRF) and Double Machine Learning (DML) in the part of machine learning methods. Finally, we employ Propensity Score Matching, OLS and BART to estimate the ATE. All estimated ATEs are significantly different from zero. The estimated ATEs are found to be minus, suggesting that these special education services may have a negative effect on students' fifth grade math scores. Obviously, this conclusion is inconsistent with the original intent of these services, which aimed to have a positive impact.

Keywords: ECLS-K, Propensity Score Matching, Ordinary Least Squares, Bayesian Additive Regression Trees, Negative Effect.

1. Introduction

The Early Childhood Longitudinal Study, Kindergarten Class of 1998-99(ECLS-K) is a nationwide study that tracks a representative group of children from kindergarten through their later school years. The ECLS-K focuses on capturing children's educational journey starting from their initial enrollment in kindergarten [1]. In the kindergarten phase, the predominant primary disability among students was identified as a speech or language impairment, accounting for 2.3 percent of the student population [2]. As students progressed through each grade, the prevalence of specific learning disability as the primary impairment within the cohort showed a progressive increase, starting from 0.5 percent in kindergarten and reaching 6.5 percent in fifth grade [2]. Therefore, special education services are provided to students necessitating supplementary assistance in academic pursuits, social-emotional growth, and other associated domains [3].

Causal inference plays a crucial role in various domains, including healthcare, marketing, and political science, offering valuable insights into real-world scenarios. The estimation of causal treatment effects, which represents a fundamental challenge in the field of causal inference, has been a subject of extensive investigation in statistics for a considerable period. Traditional treatment effect estimation may perform excellent in some cases, but such methodologies may encounter challenges in effectively managing extensive and multidimensional heterogeneous datasets [4]. In recent years, significant advancements have emerged that involve the integration of supervised machine learning (ML) techniques into estimators designed for estimating causal parameters, such as the average treatment effect (ATE). These innovations have had a notable impact and have transformed the field within the past decade. A study of an evaluation of a health insurance scheme on health care utilization in Indonesia demonstrated that ML can make their estimation more principled [5]. This paper focuses on the special education services towards the disability mentioned above, using both traditional methods and ML methods to estimate the ATE of these services on students' fifth grade math scores, trying to assess the effectiveness of these services.

We begin with traditional methods in causal inference. In this paper, we mainly introduce Donald Rubin's causal model and Propensity Score Analysis. The former is one of the most classic frameworks in causal inference, and the latter is a commonly used method to estimate the causal impact of a treatment when random assignment is not possible. The propensity score can help to balance confounders between treated and control subsamples. We also describe one of the most commonly used methods in Propensity Score Analysis, the Propensity Score Matching (PSM). We use PSM to estimate the ATE.

Then, we turn to ML methods used for estimating the treatment effect. Because they are easier to implement, and in some cases [4], they could still be effective while traditional methods are not, ML methods are prevalent in causal inference. Here we will introduce Ordinary Least Squares (OLS), Targeted Maximum Likelihood Estimation (TMLE), Bayesian Additive Regression Trees (BART), Generalized Random Forests (GRF) and Double Machine Learning (DML). We finally choose OLS and BART to estimate the ATE.

For ML methods, we use original dataset to train models, and use the trained models to predict the value of outcome under another treatment condition. By doing this, we can get the outcome for each sample under two different treatment conditions. Then we can compute the ATE for the population. However, it is not feasible to simulate a different treatment condition using traditional methods, as has been done with ML methods. Therefore, we employ PSM, a type of matching method, to estimate the ATE. We can conclude through the results that these services make negative effect to students' fifth grade math scores, which is opposed to the original intention of these services. Besides, we also compare the methods used to estimate the ATE in this paper, BART is the best method, while PSM is the worst.

2. Literature Review

Causal questions are intrinsically associated with specific interventions or treatments. Causal effects refer to the contrasts observed between potential outcomes under different treatment conditions, considering identical subjects [6]. However, it is obvious that we can't put a subject in the treatment group or the control group at the same time, which means causal inference is inherently a challenge related to missing data, making it fundamentally intertwined with the issue of incomplete or unavailable information [6].

2.1. Traditional Method

2.1.1. Donald Rubin's Model. There are two main models in causal inference, Donald Rubin's model and Causal Inference Directed Acyclic Graphs proposed by Judea Pearl.

Donald Rubin's causal model was initially proposed by Pall W. Holland [7]. Rubin's causal model finds extensive application in various disciplines such as medicine, statistics, economics and public health. The approach devised by Rubin emphasizes the importance of precisely specifying potential outcomes for each participant and the formulation of mathematically adequate assumptions to estimate the causal effect [8].

2.1.2. Propensity Score Analysis. The propensity score represents the conditional probability of assignment to a particular treatment, given a set of observed covariates [9]. In a randomized study, the assignment of treatment to participants is random, ensuring that the treated and untreated groups are, on average, equally distributed across all pretreatment covariates, both observed and unobserved. However, in practical applications, it is often inevitable that individuals who receive the treatment may exhibit systematic differences compared to those who do not receive the treatment [10], this kind of circumstance is called observational study. In an observational study, there may be confounding factors that influence both the outcome and the treatment. This phenomenon is identified as confounding, and controlling for confounding is a crucial step in the modeling process. The aim of propensity score analysis is to derive propensity score estimates that successfully achieve covariate balance between the treated and control subgroups. Attaining complete consistency is challenging; hence, our objective is to estimate the propensity score in a manner that promotes resemblance between the distributions of covariates among the treated and control units within subsamples defined by similar values of the estimated propensity score [11]. Now, propensity score methods are frequently employed to estimate the causal effect of a treatment or intervention in situations where random assignment is impossible [10].

2.2. Machine Learning Method

2.2.1. Ordinary Least Squares. Ordinary least squares (OLS), which was proposed by Adrien-Marine Legendre in 1806, is the commonest and the most typical method used to solve problems related with regression in traditional statistics. OLS can combine with some algorithms in ML, like Least Squares Support Vector Machines (LSSVM), which was proposed by J.A.K Suykens and J. Vandewalle in 1999[14] combing OLS with Support Vector Machines (SVM). Sparse solution of LSSVM is a good algorithm when the data dimension is not pretty large and the requirement for data accuracy is not particularly high [14].

2.2.2. Targeted Maximum Likelihood Estimation. Targeted Maximum Likelihood Estimation (TMLE) is a widely recognized and extensively documented approach, extensively discussed in numerous books, scholarly articles, and instructional materials. TMLE is a robust and efficient estimator that operates in two stages, incorporating double robustness principles [12], and it has shown great value of estimating the size of effect in physics, medical studies, economic and so on [13]. ML can be integrated into the TMLE procedure to facilitate the estimation of semiparametric models [12].

2.2.3. Bayesian Additive Regression Trees. Hugh A. Chipman, Edward I. George and Robert E. McCulloch proposed the Bayesian Additive Regression Trees, a kind of Bayesian Trees method combines Bayesian theory with Additive Trees model [15]. The Bayesian Additive Regression Trees (BART) estimator procedure combines many single and shallow regression trees to construct a predictive model. Regression trees, an early form of ML, serve as the foundation for BART, which incorporates two key components: a regularization prior and a sum-of-trees model. BART follows a Bayesian framework, where estimation results in a posterior distribution [15].

BART has strong flexibility in nonlinear and interactive aspects of fitting data, moreover, the method based on Bayesian probability model has more advantages than pure algorithm, and the generation ability is stronger after multi-tree integration [15].

BART also performs well in practical application. For example, the single-tree model originated from 1980s has expanded to integration-trees model using a large group of trees, these models perform well in fitting nonlinear function regression relationship, especially BART. Bonato and his colleagues adopted BART in their recent research about survival prediction, they used BART in hierarchical covariate structure [15].

In causal inference, BART has shown remarkable performance. The estimation of treatment effects using Bayesian Additive Regression Trees (BART) demonstrates significantly improved accuracy in nonlinear settings compared to other widely used approaches such as linear regression and propensity score matching with regression adjustment. Even in situations where the response surface exhibits linearity and an additive treatment effect, simulations have shown that the performance of BART is nearly indistinguishable from that of linear regression. Thus, BART is a straightforward and promising method that exhibits robustness and accuracy in estimating causal effects [16].

2.2.4. Generalized Random Forests. Generalized Random Forests (GRF) was proposed by Susan Athey, Julie Tibshirani and Stefan Wager. It was built on the notion of random forests, a ML approach introduced by Breiman. The GRF approach is a nonparametric statistical estimation method that enables the fitting of various quantities of interest by solving a collection of local moment equations. Athey and colleagues further utilized the GRF framework to introduce novel techniques for three statistical tasks: nonparametric quantile regression, estimation of conditional average partial effects, and estimation of heterogeneous treatment effects using instrumental variables [17].

2.2.5. Double Machine Learning. The Double Machine Learning (DML) framework, initially introduced by Chernzhukov, leverages modern machine learning techniques to estimate parameters. This approach is robust to model misspecification and effectively reduces bias. Building upon this framework, Yonghan Jung and colleagues extended its application scope by proposing a new and general class of estimators called DML-ID. These estimators are designed for any identifiable causal functions that exhibit DML properties. The authors have concluded that DML-ID estimators possess key properties such as debiasedness and doubly robustness. Furthermore, simulation results provided empirical support for their theoretical findings [18].

2.3. ECLS-K

The Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) is a longitudinal investigation that focuses on the initial educational experiences of children starting from kindergarten and continues to track their progress through middle school. The ECLS-K represents a significant milestone as it is the first comprehensive nationwide study to examine early education with such extensive longitudinal coverage. Its primary objective is to gather reliable and comprehensive data to effectively describe and comprehend the development and experiences of children in elementary and middle school grades. Additionally, the study aims to investigate the linkages between children's early experiences and their subsequent development, learning, and school experiences. The data collected by the ECLS-K furnish valuable insights into children's status upon school entry, their transition into the educational system, and their academic trajectory up to the 8th grade [3].

The longitudinal design of the ECLS-K dataset facilitates investigations into the associations between various factors encompassing family, school, community, and individual domains and their impact on academic achievement. Therefore, there are many related research. For example, based on the ECLS-K, Kang Jeehye's study provide evidence in support of the assertion that corporal punishment has detrimental effects on children's social development [19].

3. Methods

3.1. Propensity Score Analysis

3.1.1. Propensity Score. In a randomized study, participants are randomly allocated to either the treatment group or the control group. This random assignment guarantees that the distribution of covariates is balanced between the two groups, allowing for a direct comparison of outcomes. By comparing the outcomes of participants in these two groups, we can estimate the treatment effect. In contrast, in an observational study under typical conditions, the subjects are not assigned randomly.

Therefore, confounding arises when there exists an association between one or more covariates and both the assignment of treatment and the outcome. As a result, systematic differences may arise between the treated individuals and the control individuals prior to the administration of the treatment. Moreover, such differences would bias the actual treatment effect.

We can use the average treatment effect for the treated (ATT) to estimate the ATE in a randomized study, but we can't follow this rule in an observational study. Because in a randomized trial, the potential outcomes $(Y(0), Y(1))$ and the treatment assignment Z are assumed to be independent, thus the ATE is identical to the ATT. This relationship can be formally expressed and estimated based on the available observed data:

$$ATT = E(Y(1) - Y(0)|Z = 1) = ATE = E\big(Y(1)\big) - E(Y(0)) \tag{1}$$

However, we don't have such a condition in an observational study. In such circumstances, it is important to note that the ATE and ATT differ, thereby preventing a direct comparison of outcomes for estimating the treatment effect.

To solve the problems caused by the presence of confounding, propensity score analysis was proposed, and it circumvents many limitations in practice while other methods fail to do. The propensity score (PS) was initially introduced by Rosenbaum and Rubin and refers to the conditional probability of treatment assignment given a set of observed baseline covariates [7]. Propensity score methods are grounded in the causal model conceptualized by Rubin, providing a theoretical foundation for their application.

Rosenbaum and Rubin introduced the concept of strong ignorability, which is characterized by the fulfillment of two distinct conditions in treatment assignment. We will discuss these assumptions one by one.

The first condition, which can be called as "no unmeasured confounders", stating that the potential outcomes $(Y(0), Y(1))$ and the treatment assignment Z are conditionally independent given the observed baseline variables X. Hence, it can be inferred that when this condition is satisfied, all the confounding variables that influence both the outcome and the treatment assignment have been accounted for and measured in the set of observed baseline variables X. The second condition, which can be called as "probabilistic assignment", stating that there exists a positive probability for a subject to be assigned to either the treated group or the control group. And the description of this condition can be expressed as the following formula:

$$0 < Pr(Z = 1|X = 1) < 1 \quad [20] \tag{2}$$

In a randomized study, the situation would be simple, therefore we mainly focus on observational study. In situations where the treatment assignment in an observational study is presumed to exhibit strong ignorability, Rosenbaum and Rubin demonstrated that unbiased estimates of ATE can be acquired by conditioning on the estimated propensity score, denoted as $e(x)$, which represents the conditional probability of treatment assignment given the set of confounding variables X:

$$e(x) = P\,r(Z = 1|X = x)\ [20] \tag{3}$$

The propensity score serves as a balancing score, ensuring that the distributions of the variables X are equivalent between the treated and control groups at each value of the propensity score. This implies that the treatment assignment Z and the observed variables X are conditionally independent, given the propensity score.

Weighting, stratification and matching are three methods employed frequently in propensity score analysis in order to replicate the characteristics of a randomized trial with respect to the variables X [20]. We can generate an output dataset comprising a sample that has undergone adjustment using these three methods. Within this dataset, the distributions of the variables are equivalent between the treated and control groups. Therefore, the observed variables in both groups exhibit random differences, similar to a randomized study. Subsequently, this output dataset can be utilized to estimate the treatment effect in an outcome analysis.

3.1.2. Propensity Score Matching. The PSM estimator imputes the missing potential outcome by utilizing the observed outcome of the nearest observations from the alternative group and computes the ATE as the simple difference in means between these predicted potential outcomes [21].

After implementing the PSM, a list of matched pairs including the control group and treated group can be obtained, with the exact number of pairs depending on the match parameter set in advance. Subsequently, the ATE can be estimated based on the values in the match list.

3.2. OLS

OLS is a typical approach, and the following formula is its normal expression:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{4}$$

Y is called dependent variable, X_i is called independent variable and ε is called residual mathematically. However, in ML, to connect them with computer, we can call Y as output and X_i as input.

When we turn to causal inference, we need to change the form of this expression because of the existence of treated group and control group, the updated expression is:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \theta Z + \varepsilon[12] \tag{5}$$

where Z is called exposure variable or treatment variable, and it's a binary variable, which means its value is 0 or 1.

Therefore, if we want to estimate the treatment effect of an experiment, we just need to compute the Y for different values of Z, here we use $Y(0)$ and $Y(1)$ to denote the value of Y when $Z = 0$ and $Z = 1$. Then the treatment effect can be expressed as $Y(1) - Y(0)$, moreover, the ATE can be expressed as $E[Y(1) - Y(0)]$.

3.3. BART

BART is a much more complex model than OLS. Therefore, here we consider using BART to estimate the ATE.

3.3.1. Decision Trees

3.3.1.1.Brief Description. It is a common situation that we want to use several variables to classify objects into some classes, we can use a tree-structure to solve this kind of problems this time. One of the most important factors in Decision Tress is nodes, a tree is just a collection of nodes, where any two nodes are connected with at most one line or edge. A normal kind Decision Trees is called binary trees, each node can have at most two children nodes. The nodes without child are known as leaf nodes or terminal nodes, others are called internal nodes. Each internal node has a decision rule associated with it. When we face a practical problem, we can follow those decision rules and we will reach a terminal node, then our decision problem will be solved according to the information from this terminal node.

3.3.1.2.Mathematic Definition. Above is a kind of figurative definition, and we will give a mathematical definition below.

For the starters, we will introduce parameter T. Here T represents a binary tree comprising a collection of interior node decision rules and a collection of terminal nodes. Then we will turn to another parameter denoted as M.There are many values inside this parameter, and we denote them as $\mu_i (i = 1,2, \ldots, b)$, which means we can express M as $M = \{\mu_1, \mu_2, \ldots, \mu_b\}$[22].$M$ denotes a collection of parameter values that correspond to each of the b terminal nodes of T.

Mathematically we can say that a g decision tree is defined by these two sets, and $g(x; T, M)$ represents the function that assigns the value $\mu_i \in M$ to x. The prediction for a specific input vector x is performed in the following manner: If x is linked to terminal node i of T through the sequence of decision rules from top to bottom, it is then assigned the corresponding μ_i value associated with that specific terminal node.

3.3.1.3.Regression Trees. We have known that Decision Trees can be used for a classification problem, where $\mathrm{M_j}$ contains classes or label-values. Besides, we can also use this method for regression. In such cases, we associate a terminal node with a real number like the mean of the data points. Of course, there are alternatives to the mean of the data points, like the median of the data points. In addition to returning exact values, we can also fit a linear regression to the data points, even more complex functions.

3.3.2. Ensembles of Decision Trees. This ensemble strategy is followed both by Bayesian methods like BARTs and non-Bayesian methods like random forests. When we try to build and train a model, it's important to limit its complexity. Over-complex trees will likely not be very good at predicting new data, it's common to introduce devices to reduce the complexity of decision trees and get a fit that better adapts to the complexity of the data. One solution relies on fitting an ensemble of Decision Trees, where each individual tree is regularized to be shallow. Then, each individual tree only explains a small portion of the data.

3.3.3. BART Model. BART Model is designed to estimate a general model for the outcome Y, given by $Y = f(z, x) + \varepsilon$, where z represents the treatment assignment,x represents the observed confounding covariates, and $\varepsilon \sim N(0, \sigma^2)$[22]. BART consists of two components: a regularization prior and a sum-of-trees model. We will discuss them below.

3.3.3.1.Sum-of-Trees model. For causal inference, we need to modify the Decision Trees model mentioned above slightly, adding a parameter denoted as z, which is as same as z in $\mathrm{f(z, x)}$.To model or approximate $\mathrm{f(z, x) = E(Y|x)}$, the mean of Y given by x, we consider about building a sum of b regression trees. Then we can express the structure of BART as:

$$f(z, x) = g(z, x; T_1, M_1) + \cdots + g(z, x; T_b, M_b) \ = \ \sum_{i=1}^{b} g(z, x; T_i, M_i)[16] \tag{6}$$

Therefore, our BART model can be formulated as follows:

$$Y = \sum_{i=1}^{b} g(z, x; T_i, M_i) + \varepsilon \tag{7}$$

In contrast to the model with a single tree, the terminal node parameterμ_i, as determined by $g(z, x; T_j, M_j)$ represents only a portion of the conditional mean of Y given x when $b > 1$. These terminal parameters in the BART model can capture interaction effects when their assignment is dependent on multiple components of x, for example, there are more than one variable. Given that the structure of BART allows for trees of different sizes, the Sum-of-Trees model has the ability to incorporate direct effects as well as interaction effects of various orders. When each assignment of the terminal nodes depends solely on a single component of x, the Sum-of-Trees model simplifies to a basic additive function. Except this special case, the Sum-of-Trees model offers greater flexibility compared to conventional additive models that employ low-dimensional smoothers as components.

3.3.3.2.A Regularization Prior. Given the inherent challenges of identification and the flexible nature of the Sum-of-Trees model, the prior distribution plays a crucial role in the estimation process. It serves two important purposes: it helps to regularize the overall fit of the model, preventing overfitting, and it constrains the influence of each terminal node assignment $(\mathrm{T_i}, \mathrm{M_i})$. We can greatly simplify the complexation of prior specification by letting the $\mathrm{T_i}$ be independent and identically distributed (i.i.d), the $\mu_{\mathrm{i,j}}$(node j of tree i) be i.i.d given the set of T, and σ be independent of all T and μ.

For the tree prior, the probability of a node being nonterminal is defined as $\alpha(1 + d)^{-\beta}$, where d represents the depth of the node, $\alpha \in (0,1), \beta \in [0, +\infty]$. With this prior, we can control the depth of each node and their difference by tuning parameters α, β. However, this doesn't mean we can tune parameters randomly, we also need to correspond to the goal of BART that each T_i is a “weak learner”. In the original paper of BART, these two parameters are selected as default, their values are $\alpha = 0.95, \beta = 2$[23].

Regarding the prior on μ,we begin by adjusting and scaling the values of Y in such a way that we assign a high prior probability to $E(Y|x)$ falling within the range of (-0.5, 0.5). Subsequently, we assume a normal distribution for μ with a mean of 0 and variance of σ_μ^2. For a given T_i and a x, $E(Y|x)$ is the sum of b independent $\mu's$. The standard deviation of the sum is $\sqrt{b}\sigma_\mu$. We select the value of σ_μ such that zero falls within k standard deviations, ensuring that 0.5 also lies within this range, which means $k\sqrt{b}\sigma_\mu = 0.5$[23]. In practical problem, if we get such a k (we denote its value as k_0 for the description below), we can set k_0 as the default choice and in practical applications, it is a common practice to transform the response variable by rescaling its observed values to fall within the range of -0.5 to 0.5. Furthermore, it should be noted that as the number of trees (b) increases, the prior distribution leads to a greater shrinkage of $\mu_{i,j}$ towards zero [16].

For the prior on σ, we have mentioned above that $\varepsilon \sim N(0, \sigma^2)$, then we can choose its conjugate prior, Inverse Gama distribution, which means $\sigma \sim InvGamma(\frac{v}{2}, \frac{v\lambda}{2})$. Normally,$v$ ranges from 3 to 10.Given v, we then need to choose λ to achieve the following formula, $P(\sigma^2 < \hat{\sigma}^2) = q$.[23] Simple data-driven options of $\hat{\sigma}$ used in practice are the estimate obtained from a linear regression or the sample standard deviation of Y, and normally used values of q are 0.75,0.90,0.99[23].

3.3.4. Estimating Causal Effects. BART can be employed for the estimation of average causal effects. We mainly focus on the conditional average treatment effect (CATE) and the conditional average treatment effect for the treated (CATT). Their representations are listed below:

$$CATE = \frac{1}{n}\sum_{i=1}^{n} E(Y_i(1)|X_i) - E(Y_i(0)|X_i) = \frac{1}{n}\sum_{i=1}^{n} f(1, x_i) - f(0, x_i) \tag{8}$$

$$CATT = \frac{1}{k}\sum_{i:Z_i=1}^{n} E(Y_i(1)|X_i) - E(Y_i(0)|X_i) = \frac{1}{k}\sum_{i:Z_i=1}^{n} f(1, x_i) - f(0, x_i)[16] \tag{9}$$

We have known that the treatment effect at $X = x$ is $f(1, x) - f(0, x)$, here we define a new function as $c(x, f)$ to make only x and f contribute to the representation of the treatment effect. The joint posterior distribution of $C(f) = \big(c(x_1, f), c(x_2, f), \dots, c(x_K, f)\big)$[16]. The posterior is obtained through Markov chain Monte Carlo (MCMC). In each iteration of the BART Markov chain, a new sample of f is generated from the posterior distribution, so we denote the lth draw of f as f^l, and its joint posterior distribution of $C(f)$ denoted as $C^l = C(f^l)$.

Now we can turn to estimate the CATE and CATT. Consider a set of K observations, denoted as $\{x_i\}_1^K$, representing the empirical distribution of x from which we aim to estimate the average treatment effect (ATE). If we want to estimate for the CATE, we just need to calculate the mean of the vector C^l at each l, $\overline{C^l} = \frac{1}{K}\sum_i^K c(x_i, f^l)$[16]. For the CATT, our attention would be solely directed towards $\{i: z_i = 1\}$.

4. Result

For the starters, we need to describe the dataset and its variables briefly. The dataset we use to estimate the ATE has 7362 samples. The outcome variable Y of this dataset means fifth grade math score, and the exposure variable Z of this dataset represents special education services. As for the covariates X, it includes 34 variables in all, and it is divided into five groups [10].

In practice, we just need to replace Z with 1-Z, and substitute this new group of independent variables into the model trained by the original data. Then we can get a new group of values for Y. By doing this, with the original data, we could obtain the outcomes of one sample in two different situations, the subject is treated or not. With these results, we can compute the treatment effect for each sample as$Y_i(1) -$

$Y_i(0)$,where i denotes the ith sample. Furthermore, we could also compute the ATE for this population as $E[Y(1) - Y(0)]$.

Here we choose PSM and ML methods, OLS and the BART model to estimate the ATE. The estimated results and their standard error are listed below.

In fact, prior to analyzing the estimated values of ATE and their standard deviation, it is crucial to assess the necessity of the special education services, regardless of whether the ATE is negative or positive. Hence, besides the ATEs and their standard deviation, we also provide p-value of t-test for PSM and confidence intervals for OLS and BART. For BART, the estimated ATE is based on Bayesian approach, rather than the Hypothesis Testing Method of the Frequency School. For OLS, although conducting a significance test is reasonable, a confidence interval can provide an estimate of the true value range, which is more intuitive. Therefore, we utilize the confidence interval for these two models to appraise the necessity of the special education services by examining whether zero lies within the confidence interval.

Table 1. Results of these three methods, including the significances of these methods, the estimated ATEs, the standard errors of estimated ATEs and the confidence intervals of estimated ATEs. Blank cells in the table indicate the absence of corresponding values for the respective method.

Method	Sig	ATE	Std	CI
PSM	6.66×10^{-4}	-4.501	27.149	
OLS		-6.251	15.989	(-6.616, -5.885)
BART		-5.055	15.268	(-5.404, -4.706)

5. Conclusion

This research aimed to identify the effectiveness of the special education services by estimating the ATE. Based on the ECLS-K dataset, we use both traditional method (PSM) and ML methods (OLS and BART) to estimate the ATE of the special education services.

In the Results section, we have analyzed detailed methods used for each outcome to assess the necessity of the special education services. Consequently, we now direct our attention to the reported results. Based on the results obtained through PSM, we can infer that the ATE is significantly different from zero at a 95% confidence level. Regarding OLS and BART, both the confidence intervals exclude zero at a 95% confidence level, indicating that the estimated ATEs are significantly different from zero.

Subsequently, we can proceed with the analysis of the estimated values. Through the ATE results listed above, we can conclude that these special education services have negative effect on students' fifth grade math score. Therefore, if a school is going to take such services for its students, their fifth grade math score will decrease. Obviously, this kind of phenomenon is opposite to the original intension of these special education services.

Besides, through the values of standard error form these three methods, we can see that the standard error of OLS and BART are only half of PSM's, which means the estimates ATE of OLS and BART have much higher accuracy than PSM's. Then we can conclude that ML methods are better than traditional method in this paper, which is correspond to what we expected. Moreover, we can conclude that BART is better than OLS after comparing their standard error, this is also reasonable.

However, there are still limitations in our study. In fact, we actually have no idea about the real situation under another treatment condition, although we have simulated this through some models. Therefore, the bias is existed and we can only obtain an estimate for the ATE. The exact error or deviation is invisible. In the future, if we can develop models to simulate the situation mentioned above more than digital form, but a simulated reality, maybe this bias could be reduced to a negligible value.

References

[1] Princiotta D and Flanagan K.D 2006 J. National Center for Education Statistics 38 69. Fifth grade: findings from the fifth grade follow-up of the early childhood longitudinal study, kindergarten class of 1998-99 (ecls-k). e.d. tab. nces.

[2] Herring W.L, Mcgrath D.J, and Buckley J.A 2007 J. National Center for Education Statistics 005. Demographic and school characteristics of students receiving special education in the elementary grades. issue brief. nces.

[3] National Center for Education Statistics (n.d.) Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K). Retrieved from https://nces.ed.gov/ecls/kindergarten.asp.

[4] Cui P, Shen Z, Li S, Yao L, Li Y, Chu Z and Gao J 2020 The 26th ACM SIGKDD Conf on Knowledge Discovery and Data Mining (Association for Computing Machinery) pp 3527-28. Causal Inference Meets Machine Learning.

[5] Kreif N and Diazordaz K 2019 Machine learning in policy evaluation: new tools for causal inference arXiv stat.ML/1903.00402.

[6] Mealli F 2015 J. Observational Studies 1 291-3. Review of the book "causal inference for statistics, social, and biomedical sciences" by G.W. Imbens and D.B. Rubin.

[7] Paul W and Holland 1986 J. Journal of the American Statistical Association 81 945-60. Statistics and causal inference.

[8] West S.G and Thoemmes F 2010 J. Psychological Methods 15 18. Campbell's and Rubin's perspectives on causal inference.

[9] Rosenbaum P.R and Rubin D.B 1983 J. Biometrika 70 41-55. The central role of the propensity score in observational studies for causal effects.

[10] Keller B and Tipton E 2016 J. Journal of Educational and Behavioral Statistics 41 326-48. Propensity score analysis in R: a software review.

[11] Guido W.I and Donald B.R. 2015 Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction (Cambridge: Cambridge University Press) p 282.

[12] Mcconnell J 2019 J. Health services research 54 1273-82. Estimating treatment effects with machine learning.

[13] Dijkuis T.B and Blaauw F 2022 J. Entropy 24 1060. Transferring Targeted Maximum Likelihood Estimation for Causal Inference into Sports Science.

[14] Xuefeng Z 2007 Least Squares and Least Squares Support Vector Machines (Chengdu: University of Electronic Science and Technology of China) pp 13-27.

[15] Minghui Y 2018 Research on extension of Bayesian cumulative regression tree and its model construction in survival analysis (Wuhan: Huazhong University of Science and Technology) pp 15-61.

[16] Jennifer L.Hill 2011 J. Journal of Computational & Graphical Statistics 20 217-40. Bayesian nonparametric modeling for causal inference.

[17] Athey S, Tibshirani J and Wager S 2019 J. The Annals of Statistics 47 1179-1203. Generalized random forests.

[18] Jung Y, Tian J, and Bareinboim E 2021 International Conf on Machine Learning (PMLR) pp 5168-79. Estimating Identifiable Causal Effects on Markov Equivalence Class through Double Machine Learning.

[19] Kang J 2022 J. Child abuse & neglect 132 105817. Spanking and children's social competence: evidence from a US kindergarten cohort study.

[20] Yang Y, Yiu Fai Y and Maura S 2017 Proc of the SAS Global Forum 2017 Conf (Orlando, FL, USA) pp 2-5. Propensity Score for Causal Inference with the PSMATCH Procedure.

[21] Abadie A and Imbens G. W 2009 J. Econometrica. NBER Working Paper Series No.w15301 781-807. Matching on the estimated propensity score.

[22] Osvaldo A.M, Ravin K and Junpeng L 2022 Bayesian Modeling and Computation in Python (Boca Raton: CRC Press).

[23] Chipman H.A, George E.I and Mcculloch R.E 2006 Bayesian Ensemble Learning. Neural Information Processing Systems (Cambridge, MA: MIT Press).