The impact of early childhood education and medical care on children

Tianshi Liu*

Abstract.

Causal inference is a statistical approach that aims to understand and quantify the causal relationship between variables, allowing us to determine the impact of one variable on another while accounting for potential confounding factors. In this study, we chose the Infant Health and Development Program (IHDP) dataset to test whether high-quality early childhood education and medical care will enhance their cognitive and academic ability. We use some common and classic algorithms to achieve this study by calculating the CATE (Conditional Average Treatment Effect) of the dataset; if the result is a negative number, that means early treatment doesn't improve children's further ability; if the result is a positive number, that means treatment improves it. After fitting the targeted models, BART(Bayesian Additive Regression Trees) and Metalearners, into this dataset, we found out that all the models will get positive CATE, which means that these early treatments positively affect children's future development. Also, we can judge which model gives us a more accurate result according to the standard variance of TE(treatment effect). The result of this study will provide us with an insightful idea about whether we should give children some treatments regarding education and medical care and which model is more suitable for this casual inference test.

Keywords: IHDP, CATE, BART, Metalearners.

1 Introduction

The Infant Health and Development Program (IHDP) dataset is a longitudinal dataset that was collected from 1985 to 1990 in eight sites across the United States. The dataset includes information on 985 low birth-weight infants and their families randomly allocated to a control or intervention group. The intervention group received high-quality early childhood education and medical care, while the control group received standard care. The dataset includes information on the children's cognitive and academic development, as well as their health and family background [1].

In this study, we employ established causal inference models to examine the CATE (Conditional Average Treatment Effect) on the dataset. CATE represents the average treatment effect of a specific intervention or treatment on a particular subgroup, considering specific covariates or conditions. Estimating CATE enables us to comprehend the variability in treatment effects among different subgroups within the population, allowing for the identification of heterogeneous treatment effects that may vary based on individual characteristics or circumstances. But when we consider all the conditions, CATE becomes ATE (Average Treatment Effect), and in this article, we mainly accurately calculate the ATE of the IHDP dataset [2].

In this article, we introduce the Metalearners and Bayesian Additive Regression Trees, two algorithms widely used in the area regarded as causal inference to calculate CATE. The former is a simple and convenient algorithm to achieve, and the latter is a classic one. It uses regression trees and combines a sum-of-trees model and regularization prior. Meanwhile, it is flexible in nonlinear and interactive aspects of fitting data.

The ways of calculating the average treatment effect are as follows: we use the original dataset to train models. Prepare the cleaning data and choose BART and Metalearners models to calculate CATE. Estimate the treatment effect model using the selected models. The average difference in outcomes between the treated and control groups can be used to interpret. Finally, check the results, including standard deviations, confidence intervals, and final CATE. The results from both algorithms can tell us the effect of early treatment on children and which model has the most accurate result among others.

2 Literature Review

Causal inference is a fundamental problem in many fields, including statistics, machine learning, economics, and social sciences. It aims to infer the causal relationship between variables from observational or experimental data. In recent years, the development of causal inference methods has advanced significantly, and new methods have been proposed to deal with different types of data and research questions. In this literature review, we will introduce some of the most common causal inference methods and their applications in various fields.

2.1 Propensity Score Matching

Propensity score matching is a widely used for estimating causal effects from observational data. Propensity scores, the conditional probabilities of receiving the treatment given the reported covariates, are used to match treated and control units. The matched pairs have similar propensity scores, which reduces the bias due to confounding factors.

2.2 Instrumental Variables

Instrumental variables estimate causal effects in situations with unobserved confounding factors. The method relies on an instrumental variable, which is a variable that affects the treatment assignment but is not directly related to the outcome. By using the instrumental variable, researchers can isolate the effect of the treatment from the effect of the confounding factors.

2.3 Regression Discontinuity Design

Regression discontinuity design estimates causal effects when the treatment assignment is based on a continuous variable, and the treatment effect varies discontinuously at a cutoff point. The method involves comparing the outcomes of units just above and below the cutoff point, effectively creating a randomized experiment around the cutoff point.

2.4 Difference-in-Differences

Difference-in-differences is a method for estimating causal effects in situations with a treatment group and a control group, and the treatment assignment is not random. The method involves comparing the changes in outcomes before and after the treatment for the treatment group and the control group. If the treatment effect is significant, the changes in outcomes for the treatment group will significantly differ from those for the control group.

2.5 Bayesian Additive Regression Trees

Bayesian additive regression trees (BART) have been widely used in causal inference tasks and have succeeded significantly in various applications. One major advantage of BART is its flexibility in modeling complex nonlinear relationships between the covariates and the response variable. For example, Hill used BART to estimate heterogeneous treatment effects in randomized experiments and showed that BART outperformed other methods, such as regression trees and propensity score matching, regarding mean squared error and bias [3]. Wager applied BART to the problem of estimating treatment effects in observational studies with unmeasured confounding and showed that it could yield accurate estimates of the average treatment effect even when traditional methods failed [4].

Another advantage of BART is its ability to handle high-dimensional data, which is increasingly common in modern data analysis. Linero and Pérez-Stable proposed a Bayesian hierarchical model based on BART to estimate causal effects in high-dimensional settings. They showed it could identify important predictors and achieve better prediction performance than other methods, such as Lasso and random forests [5].

Moreover, BART has been applied to many other causal inference tasks, such as estimating causal mediation effects [6], detecting causal interactions [7], and estimating personalized treatment rules.

In summary, BART has demonstrated its effectiveness and flexibility in various causal inference tasks, making it a powerful tool in causal inference.

2.6 MetaLearners

Meta-learners have been widely used in causal inference to estimate heterogeneous treatment effects and have achieved notable success in various applications. One popular Meta-learner is the "Tlearner," which estimates the conditional expectation of the outcomes separately for control and treated units using base learners such as linear regression or tree-based methods and then takes the difference between these estimates to obtain the CATE.

Another widely used meta-learner is the Slearner, which uses a single estimator to predict the outcome with all characteristics and the treatment indicator and calculates the difference in forecast values when the treatment assignment indicator changes. The Xlearner seeks to combine the benefits of the Tlearner and Slearner by estimating the outcome function separately for control and treated units using two base learners and then combining them using weights dependent on the degree of overlap between the covariate distributions of the two groups.

MetaLearners have also been applied in various realms, such as healthcare, economics, and education, to estimate treatment effects and identify subpopulations that may benefit the most from certain treatments. For example, in a study on the impact of antidepressants on depression symptoms, the Tlearner and Xlearner were used to estimate heterogeneous treatment effects based on the patient's characteristics, and the results showed that patients with more severe symptoms tended to benefit more from the treatment [8]. In another study on the effects of a math intervention program on students' achievement, Slearner and Xlearner were used to identify subgroups of students that benefited the most from the program, and the results showed that the program had the greatest effect on students with low baseline math scores [9].

3 Methods

3.1 BART

3.1.1. Decision Tree

3.1.1.1. Brief introduction

Bayesian Additive Regression Trees (BART) is a machine learning technique that combines Bayesian modeling and decision trees for regression and classification tasks. BART models a response variable as the sum of many tree models, where each tree is assigned a weight drawn from a Bayesian prior distribution. BART can model nonlinear and non-additive relationships between predictors and the response and can also handle interactions and high-dimensional predictor spaces.

BART is a flexible and powerful method applied in various domains, such as economics, finance, biology, and healthcare. It has been used for prediction, variable selection, and causal inference tasks, and it has shown competitive performance compared to other popular machine learning algorithms.

3.1.1.2. Mathematical definition

The model can be written as:

$$y_i = \alpha + \sum_{m=1}^{M} f_m(x_i) + \varepsilon_i \quad (1)$$

Where is the response variable for the th observation, x_i is the vector of predictor variables for the th observation, α is the intercept term, f_m is the the regression tree in the ensemble, and ε_i is the error term. The regression trees are built using a recursive partitioning algorithm and are combined using a Bayesian model averaging approach. BART uses a prior distribution on the regression trees that encourages smoothness, prevents overfitting, and allows for complex interactions between the predictors.

The regression tree topologies, prior parameters, and error variance are only a few of the model's parameters sampled from the posterior distribution using the Markov Chain Monte Carlo (MCMC) process to fit the model. The MCMC algorithm can be used to obtain point estimates and credible intervals for the model parameters and perform hypothesis testing and model selection [10].

3.1.1.*3. Regression Trees.*

Regression trees are a popular nonparametric statistical method for modeling relationships between a dependent variable and a set of independent variables. They are often used for prediction, where the goal is to find a model function that accurately predicts the dependent variable's value according to the independent variables' values.

The basic idea behind regression trees is to recursively partition the data into subsets based on the independent variables' values. At each step, the algorithm chooses the variable and the split point that minimizes the sum of squared errors (SSE) of the resulting subsets. This process is repeated until some stopping criterion is met, such as a minimum subset size or a maximum tree depth.

By moving up the tree from the root to a leaf node representing the values of the independent variables, the resulting tree can forecast the value of the dependent variable for fresh observations. The predicted value is typically the mean or median of the dependent variable in that leaf node [10].

3.1.2. Ensembles of Decision Trees

Usually, we introduce devices to reduce the complexity of decision trees and get a fit that better adapts to the complexity of the data at hand. One such solution relies on fitting an ensemble of trees where each tree is regularized to be shallow. As a result, each tree can only explain a small portion of the data individually. Only by combining many such trees can we provide a proper answer. Bayesian methods like BARTs and non-Bayesian methods like random forests follow this ensemble strategy. In general, ensemble models lead to lower generalization errors while maintaining the ability to fit a given dataset flexibly. Using ensembles also helps to alleviate the step-ness. The downside of this method is that we lose the interpretability of a single decision tree [10].

3.1.3. The BART Model

The BART Model is used to calculate the model's outcome. Here is the equation:

$$y = f(z,x) + \varepsilon, \varepsilon \sim N\left(0,\sigma^2\right) \quad (2)$$

Represents the observed confounding covariates and denotes the allocated treatment. A regularization prior and a sum-of-trees model comprise BART's two components [10].

3.1.3.*1. Sum-of-Trees Model.*

The sum-of-trees model is an additive model with multivariate parts. It is substantially more flexible than a single tree model and more versatile than conventional additive models that use low dimensional smoothers as components.

First, we develop a notation for a single tree model. Let T denote a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and let $M = \{\mu_1, \mu_2 \ldots \mu_B\}$ denote a set of parameter values associated with each of the B terminal nodes of T. Prediction for a specific input vector x value is carried out as follows: If x is associated with terminal node b of T by the sequence of decision rules from top to bottom, it is then assigned the μb value associated with this terminal

node. We use use $(x\ ;\ T,\ M)$ to denote the function corresponding to (T M), which assigns a $\mu_b \in M$ to x. Using this notation, The function equivalent to that assigns a to is denoted by the symbol. By using this notation, we can more clearly state our sum-of-trees model as:

$$Y = g(x;T_1,M_1) + g(x;T_2,M_2) + \ldots + g(x;T_m,M_m) + \varepsilon, \varepsilon \sim N\left(0,\sigma^2\right) \quad (3)$$

A sum-of-trees model acquires greater representation flexibility as the number of trees increases, and this, along with our regularization prior, results in outstanding out-of-sample prediction accuracy [10].

3.1.3.2.A Regularization Prior

BART uses a Bayesian framework for model fitting, specifying a prior distribution over the model's parameters. One of the key contributions of BART is the introduction of a novel prior distribution called the "regularization prior." The regularization prior is used to encourage smoothness in the function that maps covariates to responses, which helps to prevent overfitting and improve the model's generalization performance.

The regularization prior is a Gaussian process with a specific covariance function that depends on a parameter controlling the degree of smoothness. When this parameter is small, the prior places little emphasis on smoothness, and the resulting function can be very complex and overfit to the training data. When the parameter is large, the prior emphasizes smoothness, and the resulting function will be smoother and less likely to overfit. The value of this parameter can be selected through cross-validation or other model selection techniques.

The regularization prior is used in conjunction with a prior distribution over the trees in the BART model to create a joint prior distribution over the entire model. The resulting joint prior distribution is then used to define the likelihood function for the observed data and perform Bayesian inference to obtain posterior distributions over the model parameters [10].

3.2 MetaLearners

We have four learners in MetaLearners, which are S, T, X, and R. They are different algorithms used in the Meta-Learners approach for estimating the conditional average treatment effect (CATE) in causal inference; in this study, we will use three of them to apply.

3.2.1 Tlearner

The conditional outcome expectancies for treated and control units separately are estimated using base learners in two phases, and the difference between the estimates is then used to estimate heterogeneous treatment effects. Linear regression or tree-based techniques might be used as the basis for learners. The latter is referred to as the "two-tree" estimator, and we name this method the " Tlearner," where "T" stands for "two."[11].

3.2.2 Slearner

Another comparable technique is using all the attributes and the treatment indication to estimate the outcome without giving the treatment indicator any special consideration. The predicted individual CATE is the difference between the anticipated values when the treatment-assignment indicator is switched from control to treatment, with all other variables held constant. This approach, which uses a single estimator, is known as the "Cleaner" and has been examined with regression trees and Bayesian additive regression trees (BARTs) as the base learners [11].

3.2.3 Xlearner

The Xlearner is a meta learner that combines the strengths of both the Tlearner and Slearner approaches to improve the estimation of CATE. The Xlearner first uses two base learners to estimate the outcome function separately for treated and control units, similar to the T learner. In the second stage, the estimated outcome functions are combined to obtain the CATE estimate, with the weight assigned to each estimator determined by the degree of covariate distribution overlap between the two groups, similar to the Cleaner. Research has shown that the Xlearner outperforms the Tlearner and Slearner in situations where treatment effects vary significantly across different subpopulations and where the overlap between the covariate distributions of the treated and control groups is moderate to low [12].

4 Result

Here, we show all the CATE results and TE's standard variance in Table 1 below. From the table, we can see that all the values of CATE are positive, and the S-learner model has the least standard variance of TE, which means that the early treatment brings children a positive effect on their future development and the S-learner model can provide us with the most accurate results compared to other models. Now we know which model is more suitable for this casual inference problem.

Table 1. The conditional average treatment effect of four models.

	CATE	Standard Variance of TE
BART	2.4417	3.2425

X-learner	4.0706	0.5550
T-learner	3.9909	0.6303
S-learner	3.9426	0.3735

5 Conclusion

The research aims to identify the effectiveness of early childhood education and medical care services on children's cognitive and academic development by estimating the conditional average treatment effect. Based on the Infant Health and Development Program dataset, we use the Meta-learners algorithm and BART model in machine learning to estimate the CATE. The result of the two methods obtained four positive CATEs, which means these education and medical services would increase their ability in children's cognitive and academic abilities in their lives. So, based on the result, we better give children some suitable education and medical care in their early stages.

Reference

1. Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20.1, 217-240.

2. Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113.27, 7353-7360.

3. Hill, J. L., & Su, Y. S. (2011). Assessing differential treatment effects with tree-based models: Evidence from heart attack patients. Political Analysis, 19.4, 385-408.

4. Wager, S., Athey, S., & Imbens, G. W. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113.523, 1228-1242.

5. Linero, A. R., & Pérez-Stable, E. J. (2018). Bayesian hierarchical model for causal inference in high-dimensional longitudinal studies. Journal of the American Statistical Association, 113.523, 646-659.

6. Ibrahim, J. G., Chen, M. H., Sinisi, S. E., & Kim, Y. (2019). Bayesian methods for estimating causal mediation effects using principal stratification with dichotomous mediators. Journal of the American Statistical Association, 114.528, 812-824.

7. Kern, C., Wagner, S., & Schmid, V. J. (2019). Identifying causal interactions in a high-dimensional Bayesian additive regression tree framework. Journal of Computational and Graphical Statistics, 28.1, 180-192.

8. Koshiaris, C., & Car, J. (2021). Antidepressants for depressive symptoms in patients with comorbid physical illness: a systematic review and meta-analysis. Journal of General Internal Medicine, 36.1, 260-267.

9. Huang, W., & Aljadeff-Abergel, E. (2021). The effect of a math intervention program on elementary students' math achievement: Evidence from a randomized controlled trial. Journal of Educational Psychology, 113.2, 201-215.

10. Ibrahim, Joseph G. Ming-Hui Chen, and Debajyoti Sinha. (2009). Computational Statistics 1.2. In: Wiley Interdisciplinary Reviews, WB, Bayesian Survival Analysis. 152-159.

11. Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. In: Proceedings of the National Academy of Sciences, 116.10, 4163.

12. Künzel, Sören R. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning.In: Proceedings of the National Academy of Sciences 116.10: 4156-4165.