mathematics

Article

Evaluation of Educational Interventions Based on Average Treatment Effect: A Case Study

Jingyu Liang and Jie Liu *

School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China
* Correspondence: liujie805@buaa.edu.cn

Abstract: Relative to randomized controlled trials (RCTs) with privacy and ethical concerns, observational studies are becoming dominant in education research. In an observational study, it is necessary and important to correctly evaluate the effects of different interventions (i.e., covariates) on student performance with observational data. However, these effects' evaluation results are probably derived from biased estimations because the distributions of "control" and "treatment" student groups can hardly be equivalent to those in RCTs. Moreover, the collected covariates on possible educational interventions (i.e., treatments) may be confounded with student characteristics that are not included in the data. In this work, an estimation method based on the Rubin causal model (RCM) is proposed to calculate the average treatment effect (ATE) of different educational interventions. Specifically, with the selected covariates, the propensity score (i.e., the probability of treatment exposure conditional on covariates) is considered as a criterion to stratify the observational data into sub-classes with balanced covariate distributions between the control and treatment groups. Combined with Neyman's estimation, the ATE of each sample is then obtained. We verify the effectiveness of this method with real observational data on student performance and its covariates.

Keywords: observational study; propensity score; data stratification; average treatment effect; education

MSC: 62H12

check for updates

Citation: Liang, J.; Liu, J. Evaluation of Educational Interventions Based on Average Treatment Effect: A Case Study. *Mathematics* **2022**, *10*, 4333. https://doi.org/10.3390/math10224333

Academic Editor: Liangxiao Jiang

Received: 16 October 2022
Accepted: 16 November 2022
Published: 18 November 2022

1. Introduction

In education research, the effects of educational interventions such as programs and policies on student performance need to be properly and correctly estimated to provide informative suggestions to improve education quality. Currently, the best process is to assign different treatments through randomized controlled trials (RCTs) [1]. However, incomplete and inconsistent (because missing) data sometimes coexist in RCTs. Interventions may also vary by student or school, and the data can hardly be random as designed in RCTs. The research results may thus not precisely focus on the groups of students and schools affected by the considered interventions [2]. In recent years, because of the difficulty in conducting RCTs, observational studies have been widely considered in education research [3].

Observational research is used in non-randomized comparative studies and refers to the observation and description of the characteristics of the research object in a natural state, as well as the comparative analysis of the recorded data. Considering a specified intervention (e.g., a specific education program), the observational data can generally be divided into treatment and control groups. The treatment group includes the students who take the specified intervention, while the students in the control group do not follow the specific intervention. The results of observational studies can better evaluate the actual effect of interventions, that is, a good observational study has better external authenticity than an RCT [4]. Observational studies have helped the academic community to clarify many key causal relationships, and they are increasingly being adopted in studies of comparative effectiveness. Observational studies are particularly useful when RCTs cannot be performed for reasons such as ethnicity or incomplete randomization [4]; more specifically, in education research, interventions such as individual growth environment, caste, education level, and other factors are hard to be randomly controlled in RCTs. With effective experimental design and data processing methods, however, observational studies can be more coherent with the truth [3].

In data processing, correlation analysis is widely adopted to analyze the statistical dependence between two variables [5]. Another popular method in statistics for observational study data analysis is the multivariate logistic regression [6]. Through the logistic regression method, the probability of a sample belonging to one class in a binary classification can be obtained with the collected data. However, neither of these previous methods can properly tackle the main challenges in the observational studies, that is, (1) the biased estimated effects brought by the unbalanced distribution of the covariates in the control and treatment groups and (2) unreliable estimated effects brought by unobservable student characteristics that confound student performance and the considered covariates.

In order to solve the two main problems of observational studies, propensity score (PS) analysis is considered in practice. PS is a method for balancing the distribution of covariates [7] (i.e., the observable variables related to student performance, except for the considered intervention) in the treatment and control groups. The PS is the conditional probability of the specific intervention, given the values of the covariates [8]. With the calculated PS, the observational data can be balanced with matching methods [9,10], stratification methods [11–13], or weighting [14]. Rosenbaum and Rubin [15] have pointed out that the balance of PS produces an average balance in the observed covariates. PS analysis can also yield unbiased causal effect estimates when all relevant covariates are available. The purpose of adjusting PS is to achieve a balance in the observed covariates between the treatment and control groups and thus to reproduce the expected outcomes in RCTs [16,17].

In recent years, many researchers have focused on exploring the application of PS in the education field. Marvin G. Powell et al. [18] outline the concept of propensity scores by explaining their theoretical principles and by providing two examples to identify their usefulness within the realm of educational research. Lucio Masserini and Matilde Bini [19] use PS to investigate whether university students' dropout within the first year is influenced by participation in social media groups. They use several PS matching techniques and sensitivity analyses in order to correct for selection bias due to a set of observable pretreatment covariates. Jianshen Chen and Bryan Keller [20] propose a five-step approach using PS matching and regression trees to identify subgroups with heterogeneous treatment effects in observational studies. Youmi Suk et al. [21] investigate causal forests to estimate treatment effects in multilevel observational data. In recent education research, many estimation methods with PS are used in observational studies. However, some of them only consider the PS method to balance the covariates, rather than the causal inference of the covariates. Further, some of the researchers introduce the machine learning method to estimate the effects of a covariate, without considering the balance of different covariates.

In this paper, a framework based on the Rubin causal model (RCM) is proposed for bridging the research gap in educational studies. Papers by Neyman [22] and Rubin [23] provide a statistical cornerstone for causal inference in experimental and quasi-experimental research, known as the RCM [24]. Their work provides a framework for integrating causality into a statistical model and demonstrates that statistical theory can be of great help in discussions of causal inference [25]. In addition, their work has led to a greater understanding of counterfactual causality in the statistical and social science communities, as well as advancing counterfactual causality in the estimation of treatment effects [26].

This paper introduces the PS analysis method to balance the covariates in treatment and control groups in the observational studies of the educational field; the observational data are stratified into sub-classes with respect to the PS and, in the sub-classes, the covariates are balanced in distribution. Under the framework of RCM, this paper uses Neyman's estimation method to calculate the average treatment effect (ATE) of the samples in each sub-class, then estimates the total ATE of all of the observational data. It provides detailed steps and stratification criteria for stratification and ATE estimation, and the accurate ATE value between each intervention and the student performance can be obtained, which can be used to evaluate the effectiveness of various educational interventions.

The remainder of the paper is structured as follows. Section 2 details the methods including PS calculation, stratification, and ATE calculation. Section 3 introduces the dataset used in the work. Section 4 presents the process of calculation and the results' analysis and, furthermore, compares the results with traditional methods. Finally, Section 5 summarizes the proposed method.

2. Methods

2.1. PS Calculation

An effective method for processing observational data is to estimate the PS for distribution bias and then to obtain a credible ATE estimate. The PS is usually estimated by a linear regression model, with the output being in the range from zero to one. To better tackle the distribution bias problem, the PS estimation method introduced in this paper adds the process of iterative covariate selection to the original linear regression model. The detailed steps of this method for the analysis and processing of observational data are provided below.

2.1.1. Principle of PS

The PS was first proposed by Rosenbaum and Rubin in 1983 [15]. PS is a function of multiple covariates and is used to deal with covariate distribution bias across different groups in observational studies. It is the conditional probability that the *i*-th individual is assigned to the treatment group, given the specific values for the related covariates:

$$e(X) = P(G = 1|X) \tag{1}$$

where $G = 1$ and $G = 0$ indicate that the individual is in the treatment group and the control group, respectively; X represents the covariates with $\mathrm{X} = X_1, X_2, \ldots, X_m$. The PS value is thus the conditional probability of an individual receiving an intervention ($G = 1$), given the characteristic values X.

Unlike in RCTs, observational studies have many confounding factors that can significantly bias the results. For example, a research team wants to study the factors influencing middle school students' performance in a mathematics course. In this case, the outcome variable is math test scores, focusing on the impact of using multimedia for learning on mathematics. There are other confounding factors, such as the student's gender, tutoring time, the socioeconomic status of the student's family, teacher education, and university majors. When analyzing the effect of multimedia on students' mathematics performance, one hopes to minimize the influence of confounding factors on the results of the analysis—that is, distributions of the covariates in the treatment group (using multimedia for learning) and the control group (without multimedia learning) are as identical as possible. PS can be very useful in such cases.

According to PS, we can screen the treatment and control groups, so that the confounding factors (covariates) in different groups can be balanced to achieve the purpose of control. The PS itself cannot control confounding factors, but can directly adjust confounding factors through PS matching, weighting, or stratification (among other strategies) to improve the balance degree between the treatment and control groups, thereby limiting the influence of confounding covariates on the estimation of causal effects. A simple understanding is to select the treatment and control group samples with common characteristics from a large number of sample data and then analyze the samples that meet the requirements.

In practical research, logistic regression models are usually used to estimate PS values. PS estimation is divided into the following steps: dividing the treatment and control groups, selecting covariates, calculating the PS using binary logistic regression, and trimming. In addition to the PS calculation process, this section also presents an average causal effect estimation method based on sub-classification, which can better handle the average causal effect between variables in a sample with a small amount of data. In the following paragraphs, the content and method for each step are described in detail.

2.1.2. Division of Treatment and Control Groups

Generally speaking, in observational studies, treatment groups and control groups are divided according to the intervention (variable) to be analyzed. For example, in the case mentioned above, to analyze the effect of multimedia on students' mathematics performance, one takes the students who use multimedia as the treatment group and the students who do not use multimedia as the control group.

2.1.3. Covariate Selection

The process of covariate selection is shown in Figure 1 and is described in detail in the following paragraphs.

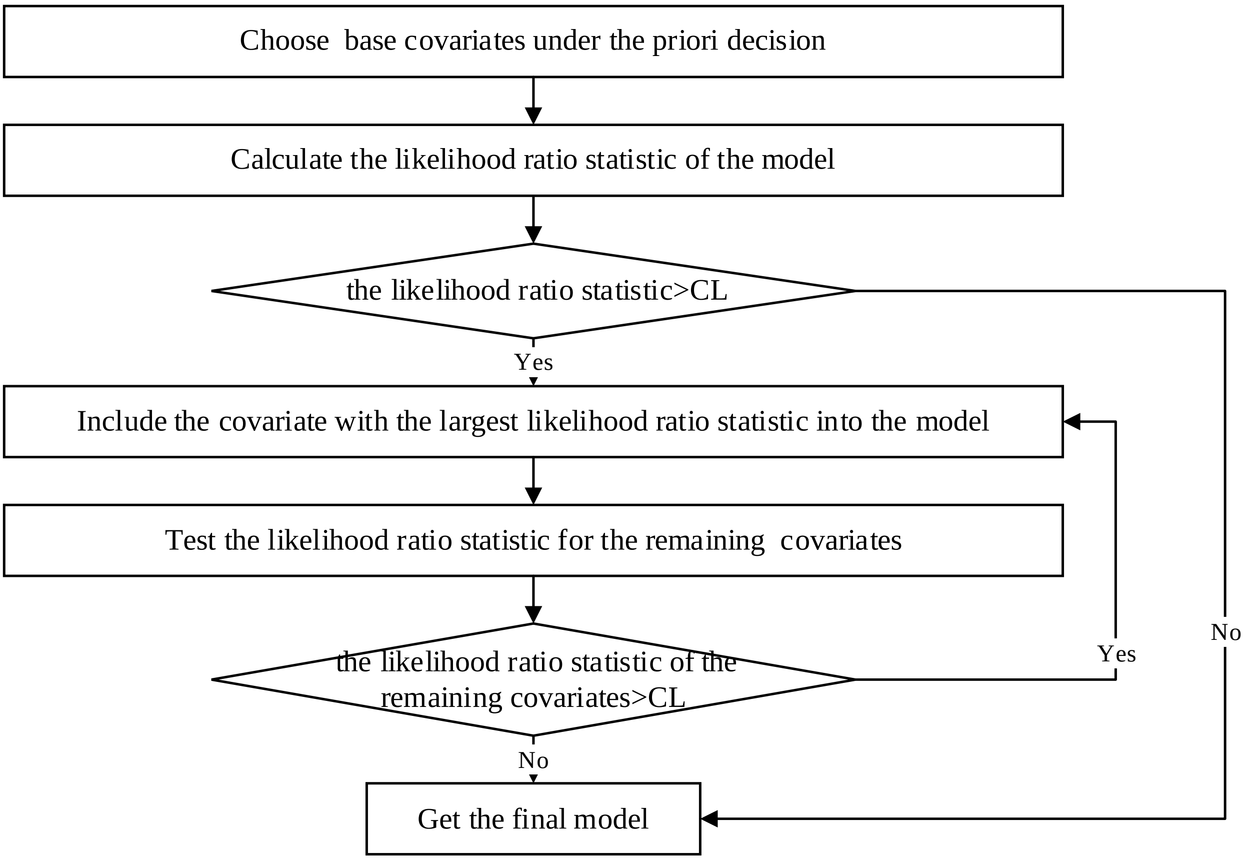

Choose base covariates under the priori decision
Calculate the likelihood ratio statistic of the model
the likelihood ratio statistic>CL
Yes
Include the covariate with the largest likelihood ratio statistic into the model
Test the likelihood ratio statistic for the remaining covariates
the likelihood ratio statistic of the remaining covariates>CL
Yes
No
No
Get the final model

Figure 1. The process of covariate selection.

(1) Choose prior covariates

We choose K_B base covariates under the a priori decision of the observational test. Two types of covariates are selected: the ones that explain the allocation mechanism and the others that are highly correlated with the potential outcome. Both types of covariates depend on prior information. If the researcher has no prior information on the covariates, $K_B = 0$ can be set.

(2) Add linear combination terms

In the second step, we select the remaining covariates and gradually integrate them into the PS estimation model. The remaining $K - K_B$ covariates are not all introduced into the model; only one of them is considered for inclusion in the logistic regression model at a time. Assume that K_L' covariates have been selected at a certain time as the linear combination item in the logistic regression model, including the basic covariate with dimension K_B. Each of the remaining $K - K'_L$ covariates is incorporated into the logistic regression model and the likelihood ratio statistic is calculated to evaluate the null hypothesis that the newly included covariate has a coefficient of zero in the regression model.

If all covariates are included in the original regression model and the likelihood ratio statistic is lower than C_L, the remaining covariates are not included in the regression model. At this point, only the linear combination of K'_L covariates from the original data is included in the PS estimation model. If the likelihood ratio statistic of at least one covariate after inclusion in the model exceeds C_L, the covariate with the largest likelihood ratio statistic is added into the model. Then, we have $K'_L + 1$ covariates in the model and we should test the likelihood ratio statistic for the remaining $K - K'_L - 1$ covariates to determine whether they should be included in the existing model. This process is iterated until the likelihood ratio statistics of the remaining covariates are less than C_L, and the covariate dimension of the final model of the iteration is denoted as K_L.

2.1.4. Calculation of PS Value for Individuals

In an observational study, proper regression methods are generally used to estimate the PS value for each individual. A logistic regression model is established to derive the probability of an individual belonging to the treatment group. Logistic regression has categorical outputs and the PS is the probability derived by the logistic regression model. The formula for the logistic regression is as follows:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \tag{2}$$

where P is the probability of an individual being in the treatment group and X is a potentially confounding variable. Linear regression is established to characterize the relationship between the input variables X and $\ln\left(\frac{P}{1-P}\right)$. Based on the above equation, the value of P can be expressed as follows:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}} \tag{3}$$

SPSS software is adopted to calculate the PS in this study. In SPSS, the steps for PS calculation are as follows: analysis, regression, binary logistic regression, setting dependent and independent variables, setting categorical covariate, and output predicted value probability.

2.1.5. Trimming

If the PS is close to 0, it is difficult for the control group to find matching members in the treatment group, while if the PS is close to 1, it is difficult for the treatment group to find matching members in the control group. A more practical approach is thus to exclude individuals with a PS close to 0 or 1, which is named as trimming. Trimming changes the measure and compromises external validity, in which case the average causal effect estimated with the truncated sample cannot represent the average causal effect of the original samples. However, trimming improves internal validity and the average causal effect estimate in the trimmed sample is more precise and credible than the untrimmed sample [27].

2.2. Stratification

Stratification is a PS analysis method. After obtaining the PS for all of the samples, the covariates can be balanced using a stratified approach. This paper presents the standard and specific steps for sample stratification.

To ensure that there are comparative pairs between the treatment and control groups, it is necessary to overlap the two groups of data, and we use the PS stratification method to process the truncated samples. The method for constructing the PS stratification is as shown in Figure 2.

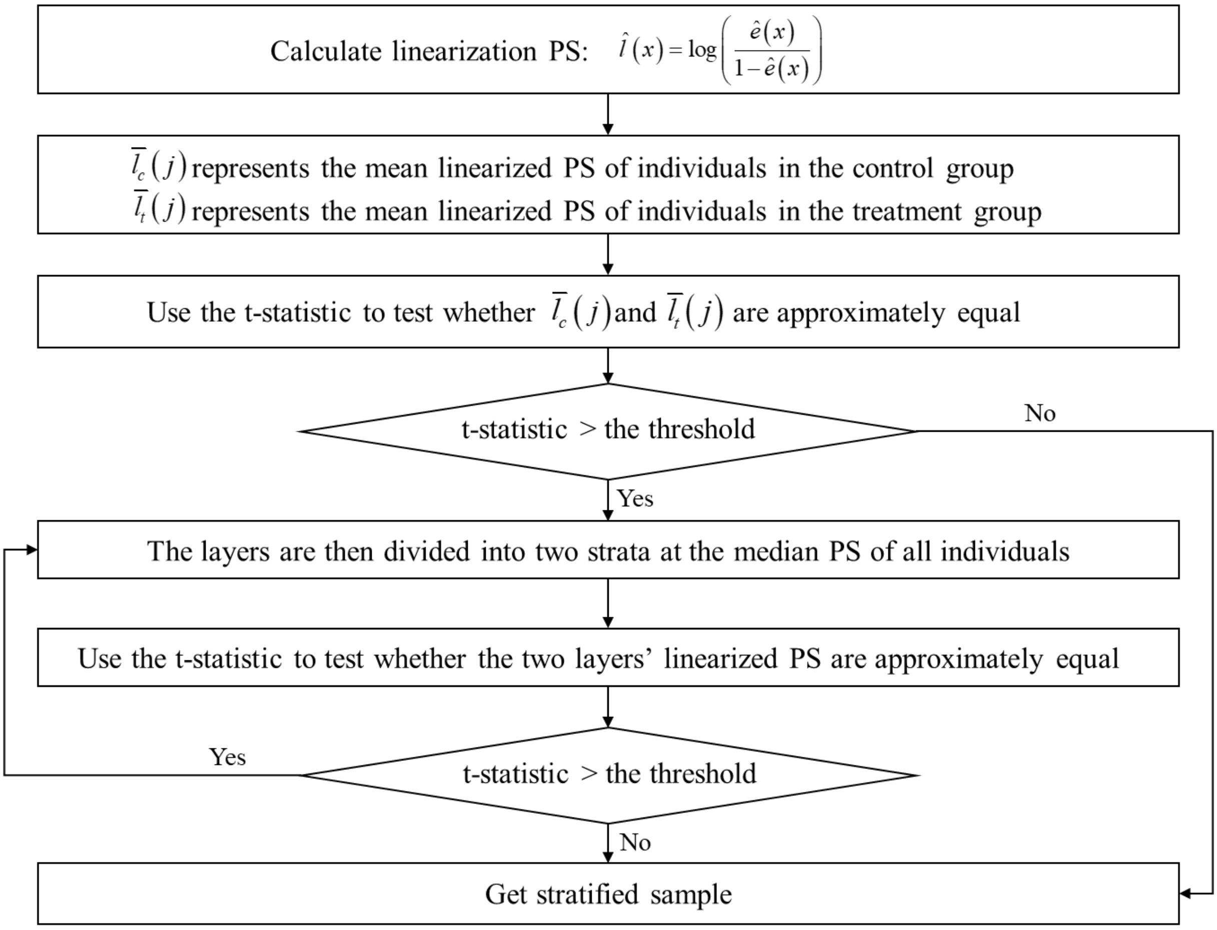

Calculate linearization PS: $\hat{l}(x)=\log\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$
$\overline{l}_c(j)$ represents the mean linearized PS of individuals in the control group
$\overline{l}_t(j)$ represents the mean linearized PS of individuals in the treatment group
Use the t-statistic to test whether $\overline{l}_c(j)$ and $\overline{l}_t(j)$ are approximately equal
t-statistic > the threshold
No
Yes
The layers are then divided into two strata at the median PS of all individuals
Use the t-statistic to test whether the two layers' linearized PS are approximately equal
Yes
t-statistic > the threshold
No
Get stratified sample

Figure 2. The process of stratification.

(1) Linearization of PS values

We linearize the PS values by taking the log ratio of PS as in Equation (4):

$$\hat{l}(x) = \log\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right) \tag{4}$$

(2) Stratification standard and *t*-statistic test

We define

$$N_c(j) = \sum_{i=1}^{N} (1-W_i)B_i(j) \tag{5}$$

$$N_t(j) = \sum_{i=1}^{N} W_i B_i(j) \tag{6}$$

$$\overline{l}_c(j) = \frac{1}{N_c(j)} \sum_{i=1}^{N} (1-W_i)\cdot B_i(j)\cdot \hat{l}(X_i) \tag{7}$$

$$\overline{l}_t(j) = \frac{1}{N_t(j)} \sum_{i=1}^{N} W_i\cdot B_i(j)\cdot \hat{l}(X_i) \tag{8}$$

where $N_c(j)$ represents the number of individuals in the control group; $N_t(j)$ represents the number of individuals in the treatment group; $W_i = 1$ represents the i-th individual belonging to the treatment group; $W_i = 0$ represents the i-th individual belonging to the control group; $B_i(j)$ represents a stratification index, $B_i(j) = 1$ means the j-th sample is within i-th layer, and $B_i(j) = 0$ means the j-th sample is not in i-th layer; $\bar{l}_c(j)$ represents the mean linearized PS of individuals in the control group; and $\bar{l}_t(j)$ represents the mean linearized PS of individuals in the treatment group. Based on Equations (5)–(8), we use the t-statistic to test whether $\bar{l}_c(j)$ and $\bar{l}_t(j)$ are approximately equal.

(3) The t-statistic is a common measure of the difference of mean values of two groups of data. Considering that the distribution bias in our case study refers to the difference in the mean values between the treatment and control groups, the t-statistic is considered to judge the significance of the difference between treatment and control groups. If the obtained t-statistic exceeds the predefined threshold, it indicates that the PS values in treatment and control groups exhibit a significant difference in this layer and need to be further stratified.

(4) The layers are then further divided into two strata at the median PS of $N_c(j) + N_t(j)$ individuals. The two new layers must satisfy the following criterion: the number of control groups and treatment groups in each new layer is not less than $K + 2$, where K is the number of covariates. For example, in the process of stratification, if the i-th layer is divided and it is found that the number of samples in treatment/control groups in the new layers is less than $K + 2$, the stratification is stopped even if the t-statistic is not satisfied. This is a stop criterion for the stratification process and the influence is limited as the other stratification process satisfies the t-statistic threshold.

2.3. Calculation of the Average Treatment Effect

The ATE is the expected difference between the treatment and control groups for each individual. The larger the value, the greater the influence of the intervention variable on the student performance. We thus calculate the ATE to obtain the relationship between the intervention variable and the response variable. The ATE based on the Neyman estimate is calculated as follows. First, the estimated ATE between individuals within each layer needs to be calculated. We can use weighted estimators to estimate it as Equation (9):

$$\hat{\tau}^{\text{dif}}(j) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i^{obs}}{\hat{e}(X_i)} - \frac{(1 - W_i) \cdot Y_i^{obs}}{1 - \hat{e}(X_i)}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(W_i - \hat{e}(X_i))Y_i^{obs}}{\hat{e}(X_i)(1 - \hat{e}(X_i))}\right) \tag{9}$$

where $\hat{\tau}^{\text{dif}}(j)$ is the estimated ATE difference within the j-th layer; $\hat{e}(X_i)$ is the estimated PS; $W_i = 1$ represents the i-th individual belonging to the treatment group; $W_i = 0$ represents the i-th individual belonging to the control group; and Y_i^{obs} represents the observed value of the i-th individual.

The ATE based on sub-classification is estimated as follows:

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j)\hat{\tau}^{\text{dif}}(j) \tag{10}$$

$$q(j) = \frac{N(j)}{N} \tag{11}$$

where $\hat{\tau}^{\text{strat}}$ represents the final ATE; $\hat{\tau}^{\text{dif}}(j)$ represents the ATE of the j-th layer; $N(j)$ represents the number of individuals in this layer; and N represents the total number of individuals in the dataset. This method of estimating the difference in the ATE is called the Neyman inference for sub-classification estimation. The ATE of each layer is calculated and then weighted according to the number of samples and, finally, the ATE of the entire dataset is obtained.

3. Data Description

In this paper, the UCI public dataset for student performance in an entrance examination is considered. The data are obtained from the Common Entrance Examination (CEE) of Dibrugarh University for a given year at a medical school in the Indian state of Assam [28]. This dataset was obtained from observational studies in education research, and the relationship between multiple intervention variables and response variables is considered. The sample size of this dataset is small, but, because of the particularity of the individual growth environment, the balance between the treatment group and the control group is poor. The ATE can thus be calculated using the method based on sub-classification to obtain a more balanced value for the ATE.

The original dataset contains 12 variables and 666 samples, of which performance in the CEE is set as the response variable and the other variables are interventions, that is, the performance of students in the CEE is related to factors such as caste, gender, coaching, or father's occupation. The cause–effect relationship between variables is shown in Figure 3. This paper uses the above method to calculate the ATE value between students' entrance examination performance and other intervention variables and then analyzes the influencing factors according to the results. The basic dataset information is shown in Table 1.

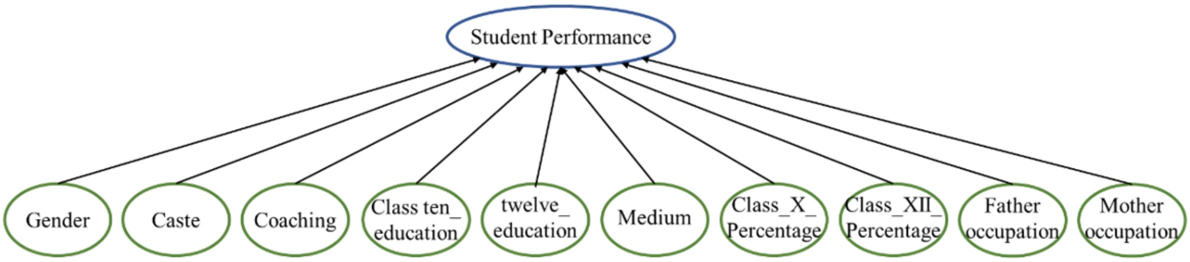


Figure 3. Relationship between response variable and intervention variables.

Table 1. Dataset information [28].

Variable	Description	Values
Performance	Performance in Common Entrance Examination (CEE)	{'Excellent', 'Vg', 'Good', 'Average'} If the percentage is top 100, Excellent If the percentage is next 200, Very Good (Vg) If the percentage is next 200, Good Remainder, Average Here, the percentage means the percentage score on the CEE Examination
Gender	Gender of the candidate	{'male', 'female'}
Caste	Caste of the candidate	{'General', 'OBC', 'SC', 'ST'} OBC—Other Backward Caste SC—Schedule Caste ST—Schedule Tribes
Coaching	Whether or not the candidate attended any coaching classes within Assam or outside Assam	{'NO', 'WA', 'OA'} No—No Coaching WA—Within Assam OA—Outside Assam
Time	The length of time students received coaching	{'ONE', 'TWO', 'THREE', 'FOUR', 'FIVE', 'SEVEN'}
Class ten_ education	Name of the board where the candidate studied at Class X level	{'SEBA', 'OTHER', 'CBSE'}

Table 1. *Cont.*

Variable	Description	Values
twelve_ education	Name of the board where the candidate studied at Class XII level	{'AHSEC', 'CBSE', 'OTHER'}
medium	Medium of instruction for the study at the Class XII level	{'ENGLISH', 'OTHER', 'ASSAMESE'}
Class_X_Percentage	The percentage secured by the candidate at Class X standard	{'Excellent', 'Vg', 'Good', 'Average'} If the percentage is above 80%, Excellent If 70% ≤ percentage < 80%, Very Good (Vg) If 60% ≤ percentage < 70%, Good The remainder, Average
Class_XII_Percentage	The percentage secured by the candidate at Class XII standard	{'Excellent', 'Vg', 'Good', 'Average'} If the percentage is above 80%, Excellent If 70% ≤ percentage < 80%, Very Good (Vg) If 60% ≤ percentage < 70%, Good The remainder, Average
Father occupation	The occupation of the father of the candidate	{'DOCTOR', 'SCHOOL_TEACHER', 'BUSINESS', 'COLLEGE_TEACHER', 'OTHER', 'BANK_OFFICIAL', 'ENGINEER', 'CULTIVATOR'}
Mother occupation	The occupation of the mother of the candidate	{'OTHER', 'HOUSE_WIFE', 'SCHOOL_ TEACHER', 'DOCTOR', 'COLLEGE_TEACHER', 'BANK_ OFFICIAL', 'BUSINESS', 'CULTIVATOR', 'ENGINEER'}

4. Calculation and Results

4.1. Average Treatment Effect Calculation

4.1.1. Division of Treatment and Control Groups

The ATE value is calculated between all interventions and the response variable using the method based on sub-classification estimation. The response variable, i.e., student performance, is divided into four categories. We encode "excellent" as 1, "very good" as 2, "good" as 3, and "average" as 4. These four values are used as the observed value for subsequent ATE calculations.

Some intervention variables are not binary, so they need to be preprocessed when dividing the control and treatment groups. For example, for the "medium" intervention variable, after classification processing, the medium was set to "English" for the treatment group and to "not English" for the control group. The division of intervention variables into the control and treatment groups is shown in Table 2.

Table 2. Grouping of intervention variables.

Intervention Variables	Treatment Group	Control Group
Gender	Male	Female
Caste	General, OBC	SC, ST
Coaching	WA, OA	NO
Time	THREE, FOUR, FIVE, SEVEN	ONE, TWO
Class_ten_education	SEBA	Other, CBSE
twelve_education	CBSE	AHSEC, Other

Table 2. *Cont.*

Intervention Variables	Treatment Group	Control Group
Medium	English	Other, Assamese
Class_X_Percentage	Excellent	Vg, Good, Average
Class_XII_Percentage	Excellent	Vg, Good, Average
Father's occupation	SCHOOL_TEACHER, COLLEGE_TEACHER	DOCTOR, BUSINESS, OTHER, BANK_OFFICIAL, ENGINEER, CULTIVATOR
Mother's occupation	HOUSE_WIFE	SCHOOL_TEACHER, DOCTOR, COLLEGE_TEACHER, BANK_OFFICIAL, BUSINESS, CULTIVATOR, ENGINEER, OTHER

Here, we calculate the ATE between intervention variable *coaching* and student entrance exam performance, as an example. The control and treatment groups are set as follows: individuals who *Attended Coaching* as treatment group; individuals with *No Coaching* as control group. The remaining intervention variables that could be included as covariates are Gender, Caste, Class_ten_education, Twelve_education, Medium, Class_X_Percentage, Class_XII_Percentage, Father's occupation, and Mother's occupation.

4.1.2. Select Covariates for Inclusion in Logistic Regression Models

According to the covariate selection method described in Section 2.1.3, the likelihood ratio statistic is used to select the covariates for the logistic regression model. Ten covariates are not introduced into the model at the beginning, and these covariates are included in the logistic regression model one by one with respect to the calculated likelihood ratio statistic. Taking the intervention *Coaching* as an example, we set the threshold $C_L = 1$ according to the stepwise regression, and then calculate the likelihood ratio statistic for each covariate and incorporate the covariate with the largest likelihood ratio statistic for each step into the logistic regression model. The results of the likelihood ratio statistic are shown in Table 3. In the first round, the covariate *Caste* with largest likelihood ratio statistic is included in the model. By repeating this process with respect to Table 3, the covariate with the largest likelihood ratio statistics (highlighted in yellow) is recursively added to the model in each step. From Table 3, one may observe that all the covariates are integrated in the model for PS calculation, as the likelihood ratio statistic is always above the predefined threshold.

Table 3. Results of the likelihood ratio statistic for each step.

Covariate	Step									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Gender	5.012	3.923	5.926	5.935	6.070	7.593	7.173	5.831	5.729	6.317
Caste	474.493									
Time	21.36	18.071	16.781	18.541	26.074					
Class_ten_education	12.164	11.521	7.322	7.542	6.564	7.592	7.048	16.441		
twelve_education	8.894	6.437	4.298	4.468	4.216	7.319	8.962			
Medium	31.220	20.703	20.075	19.472	18.338	18.760				
Class_X_Percentage	40.910	11.458	11.762	11.950	6.355	2.943	3.067	3.703	15.242	
Class_XII_Percentage	72.628	24.394	21.414	21.854						
Father's occupation	64.999	38.281								
Mother's occupation	46.597	37.237	32.471							

4.1.3. Calculate PS for Sample Data

Here, we set the treatment group (*Attended Coaching*) = 1 and the control group (*No Coaching*) = 0. The observation results are set to "*excellent*" = 1, "*very good*" = 2, "*good*" = 3, and "*average*" = 4. The PS of the sample is estimated using the SPSS multinomial logistic regression model—that is, the probability value obtained in that model.

4.1.4. Trimming

For this dataset, we exclude samples with a PS very close to 0 and very close to 1 to ensure linearized PSs that are not too extreme. In total, 649 samples of data are left, with 500 and 149 individuals in the treatment group and control group, respectively.

4.1.5. Stratification

The samples are stratified according to PS values and the final stratified results had to have at least 12 samples of data per group. The main steps of this process are shown as below.

(1) The estimated PS is linearized according to Equation (4).
(2) The *t*-statistic is used to test whether the linear PSs of the treatment and control groups are approximately equal. Before conducting the *t*-test, the F-statistic is constructed to test whether the variances in the treatment and control groups are comparable. In this case, equal variances are justified with an F-test within each level, so the variances of the two groups are considered equal in the calculation of the *t*-statistic.

The *t*-statistic calculation results of all samples are shown in Table 4. As the *t*-statistic exceeds 1 [29], it is divided into two layers at the median, and then the *t*-statistic is calculated separately for each layer.

Table 4. The *t*-statistic calculation results of all samples.

Step	Layer	Lower Bound	Upper Bound	Interval Width	Number in Control Group	Number in Treatment Group	*t*-Statistic
1	1	0.04	0.98	0.94	149	500	12.715

(3) The first stratification

The results of the first stratification are shown in Table 5. After the first stratification, the *t*-statistic of the first layer is greater than 1, thus it is divided into two layers at the median. Because there are only 30 samples left in the control group in the second layer, continuing stratification would make the number of samples in the group less than 12, so no further stratification is required.

Table 5. The results of the first stratification.

Step	Layer	Lower Bound	Upper Bound	Interval Width	Number in Control Group	Number in Treatment Group	*t*-Statistic
2	1	0.04	0.86	0.82	119	206	7.838
2	2	0.86	0.98	0.12	30	294	1.040

(4) The second stratification

The results of the second stratification are shown in Table 6. According to the stratification results, the *t*-statistics of the first and second layers are all greater than 1, but further stratification in the second layer would lead to less than 12 samples in the control group, so only the first layer would continue to be stratified.

Table 6. The results of the second stratification.

Step	Layer	Lower Bound	Upper Bound	Interval Width	Number in Control Group	Number in Treatment Group	*t*-Statistic
3	1	0.04	0.59	0.55	86	76	2.419
3	2	0.59	0.86	0.27	33	130	3.631
3	3	0.86	0.98	0.12	30	294	1.040

(5) The third stratification

The results of the third stratification are shown in Table 7. So far, the *t*-statistic of each layer is less than 1. Two of the *t*-statistics are still greater than 1, but, because of the minimum sample size requirement, the stratification could not be continued. Therefore, we considered that these four layers met the necessary conditions. The whole process of stratification in this case study can be seen from the Figure 4.

Table 7. The results of the third stratification.

Step	Layer	Lower Bound	Upper Bound	Interval Width	Number in Control Group	Number in Treatment Group	*t*-Statistic
4	1	0.04	0.49	0.45	29	52	0.949
4	2	0.49	0.59	0.1	34	47	0.054
4	3	0.59	0.86	0.27	33	130	3.631
4	4	0.86	0.98	0.12	30	294	1.040

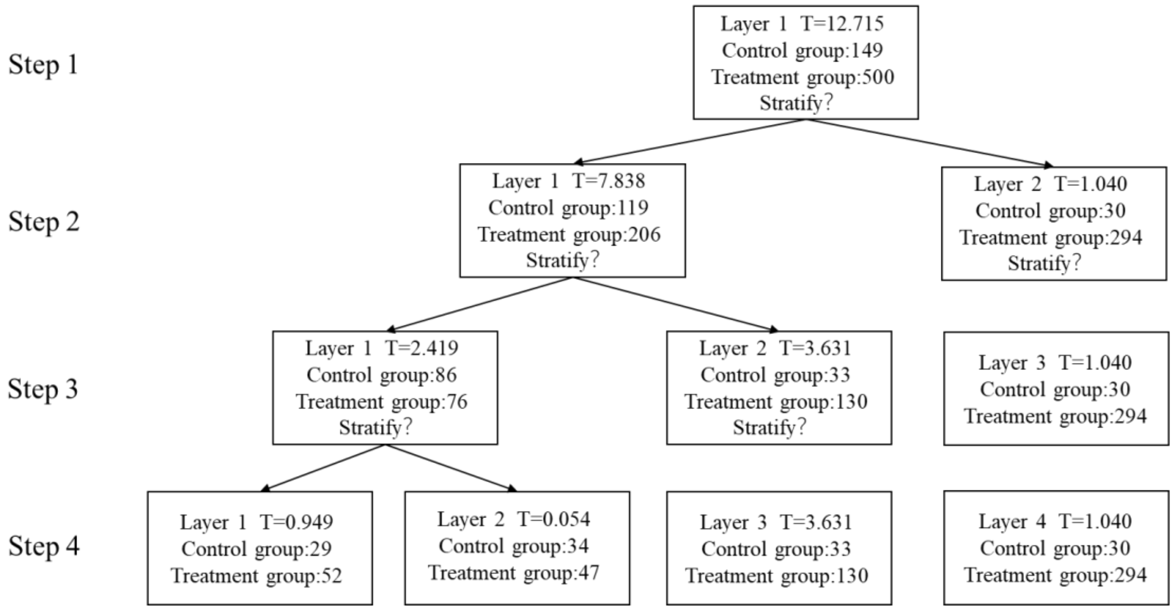


Figure 4. The process of stratification in this case study.

4.1.6. Calculation of Average Treatment Effect

When calculating the difference in the ATE, we set the treatment group (*Attended Coaching*) = 1, the control group (*No Coaching*) = 0, and the observation result as the student entrance exam performance. We substitute the value into Equation (9) and calculate the value of the ATE between coaching and student performance in each layer. The value of the ATE between coaching and student performance in each layer is shown in Table 8.

Table 8. The value of the ATE between coaching and student performance in each layer.

Layer	1	2	3	4
Value	−0.21	0.27	0.45	−0.6

Using Neyman inference based on sub-classification, referring to Equations (10) and (11), the average causal effect difference estimate of the sample population is obtained as follows:

$$\begin{aligned}\hat{\tau}^{\text{strat}} &= \left(-0.21 \times \frac{81}{649}\right) + \left(0.27 \times \frac{81}{649}\right) + \left(0.45 \times \frac{163}{649}\right) + \left(-0.6 \times \frac{324}{649}\right) \\ &= -0.18\end{aligned} \tag{12}$$

According to Equation (12), the ATE between coaching and student performance is −0.18.

4.2. Average Causal Effect between Response and Other Intervention Variables

Repeating the procedure in Section 4.1 for other interventions, we obtain the values of the ATE between each intervention and the response variable, as shown in Table 9.

Table 9. Average treatment effect values between intervention and response variables.

Intervention Variables	ATE Value
Gender	−0.13
Caste	−1.37
Coaching	−0.18
time	0.45
Class_ten_education	−0.47
twelve_education	−0.54
medium	0.41
Class_X_Percentage	0.15
Class_XII_Percentage	0.31
Father_occupation	0.14
Mother_occupation	0.0022

As the ATE value measures the magnitude of the causal relationship between variables in terms of absolute value, its absolute value can be graphed in the following Figure 5.

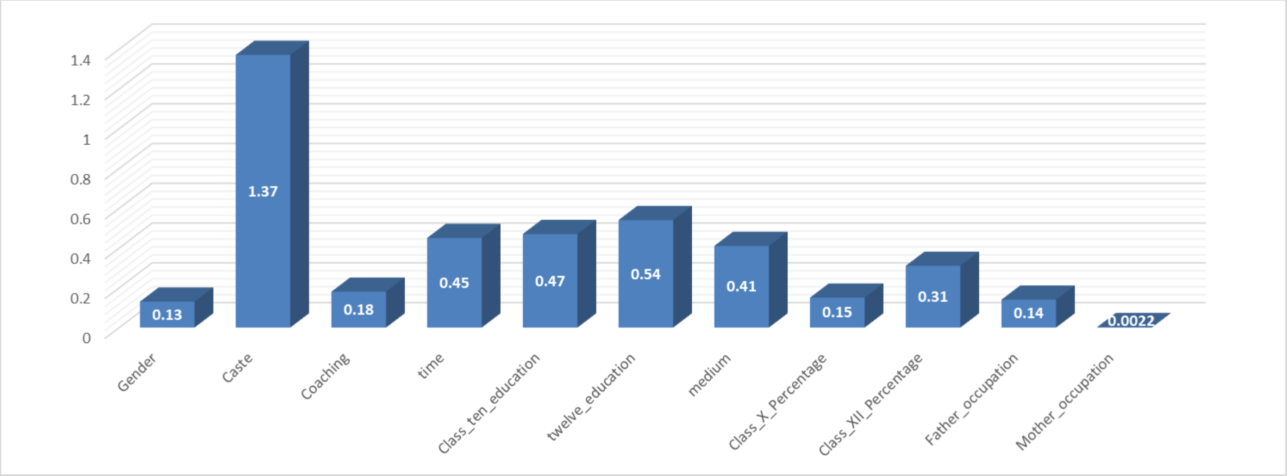


Figure 5. ATE value of the intervention variables.

4.3. Case Study Result Analysis

According to the above results, among the intervention variables, the ATE between caste and student performance is the largest. The intervention variables *time*, *Class_ten_education*, *Twelve_education*, *medium*, and *Class_XII_Percentage* have similar important effects on student performance; however, the variables of *father's occupation*, *mother's occupation*, *gender*, *coaching*, and *Class_X_Percentage* have less influence on student performance.

(1) The data are obtained from the Dibrugarh University CEE for a given year at a medical school in the Indian state of Assam. In India, *caste* determines the environment in which individuals are born, the environment and quality of education, as well as other factors. Therefore, the two races 'SC—Schedule Caste' and 'ST—Schedule Tribes' have lower social status than 'general'; thus, overall, they have a lower social status in India. There is, not surprisingly, also a large gap in performance on entrance exams based on *caste*.

(2) Several intervention variables such as *time*, *Class_Ten_education*, *Twelve_education*, and medium are relatively direct influencing factors in the learning process. From the perspective of education, these factors have a certain impact on student performance. This also indicates, however, that students who perform well in Class XII courses also perform well in the entrance examination.

(3) Factors such as *parental occupation* and *gender* appear to have little impact on student performance, which indicates that parents' education level, work environment, and other such factors have little effect on a student's learning ability or that these factors have an indirect influence rather than a direct one.

(4) To improve student exam performance, the most influential direct factor appears to be *caste*—that is, social class—but this factor is difficult to change. Therefore, in the field of education, teachers can try to increase the training time of students and choose the appropriate *Class_ten_education*, *Twelve_education*, and *medium of instruction*.

The above analysis only concerns this particular dataset, which has obvious social background characteristics; however, it can be seen from this case that, in observational studies with a small sample size, the use of the ATE estimation method based on subclassification is effective. This method can well estimate the value of the ATE between the intervention variable and the response variable. In the field of education, the value of the ATE can be combined with the social background to determine the factors that have a greater impact on students' learning performance among the intervention variables to clarify the necessary direction of improvement.

4.4. The Results of Traditional Methods

In order to compare to the method proposed in this paper, we introduce the commonly used correlation analysis method and traditional ATE calculation method to obtain the calculation results.

4.4.1. Correlation Analysis

Correlation analysis is often used in observational studies to judge the relationship of the covariates and the response variable. The software SPSS is used to calculate the correlation between each intervention and the response variable. The result is shown as follows in Table 10.

The results of the correlation analysis indicated that the three intervention variables, caste, Class_X_Percentage, and Class_XII_Percentage, are significantly correlated with the response variable at the 0.05 significance level, and time and Class_ten_education are significantly correlated with the response variable at the 0.01 significance level. Correlation analysis considers solely the correlation between the response variable and one of the intervention variables, which can be misled by the confounding variables. Such a treatment amounts to ignoring the relationship between their intervening variables, and it is untrustworthy in the analysis of observational studies in the education field.

Table 10. Correlation values between intervention and response variables.

Intervention Variables	Correlation Value
Gender	0.058
Caste	0.598 **
Coaching	0.031
time	−0.076 *
Class_ten_education	−0.096 *
twelve_education	0.067
medium	−0.006
Class_X_Percentage	0.204 **
Class_XII_Percentage	0.280 **
Father_occupation	0.031
Mother_occupation	−0.061

** represents significant correlation at the 0.01 level; * represents significant correlation at the 0.05 level.

4.4.2. ATE Calculation Method Based on DoWhy

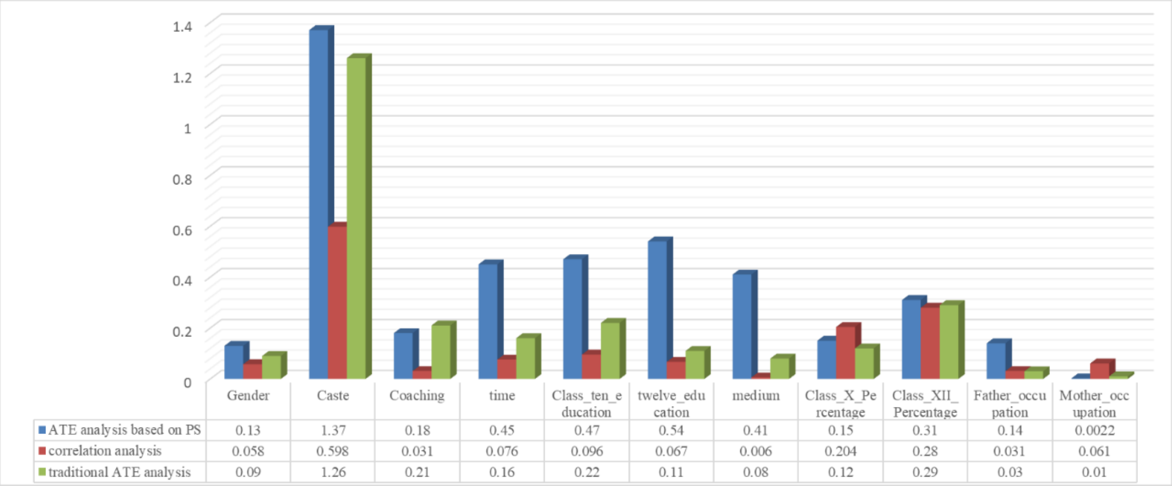
The treatment effect calculation method based on observed data is also considered as a benchmark method to calculate the ATE result of the intervention variables. The calculation can be divided into two steps, that is, identification and estimation. Identification represents the causal effects as a statistic, then ATE can be obtained from the model [30]. In this work, the causal model is shown in Figure 3. Code form [31] named DoWhy is adopted to calculate the ATE and the results are shown in Table 11.

Table 11. ATE values between intervention and response variables based on DoWhy.

Intervention Variables	ATE Value
Gender	0.09
Caste	1.26
Coaching	0.21
time	−0.16
Class_ten_education	−0.22
twelve_education	−0.11
medium	−0.08
Class_X_Percentage	−0.12
Class_XII_Percentage	0.29
Father_occupation	0.03
Mother_occupation	−0.01

The traditional ATE calculation method leads to the conclusion that caste has the greatest effect on student performance, followed by Class_XII_Percentage, Class_ten_education, Coaching, and time, with the remaining intervention variables having a limited effect. This result still differs from the ATE calculation based on PS proposed in this work. The main reason is that, in the process of the traditional ATE calculation, its causal effect is only treated as a statistic, and the confounding factors among its variables are not considered and treated. Therefore, there is still some error in the calculation of its causal effect.

The comparison of the three methods can be seen in the Figure 6. It can be seen from the bar chart that the results of the three methods are different in some ways. The method proposed in this work takes into account the relationship among the intervention variables, while treating the problem of confounding variables in the data. Compared with both the traditional correlation analysis and ATE calculation methods, it has higher reliability and is more adaptable to the analysis of observational study data in education.



	Gender	Caste	Coaching	time	Class_ten_education	twelve_education	medium	Class_X_Percentage	Class_XII_Percentage	Father_occupation	Mother_occupation
🟦 ATE analysis based on PS	0.13	1.37	0.18	0.45	0.47	0.54	0.41	0.15	0.31	0.14	0.0022
🟥 correlation analysis	0.058	0.598	0.031	0.076	0.096	0.067	0.006	0.204	0.28	0.031	0.061
🟩 traditional ATE analysis	0.09	1.26	0.21	0.16	0.22	0.11	0.08	0.12	0.29	0.03	0.01

Figure 6. Comparison of the three methods.

5. Conclusions

Considering the unbalanced distributions of control and treatment groups in an observational education study, this paper proposes a method for calculating the ATE based on observational study samples—that is, a sub-classification method based on PS analysis. The proposed method can correct the biased intervention effect estimation results. Compared with the correlation analysis method and the ATE calculation method based on DoWhy in observational studies, the method proposed in this paper takes into account both the unobservable confounding factors in the treatment variables and the causal relationships between the response variable and intervention variables. In addition, the distribution bias between the treatment and control groups is balanced by stratification with respect to PS. In summary, this method is more credible than the traditional methods in observational studies with biased distribution and unobservable confounding factors.

The UCI public dataset for student performance in an entrance examination in this work is used to test the credibility of the proposed method in observational education study. The data are divided into subgroups through trimming and stratification methods with respect to the PS. The ATE of each layer of samples is calculated using the Neyman estimation method. Then, the ATE on the response variable in the whole dataset is derived. This value may reflect the effect of educational intervention policies, indicating that this method could be suitable for application in the formulation of intervention policies in education research.

Author Contributions: Conceptualization, J.L. (Jingyu Liang) and J.L. (Jie Liu); methodology, J.L. (Jingyu Liang); software, J.L. (Jingyu Liang); validation, J.L. (Jingyu Liang); formal analysis, J.L. (Jingyu Liang); investigation, J.L. (Jie Liu); resources, J.L. (Jie Liu); data curation, J.L. (Jingyu Liang); writing—original draft preparation, J.L. (Jingyu Liang); writing—review and editing, J.L. (Jie Liu); visualization, J.L. (Jingyu Liang); supervision, J.L. (Jie Liu); project administration, J.L. (Jie Liu); funding acquisition, J.L. (Jie Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the dataset used in this paper are from UCI Machine Learning Repository, a publicly available dataset. This dataset contains data of the candidates who qualified the medical entrance examination for admission to medical colleges of Assam of a particular year and collected by Jiten Hazarika. There are no ethical issues involved.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Springer at https://doi.org/10.1007/978-3-319-91192-2_21, reference number [28].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Clearinghouse, W.W. *Standards Handbook*; Version 4.0; Institute of Education Sciences: Washington, DC, USA, 2017.
2. Weiss, M.J.; Bloom, H.S.; Savitz, N.V.; Gupta, H.; Vigil, A.E.; Cullinan, D.N.; Cullinan, D. How Much Do the Effects of Education and Training Programs Vary Across Sites? Evidence From Past Multisite Randomized Trials. *J. Res. Educ. Eff.* **2017**, *10*, 843–876. [CrossRef]
3. Tipton, E.; Olsen, R.B. A Review of Statistical Methods for Generalizing from Evaluations of Educational Interventions. *Educ. Res.* **2018**, *47*, 516–524. [CrossRef]
4. Bosdriesz, J.R.; Stel, V.S.; Van Diepen, M.; Meuleman, Y.; Dekker, F.W.; Zoccali, C.; Jager, K.J. Evidence-based medicine—When observational studies are better than randomized controlled trials. *Nephrology* **2020**, *25*, 737–743. [CrossRef] [PubMed]
5. Wong, V.C.; Valentine, J.; Miller-Bains, K. Empirical Performance of Covariates in Education Observational Studies. *J. Res. Educ. Eff.* **2017**, *10*, 207–236. [CrossRef]
6. Cook, C.; Engelhard, C.; Landry, M.D.; McCallum, C. Modifiable variables in physical therapy education programs associated with first-time and three-year National Physical Therapy Examination pass rates in the United States. *J. Educ. Eval. Health Prof.* **2015**, *12*, 44. [CrossRef]
7. Titus, M.A. Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Res. High. Educ.* **2007**, *48*, 487–521. [CrossRef]
8. Chiteng Kot, F. The Impact of Centralized Advising on First-Year Academic Performance and Second-Year Enrollment Behavior. *Res. High. Educ.* **2014**, *55*, 527–563. [CrossRef]
9. Rubin, D.B.; Thomas, N. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **1992**, *79*, 797–809. [CrossRef]
10. Rubin, D.B.; Thomas, N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* **1996**, *52*, 249. [CrossRef]
11. Lunceford, J.K.; Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **2004**, *23*, 2937–2960. [CrossRef]
12. Myers, J.A.; Louis, T.A. *Optimal Propensity Score Stratification*; Johns Hopkins University, Department of Biostatistics: Baltimore, MD, USA, 2007.
13. Rosenbaum, P.R.; Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **1984**, *79*, 516–524. [CrossRef]
14. Hirano, K.; Imbens, G.W. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Serv. Outcomes Res. Methodol.* **2001**, *2*, 259–278. [CrossRef]
15. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
16. Turk, J.M. Estimating the Impact of Developmental Education on Associate Degree Completion: A Dose–Response Approach. *Res. High. Educ.* **2019**, *60*, 1090–1112. [CrossRef]
17. Vaughan, A.L.; Lalonde, T.L.; Jenkins-Guarnieri, M.A. Assessing Student Achievement in Large-Scale Educational Programs Using Hierarchical Propensity Scores. *Res. High. Educ.* **2014**, *55*, 564–580. [CrossRef]
18. Powell, M.G.; Hull, D.M.; Beaujean, A.A. Propensity Score Matching for Education Data: Worked Examples. *J. Exp. Educ.* **2020**, *88*, 145–164. [CrossRef]
19. Masserini, L.; Bini, M. Does joining social media groups help to reduce students' dropout within the first university year? *Socio-Econ. Plan. Sci.* **2021**, *73*, 100865. [CrossRef]
20. Chen, J.; Keller, B. Heterogeneous Subgroup Identification in Observational Studies. *J. Res. Educ. Eff.* **2019**, *12*, 578–596. [CrossRef]
21. Suk, Y.; Kang, H.; Kim, J.-S. Random Forests Approach for Causal Inference with Clustered Observational Data. *Multivar. Behav. Res.* **2021**, *56*, 829–852. [CrossRef]
22. Neyman, J.; Iwaszkiewicz, K. Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.* **1935**, *2*, 107–180. [CrossRef]
23. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **1974**, *66*, 688–701. [CrossRef]
24. Holland, P.W. Statistics and causal inference. *J. Am. Stat. Assoc.* **1986**, *81*, 945–960. [CrossRef]
25. Kaplan, D. Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assess. Educ.* **2016**, *4*, 7. [CrossRef] [PubMed]
26. Morgan, S.L.; Winship, C. *Counterfactuals and Causal Inference*; Cambridge University Press: Cambridge, MA, USA, 2015.
27. Yang, S.; Ding, P. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **2018**, *105*, 487–493. [CrossRef]

28. Hussain, S.; Atallah, R.; Kamsin, A.; Hazarika, J. Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study. In Proceedings of the Computer Science On-line Conference 2018, Vsetin, Czech Republic, 25–28 April 2018; pp. 196–211. [CrossRef]
29. Imbens, G.W.; Rubin, D.B. *Causal Inference in Statistics, Social, and Biomedical Sciences*; Cambridge University Press: Cambridge, MA, USA, 2015.
30. Neal, B. Introduction to Causal Inference from a Machine Learning Perspective; Course Lecture Notes (Draft). 2020. Available online: https://scholar.google.co.jp/scholar?hl=zh-TW&as_sdt=0%2C5&q=Introduction+to+Causal+Inference+from+a+Machine+Learning+Perspective&btnG=#d=gs_cit&t=1668758477310&u=%2Fscholar%3Fq%3Dinfo%3ATFJuQPUjj00J%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Dzh-TW (accessed on 13 November 2022).
31. Sharma, A.; Kiciman, E. DoWhy: An end-to-end library for causal inference. *arXiv* **2020**, arXiv:2011.04216.