Assessing the Treatment Effect on the Causal Models via Parametric Approaches with Applications to the Study of English Educational Effect in Japan

Emura, Takeshi and Katsuyama, Hitomi and Wang, Jinfang

Graduate Institute of Statistics, National Central University, Taiwan, Department of English, Kawamura Gakuen Women's University, Japan, Department of Mathematics and Informatics, Graduate School of Science, Chiba University, Japan

27 April 2010

Assessing the Treatment Effect on the Causal Models via Parametric Approaches with Applications to the Study of English Educational Effect in Japan

Takeshi Emura;

emura@stat.ncu.edu.tw

Graduate Institute of Statistics, National Central University, Taiwan

Hitomi Katsuyama;

Department of English, Kawamura Gakuen Woman's University, Japan

Jinfang Wang;

Department of Mathematics and Informatics, Graduate School of Science,

Chiba University, Japan

Abstract: Observational studies are widely used to evaluate the effect of treatment when it is not feasible to conduct controlled experiment. This article considers the use of parametric analyses for estimating the causal treatment effect. The proposed approach is an alternative to the widely used stratification estimator as well as Robins' double robust estimator both of which are consistent under the key assumption of strong ignorability. To relax the assumption of strong ignorability, we instead impose fully parametric structures on the causal models to identify the causal treatment effect. The proposed parametric framework provides a likelihood ratio test for checking the assumption of strong ignorability. Simulations are conducted to investigate the performance of the proposed estimator as well as the power of the likelihood ratio test. We demonstrate how the proposed method can be used for data from an observational study for measuring English educational effect on Japanese elementary school students.

KEYWORDS: Counterfactual model of causality; Independence test; Likelihood ratio test; Missing data; Model checking; Propensity score

1 Introduction

The Japanese government has not yet decided to include English as a mandatory subject at elementary schools, unlike Taiwan, where it has been a mandatory subject from the 5th grade since 2001 and South Korea, where it has been a mandatory subject from the 3rd grade since 1997. In March 2006, the Central Council of Education in Japan suggested that English education should start from the 5th grade as a compulsory subject, which may start in 2010 at the earliest. The other reason why the Japanese government has not yet decided to make English as part of the formal curriculum at the elementary school level is that there are many people including some authorities who wonder if English education at elementary schools is an effective use of valuable classroom time (Otsu, 2004; 2005). Our research is motivated by the question for the effectiveness of English education in Japanese elementary schools and the dataset from this study plays an important role in explaining the proposed statistical methodologies in this article.

Observational studies are often used to study the effect of a treatment or policy when it is not feasible to use controlled experiment. Since the original work of Cochran (1965), substantial research efforts have been devoted to developing methodologies suited for observational studies. Rosenbaum (2002) summarizes the concepts of observational studies with a formal mathematical framework, illustrated with plenty of real examples. Applications of observational studies arise in biology, econometrics, education and sociology to name but a few. Dehejia and Wahba (1999) studied the effect of a labor training program in the United States. They compared the after-intervention income of the program participants with that of non-participants using a database. In a series of papers of Coleman, Hoffer and Kilgore (1982), Goldberger and Cain (1982) and Morgan (2001), the effect of attending Catholic schools, compared to private and public schools, has been studied under non-experimental settings. In Japan, the efficacy of early introduction of English training programs at elementary schools is of substantial interest to both the general public and the government (Katsuyama, Nishigaki and Wang 2006, 2008). In most observational studies, the typical focus of interest is the effect of treatment, such as the efficacies of

training programs or medical treatments.

A modern approach to investigating the treatment effect in observational studies is based on the counterfactual models of causality. Although some model assumptions are necessary, the counterfactual models do provide a systematic way to define the treatment effect. In particular, the successful development in both theory and application on propensity score (Rosenbaum and Rubin 1983) increased the value of the counterfactual modeling. Rosenbaum and Rubin (1983) reviews how the point estimates for the treatment effect can be constructed based on stratification and matching on propensity scores. In recent years, the counterfactual models have been adopted by many researchers who aim to analyze problems in various fields; see Heckman, Ichimura, Smith and Todd (1998) for econometric applications, Winship and Morgan (1999) for sociological applications and Pearl (2001) for applications to the health sciences. All these methods based on the counterfactual models heavily depend on an identifiability assumption called the strongly ignorable treatment assignment.

It is not unusual in the counterfactual models that some assumptions are not statistically testable by data. As we will discuss in Section 2.3, the fundamental assumption of strong ignorability is not statistically testable from data in a nonparametric way. To obtain a consistent estimator of the treatment effect, most existing methods are built on this untestable assumption and many practitionars use their subject matter knowledge on the validity of the assumption. In the problem of estimating the English educational effect, counterfactual models do provide a model to assess the causal treatment effect of interest. However, the assumption of strong ignorability may be questionable to many people who analyse data or who interpret the resulting educational effect. So far, the majority of case studies uses the sensitivity analysis (Rosenbaum 2002) to assess the robustness of estimate against the violation of the strong ignorability.

This article presents a different way to assess the validity of existing estimates and the assumption of strong ignorability. Specifically, we impose fully parametric models on the counterfactual models and then obtain the estimates of the treatment effect without the assumption of strong ignorability. Raughly speaking, if the parametric estimate is close to

the existing estimates, then the existing estimates are shown to be a reliable estimate of the treatment effect. In addition, the current approach easily incorporates a statistical test of the assumption of strong ignorability. If the test does not reject the assmption of strong ignorability, then the existing estimates can be considered reliable. On the other hand, if the parametric estimate departs from the existing estimates, or if the assumption of strong ignorability is rejected, then the validity of the existing estimates could be questionable. We believe that not many practitionars have appropriate konwledge for interpreting the sensitivity analysis. Instead of presenting the sensitivity analysis, one can present the parametric estimate and the p-value of the strong ignorability test to suppliment the existing estimates.

The paper is organized as follows. Section 2 introduces the background for the counterfactual models and review some existing inference procedures. Also, it is shown that the assumption of strong ignorability is not identifiable nonparametrically. Section 3 introduces the proposed parametric methods for the counterfactual models, which do not requires the assumption of strong ignorability. Section 4 reports on simulation results that investigate the performance of the proposed approaches. Section 5 presents data analysis for the English educational effect based on a survey research. Section 6 concludes the article.

2 Background and motivation

2.1 The counterfactual models of causality

The counterfactual models of causality assume that each unit has two potential outcomes. Let $Y_i(0)$ denote the response for unit i when he/she were assigned to a control group, and $Y_i(1)$ denote the response for unit i when he/she were assigned to a treatment group. We call the pair, $(Y_i(0), Y_i(1))$, potential outcomes since only one of them is observed and the other is a hypothetical latent variable for each i. A component $Y_i(t)$ ($t = 0$ or 1) of the potential outcome $(Y_i(0), Y_i(1))$ is observed if and only if the unit is assigned to the treatment $T_i = t$. Thus, the observed response variable can be expressed

as $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. Suppose that, for each unit, a p dimensional covariate vector $\boldsymbol{X}_i$ is measured prior to the treatment assignment. Observations consist of $(Y_i, T_i, \boldsymbol{X}_i)$; $i = 1, \ldots n$, independent and identical replica from the population distribution of $(Y(0), Y(1), T, \boldsymbol{X})$. The average treatment effect describes the expected gain due to the treatment and it is defined as

$$\tau = E\{Y(1)\} - E\{Y(0)\}, \tag{1}$$

where $E(\cdot)$ denotes expectation in the population. A covariate vector $\boldsymbol{X}$ is a potential confounder that may relate to the distributions of both $(Y(0), Y(1))$ and T.

2.2 Statistical inference on the counterfactual models

A majority of statistical methods developed on the counterfactual models of causality relies on some assumption on the treatment assignment process. To construct an unbiased estimator for the average treatment effect, Rosenbaum and Rubin (1983) imposed the following assumption:

Definition I *Treatment assignment is strongly ignorable given the observed covariate vector* $\boldsymbol{X}$ *if*

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid \boldsymbol{X} = \boldsymbol{x} \quad \textit{for all} \quad \boldsymbol{x}$$

and $0 < \Pr(T = 1|\boldsymbol{X} = \boldsymbol{x}) < 1$ *for all* $\boldsymbol{x}$.

Here, the notation $A \perp\!\!\!\perp B \mid C$ represents independence between variables A and B given a variable C (Dawid 1979). This condition states that, for those having $\boldsymbol{X} = \boldsymbol{x}$, the assignment rule is determined by an independent Bernoulli random variable having a probability of success $\Pr(T = 1|\boldsymbol{X} = \boldsymbol{x})$, as in a randomized experiment applied to the stratum $\boldsymbol{X} = \boldsymbol{x}$.

We extract the core statement from Definition I:

$$H_0^{Ign} : (Y(0), Y(1)) \perp\!\!\!\perp T \mid \boldsymbol{X}. \tag{2}$$

Also, a major interest in observational studies is to test no treatment effect

$$H_0^{Trt} : \tau = 0.$$

Under H_0^{Ign} it is easy to show that

$$\tau = E\{E(Y(1)|\boldsymbol{X}, T = 1) - E(Y(0)|\boldsymbol{X}, T = 0)\}.$$

Assuming that the covariate vector $\boldsymbol{X}$ is discrete, the above equation allows one to obtain a stratification estimator for τ:

$$\hat{\tau} = \sum_x \frac{n_x}{n} \left\{ \frac{\sum_i Y_i T_i \mathrm{I}(\boldsymbol{X}_i = \boldsymbol{x})}{\sum_i T_i \mathrm{I}(\boldsymbol{X}_i = \boldsymbol{x})} - \frac{\sum_i Y_i (1 - T_i) \mathrm{I}(\boldsymbol{X}_i = \boldsymbol{x})}{\sum_i (1 - T_i) \mathrm{I}(\boldsymbol{X}_i = \boldsymbol{x})} \right\}, \tag{3}$$

where n_x is the number of units in the stratum $\{j : \boldsymbol{X}_j = \boldsymbol{x}\}$ and $\mathrm{I}(\cdot)$ is the indicator function.

In studying the problem of evaluation, Heckman, Ichimura, Smith and Todd (1998) considers the effect of treatment on the treated, in our notation

$$\Delta(\boldsymbol{x}) = E(Y(1)|\boldsymbol{X} = \boldsymbol{x}, T = 1) - E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 1).$$

For this parameter to be interpretable they require that conditional distribution of $\boldsymbol{X}$ satisfies

$$\Pr(\boldsymbol{X} \leq \boldsymbol{x}|Y(0), Y(1), T) = \Pr(\boldsymbol{X} \leq \boldsymbol{x}|Y(0), Y(1)),$$

that is, conditional on the potential outcomes, realized T does not predict $\boldsymbol{X}$. This condition is closely related to the strong ignorability assumption. When the effect of treatment on the treated is of interest, a fundamental problem is on estimation of $E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 1)$, about which direct information is lacking. When $E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 0) = E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 1)$ holds, one can utilize data on the control group to estimate $E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 1)$. Otherwise selection bias

$$B(X) = E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 1) - E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, T = 0)$$

arises. This bias will vanish however when, for all $\boldsymbol{x}$,

$$Y(0) \perp\!\!\!\perp T|\boldsymbol{X} = \boldsymbol{x}$$

which is a condition weaker than the strong ignorability defined in Definition I. For details see Heckman, Ichimura, Smith and Todd (1998).

If $\boldsymbol{X}$ is of high dimension or contains continuous measurements, each of the n units may have a different value of $\boldsymbol{X}$, so no stratum can contain a treated and control unit with the same $\boldsymbol{X} = \boldsymbol{x}$. The application of the propensity score method is the standard way of overcoming this difficulty. Rosenbaum and Rubin (1983) showed that if the assignment is strongly ignorable given $\boldsymbol{X}$, it is also strongly ignorable given the propensity score $P(\boldsymbol{X})$ defined by $P(\boldsymbol{x}) = \Pr(T = 1|\boldsymbol{X} = \boldsymbol{x})$. Stratification method can then be implemented on $P(\boldsymbol{X})$ rather than on $\boldsymbol{X}$. In practice, $P(\boldsymbol{X})$ is replaced by the estimates $\hat{P}(\boldsymbol{X})$ obtained from a regression model, such as the logistic regression. Dehejia and Wahba (1999) gives case studies for implementing the propensity score method.

Another approach is the inverse probability weighting (IPW) method of Robins et al. (1994), which leads to

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_i \left\{ \frac{\sum_i Y_i T_i - \{T_i - \hat{P}(\boldsymbol{X}_i)\}\hat{m}_{1i}}{\hat{P}(\boldsymbol{X}_i)} - \frac{\sum_i Y_i(1-T_i) + \{T_i - \hat{P}(\boldsymbol{X}_i)\}\hat{m}_{0i}}{1-\hat{P}(\boldsymbol{X}_i)} \right\}, \quad (4)$$

where $\hat{m}_{ki} = m_k(\hat{\boldsymbol{\alpha}}_k; \boldsymbol{X}_i)$, $m_{ki}(\boldsymbol{\alpha}_k; \boldsymbol{X}_i) = E\{Y_i(k)|T_i = k, \boldsymbol{X}_i\}$ is the mean-regression structure and $\hat{\boldsymbol{\alpha}}_k$ is the regression estimator based on subject $\{j : T_j = k\}$. According to the theorem of Robins et al., $\hat{\tau}_{IPW}$ has the smallest asymptotic variance if both the mean-regression structures and the propensity score model are correctly specified. Furthermore, the 'double-robustness' property ensures that $\hat{\tau}_{IPW}$ is consistent if either (i) the two mean-regression structures are correctly specified but the propensity score model is not, or (ii) the propensity score model is correctly specified but the mean structures are not. All methodologies mentioned in this subsection rely on H_0^{Ign} or its similar versions.

2.3 The non-identifiability of H_0^{Ign}

Although testing the assumption of strong ignorability is a problem of considerable importance, there are few literature on systematic treatment of the assumption. Unfortunately, it turns out that H_0^{Ign} is not identifiable nonparametrically. To illustrate the non-identifiability, we consider two different probability models on $(Y(0), Y(1), T, \boldsymbol{X})$. In

the following two models, let $Y(0)$ be an arbitrary random variable having a density function $f(y(0)|\boldsymbol{X})$ with respect to some dominating measure, given the covariate vector $\boldsymbol{X}$.

Model A *Given the covariate vector* $\boldsymbol{X}$,

$$T = \begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2 \end{cases}, \quad Y(1) = \begin{cases} Y(0) + \tau & \text{if } T = 1 \\ Y(0) & \text{if } T = 0 \end{cases}.$$

Model B *Given the covariate vector* $\boldsymbol{X}$,

$$Y(1) = Y(0) + \tau, \quad T = \begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2 \end{cases}, \quad Y(0) \perp\!\!\!\perp T.$$

Here, in Model A and B, τ *is an arbitrary, but fixed value.*

In Model A, units enjoy the causal treatment effect τ if they are actually assigned to the treatment group while those who are assigned to the control group have no causal treatment effect. Thus, the assignment process depends on the efficacy of the causal quantity. In Model B, on the other hand, the assignment T is determined by an independent flip of coin, irrespective of the response value $(Y(0), Y(1))$, and the model is strongly ignorable given $\boldsymbol{X}$. It is easy to see that the two different models on $(Y(0), Y(1), T, \boldsymbol{X})$ yield the same density function for $(Y, T|\boldsymbol{X})$:

$$\frac{1}{2} f(Y - \tau T|\boldsymbol{X}).$$

Thus, one can not distinguish the strongly ignorable treatment assignment model, Model B, from Model A, based on the observed data $(Y_i, T_i, \boldsymbol{X}_i);\ i = 1, \ldots, n$. The identifiability dilemma stated above explains why it is difficult to assess the assumption of strong ignorability based on observed data.

A conventional way to alliviate this dilemma is based on the sensitivity analysis (Rosenbaum, 2003). This paper, however, presents an alternative way to the widely-used sensitivity analysis. The present approach utilizes a fully parametric model on the causal models to assess the assumptions H_0^{Trt} without assuming H_0^{Ign}. Also, the parametric approach allows one to test H_0^{Ign} using the likelihood ratio test.

3 Parametric inference on counterfactual models

3.1 Likelihood construction

The nonidentifiability aspect of τ arises partly because we never observe $Y_i(0)$ and $Y_i(1)$ jointly. The following assumption is fundamental in the proposed approach:

Assumption I (Constant Effect, Holland 1986):

$$Y(1) - Y(0) = \tau.$$

Assumption I states that the treatment uniformly changes the response $Y(0)$ by a magnitude τ. Assumption I is perhaps the simplest model for causal inference, but it is obviously a strong assumption. It is interesting to point out that Assumpton I implies $Var\{Y(0)\} = Var\{Y(1)\}$, which is commonly assumed in the analysis of variance for comparing two groups. There appears several methods for checking the validity of Assumption I, including a method mentioned in Holland (1986). We will develop an empirical method for checking Assumption I based on real data in Section 5.2 and the plausibility of Assumption I is further discussed in Section 6.

Let g be a strictly increasing and twice differentiable function. The distribution on $Y_i(0)$ may be parameterized as

$$Y_i(0) = g(\alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i) + \epsilon_i, \tag{5}$$

where ϵ_i follows a zero-mean distribution with a density $f_\epsilon(\cdot|\psi)$, where $\psi > 0$ is an unknown parameter. We further assume that ϵ_i is independent of $\boldsymbol{X}_i$ so that $E(\epsilon_i|\boldsymbol{X}_i) = 0$ for all i. An example includes $f_\epsilon(x|\psi) = e^{-x^2/(2\psi)}/\sqrt{2\pi\psi}$. Under Assumption I, the bivariate random variables $(Y(0), Y(1))$ is degenerated to a single random variable. Therefore, to model the non-ignorable treatment assigment process, it is convenient to assume

$$\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0)) = h\{\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i(0)\} \equiv P_i(Y_i(0)),$$

where h is a strictly increasing and twice differentiable function. Let

$$Q_i(Y_i(0)) \equiv f_\epsilon\{Y_i(0) - g(\alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i)|\psi\}.$$

Under Assumption I, the likelihood for $\{(Y_i, T_i) : i = 1, \ldots, n\}$ can be written as

$$\prod_i L_i(\theta) = \prod_i Q_i(Y_i - \tau T_i) P_i(Y_i - \tau T_i)^{T_i} \{1 - P_i(Y_i - \tau T_i)\}^{1-T_i}. \tag{6}$$

This is maximized for $\theta = (\tau, \alpha_0, \boldsymbol{\alpha}', \psi, \beta_0, \boldsymbol{\beta}', \gamma)' \in \Theta \subset \boldsymbol{R}^{p+2} \otimes (0, \infty) \otimes \boldsymbol{R}^{p+2}$. Let $\hat{\tau}_1$ be the MLE based on (6) and $\hat{\tau}_0$ be the MLE based on (6) under $\gamma = 0$. Let

$$i_n(\theta) = -\frac{1}{n} \sum_i \frac{\partial^2}{\partial\theta\partial\theta'} \log L_i(\theta).$$

We assume the following three conditions:

(A) $i_n(\theta)$ exists and continuous in a compact set Θ,

(B) $i_n(\theta)$ is positive definite with probability going to one as $n \to \infty$, and

(C) The convergence in probability of $i_n(\theta)$ to $-E\{\partial^2 \log L_i(\theta)/\partial\theta\partial\theta'\}$ is uniform in Θ.

The consistency of the maximum likelihood estimator follows directly from Theorem 2 of Foutz (1977).

Theorem I: *Suppose that (A), (B) and (C) hold. If Assumption I holds and functional forms of $f_\epsilon(x|\psi)$, g and h are correctly specified, then the assumption H_0^{Ign} is equivalently written as $H_0^{Ign} : \gamma = 0$. If the assumption of H_0^{Ign} holds in the underlying probability model, the two estimators $\hat{\tau}_1$ and $\hat{\tau}_0$ are consistent for the true treatment effect τ.*

Theorem I implies that, as $n \to \infty$, the two point estimates $\hat{\tau}_1$ and $\hat{\tau}_0$ get close to each other if the assumption of strong ignorability holds in the population. Notice that $\hat{\tau}$ and $\hat{\tau}_{IPW}$ are also consistent under the assumption of strong ingrability. Comparison of the four estimates $\hat{\tau}$, $\hat{\tau}_{IPW}$, $\hat{\tau}_1$ and $\hat{\tau}_0$ provides a method for checking the validity of the assumption of strong ignorability and therefore the validity of $\hat{\tau}$. Later we will compare the four estimators using the English test score data. Conditions (A), (B) and (C) may be chekced when $f_\epsilon(\cdot|\psi)$, g, h and the density of $\boldsymbol{X}_i$ are given. We will examine the validity of (A), (B) and (C) under a normal distribution model. The consistency of the two estimators in Theorem I is also studied by simulations.

For its popularity in applications and mathematical tractability, we recommend choosing $h(x) = e^x/(1+e^x)$ where the model on T becomes

$$\text{logit}\{\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0))\} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i(0). \tag{7}$$

3.2 An example for normal distribution model

This subsection provides an example of the proposed likelihood analysis when the response $Y(0)$ follows a normal distribution. Setting $f_\epsilon(x|\psi) = e^{-x^2/(2\psi)}/\sqrt{2\pi\psi}$, $\psi = \sigma^2$ and $g(x) = x$, the logarithm of (6) can be written as

$$\begin{aligned}\sum_i \log L_i(\theta) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i)^2 \\ &+ \sum_i \big[T_i\{\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma(Y_i - \tau T_i)\} - \log\{1 + \exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma(Y_i - \tau T_i))\}\big].\end{aligned} \tag{8}$$

The maximization for (8) can be done easily using a Newton-type algorithm. For instance, a nonlinear miminization command "nlm" implemented in R can be applied to the minus of (8). It also provides the Hessian matrix at the solution, which equals the observed Fisher information matrix. The maximization under $H_0^{Ign} : \gamma = 0$ can also be obtained similarly.

For Theorem I to be applied, one need to check conditions (A), (B) and (C). Condition (A) holds since (8) is continuously twice differentiable as can be seen from the explicit form of the observed Fisher information matrix provided in Appendix. Condition (B) is not easy to evaluate analytically. Nevertheless, our simulation results support the validity of (B) in that the observed Fisher information matrix is always positive definite. More details can be found in Section 4. To check (C), one need to show that the law of large number holds unifromly in the parameter space. Note that $\theta \to \partial^2 \log L_i(\theta)/\partial\theta\partial\theta'$ is continuous in Θ and thus bounded. This implies that the class of functions $\{\partial^2 \log L_i(\theta)/\partial\theta\partial\theta' : \theta \in \Theta\}$ is "Glivenko-Cantelli" (page 272 of Van der Vaart, 1998). The uniform version of the low of large number follows from the Glivenko-Cantelli's theorem.

If the assumed normal model is incorrect, the likelihood may not provide a consistent estimator for the treatment effect. For the results to be believable, the adequecy for the

forms $f_\epsilon(\cdot|\psi)$, g and Assumption I must be checked in details. For normal distribution models, some model diagnostics procedure to assess these assumptions are available from standard regression procedures. However, since the proposed method of model diagnostics is somewhat ad-hoc, it should be supplemented by subject matter knowledge on the distributional form. Such methods are illustrated in Section 5.2 through real data analysis.

3.3 Likelihood ratio test

Let $\hat{l}_0$ and $\hat{l}_1$ be the maximized log-likelihood of (6) under $H_0^{Ign} : \gamma = 0$ and without $H_0^{Ign} : \gamma = 0$ respectively. The likelihood ratio statistic,

$$2(\hat{l}_1 - \hat{l}_0), \tag{9}$$

can be used as a statistic for testing H_0^{Ign}. Properties of the proposed test directly follow from the general theory of the likelihood ratio test. The likelihood ratio statistic (9) has an approximate chi-squared distribution with one degree of freedom. The likelihood ratio test is asymptotically optimal for testing H_0^{Ign} when the one-sided alternative hypothesis $H_1 : \gamma > 0$ or $H_1 : \gamma < 0$ is specified (Van Der Vaart 1998). A test based on a confidence interval for γ is asymptotically equivalent to the likelihood ratio test. However, it is often desirable to consider a test based on the likelihood ratio rather than on a confidence interval for γ for accuracy in small sample sizes (McCullagh and Nelder 1989, p.471). It is still helpful to use the value of $\hat{\gamma}$ to see how $H_0^{Ign} : \gamma = 0$ is violated.

4 Simulation results

We conducted simulation studies to examine the performance of the proposed methods in finite samples. The covariate X_1 is a Bernoulli random variable with $\Pr(X_1 = 1) = \Pr(X_1 = 0) = 1/2$, and X_2 is uniformly distributed on (0,1). These variables consist of independent covariate vectors $\boldsymbol{X}' = (X_1, X_2)$. The counterfactual response $Y(0)$ were generated from the normal distribution with mean $\alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}$ and variance $\sigma^2 = 1$. Four configurations were considered to set up the symmetry in the effect of X_1 and X_2 on $Y(0)$

with $\alpha_0 = 1$; $\boldsymbol{\alpha}' = (1,1)$, $\boldsymbol{\alpha}' = (1,-1)$, $\boldsymbol{\alpha}' = (-1,1)$ and $\boldsymbol{\alpha}' = (-1,-1)$. A nonignorable assignment process is modeled through the logistic model

$$\text{logit}\{\Pr(T = 1|X_1, X_2, Y(0))\} = -1 + X_1 + X_2 + \gamma Y(0),$$

Letting $Y_i(1) = Y_i(0) + \tau$ and $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$, we generated $(Y_i, T_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$ and then computed the MLE $(\hat{\tau}_1, \hat{\gamma})$ that maximizes (8). We denote $\hat{\tau}_0$ for the MLE that maximize (8) under $H_0^{Ign} : \gamma = 0$. We also computed the minimum eigen value of the observed Fisher information to check the correctness of the regularity condition (B).

Biases and simulated 95% coverage probabilities for $(\hat{\tau}_1, \hat{\gamma})$ based on 1,000 replications are reported in Table 1. As shown, the biases of $\hat{\gamma}$ are almost negligible. There exist modest biases of $\hat{\tau}_1$, which gets small as the sample size becomes large. The simulated coverage probabilities for $\hat{\gamma}$ are very close to 0.95 while those of $\hat{\tau}_1$ are slightly smaller than 0.95 . Sample sizes more than 800 appear to be neccessary to get an accurate confidence interval for $\hat{\tau}_1$. The positive definiteness of the observed Fisher information matrix is guaranteed from the rightmost column. Furthermore, the smallest eigen values in the observed Fisher information matrix get larger when the sample size increases, which implies that (B) is correct at the maxima.

Next, we compare the biases of $\hat{\tau}_1$ and $\hat{\tau}_0$ as shown in Table 2. As expected from Theorem I, both $\hat{\tau}_1$ and $\hat{\tau}_0$ has negligible biases under $H_0^{Ign} : \gamma = 0$. However, under $H_0^{Ign} : \gamma \neq 0$, $\hat{\tau}_0$ produces systematic biases while $\hat{\tau}_1$ does not. The directions of biases are determined by the signs of γ. Thus, as the value of γ departs from 0, the biases of $\hat{\tau}_0$ increase and therefore, $|\hat{\tau}_1 - \hat{\tau}_0|$ gets large.

Finally, we compared the power of the likelihood ratio test on $\gamma \in (-3.5, 3.5)$ based on 5% level from 1,000 replications, for a chosen value of $\tau = 1.5$. Figure 1 shows the power curves with sample sizes 200, 400 and 600. For all sample sizes considered here, the powers at $\gamma = 0$ are very close to 5%, indicating that the test has accurate type I error rates. The power trajectories in all configurations are what is expected; the probability that rejects $H_0^{Ign} : \gamma = 0$ increases as γ deviates from zero. Larger sample sizes provide uniformly higher power than smaller sample sizes. Simulation results for the other values

of τ, such as $\tau = 0$ and $\tau = 3$, exhibit virtually the same tendency as Figure 1, so we do not report those results here.

5 Data analysis

There has been substantial interest in quantifying the efficacy of English education at the elementary school levels in Japan. A major challenge in program evaluation is that it is typically impossible to study the effect of English education using controlled randomized experimentation. In Section 5.1, we introduce the survey data in Katsuyama et al. (2006), and in Section 5.2, we demonstrate the utility of the proposed method using this dataset.

5.1 Test score data from Japanese elementary schools

The primary interest in Katsuyama et al. (2006) is to measure the effect of an English educational program applied to a Japanese elementary school, School A in Chiba prefecture. As a control group, they choose School B located in the same school district but without any English educational program.

English test scores and some background information for students were collected at School A from December 15th to 17th, 2003 and at School B from February 24th and 27th, 2004, respectively. Both datasets were for the same school year of 2003. The participants include 369 students from School A and 146 students from School B. The observed pretreatment covariates are the scholastic years and the English study experiences at kindergarten, which are obtained from questionnaires. Specifically, we define the pretreatment covariate vector $\boldsymbol{X}' = (X_1, X_2)$ as follows:

- X_1: A categorical variable representing student's scholastic year (from 2 to 6 years)
- X_2: An indicator variable for learning English at kindergarten

Let T be the indicator variable such that $T = 1$ if the student was educated in School A and $T = 0$ if the student was educated in School B. Also let $Y(0)$ denote the potential

score for students without English educational program and $Y(1)$ denote the potential score for students with English educational program.

5.2 Data analysis using test score data

A primary concern in the study of English educational effect based on Schools A and B is that, even after adjusting for the observed $\boldsymbol{X}$, we still cannot compare the two schools. In other words, more background information may be necessary for adjustments. If H_0^{Ign} is correct, then the adjustment by $\boldsymbol{X}$ is sufficient for comparing these two schools.

Before applying the proposed parametric approach, we temporalily impose H_0^{Ign}. Then, $\hat{\tau}$ is a valid point estimate for the program effect. We obtain $\hat{\tau} = 1.525$ (SD=0.411), as shown in Table 1. If either models $E\{Y_i(k)|T_i = 1, \boldsymbol{X}_i\} = \alpha_{k0} + \boldsymbol{\alpha}_k'\boldsymbol{X}_i$ or the logistic model on the propensity score is correct, then $\hat{\tau}_{IPW}$ is also a consistent estimator. The estimate is 1.770 (SD=0.393). The $\hat{\tau}_{IPW}$ has smaller standard deviation than $\hat{\tau}$ since it utilize some model assumptions and is semiparametirc efficient. The 95% confidence intervals are away from zero for both estimators. Thus it provides a significant evidence for the positive educational effect for School A. Here, we must emphasize that this conlcusion is valid under the untested assumption of H_0^{Ign}.

To apply the proposed likelihood analysis, one needs to check Assumption I and functional forms of $f_\epsilon(\cdot|\psi)$ and g. Since it is a common practice to approximate the distribution of test scores by a normal distribution, we fit the normal linear model specified as $f_\epsilon(x|\psi) = e^{-x^2/(2\psi)}/\sqrt{2\pi\psi}$ and $g(x) = x$ as in Section 3.2. Under Assumptions I and $g(x) = x$, one can obtain equations

$$Y_i = Y_i(0) + \tau T_i,$$
$$Y_i(0) = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}_i + \epsilon_i,$$

for all i. Combining these equations, we have

$$Y_i = \alpha_0 + \tau T_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \epsilon_i. \tag{10}$$

If the assumption of strong ignorability holds, then $E(\epsilon_i|\boldsymbol{X}_i, T_i) = E(\epsilon_i|\boldsymbol{X}_i) = 0$. Therefore, equation (10) can be interpreted as a usual linear regression of Y_i on $(T_i, \boldsymbol{X}_i)$. The

validity of (10) can be checked by the residual plot, a popular tool for regression analyses. If Assumption I does not hold or the form of $g(x) = x$ is wrong, equation (10) will not hold in general. The residuals shown in Figure 2 seem to be scattered randomly around the horizontal axis and systematic departure from the linear assumption cannot be found. Next, we check the functional form of $f_\epsilon(x|\psi) = e^{-x^2/(2\psi)}/\sqrt{2\pi\psi}$. Note that this assumption is equivalent to the normality assumption for ϵ_i under the linear model. We display the Q-Q plot in Figure 3 to check the normality of the residuals. The plot forms an approximately straight line except a few outlying observations in the tails. Thus, we find no enough evidence for rejecting the assumed parametric models.

Next, we fit the proposed parametric models without H_0^{Ign}. Table 1 shows that the point estimate is $\hat{\tau}_1 = 1.921$ (SD=1.388). The estimate may be somewhat close to $\hat{\tau}$ and $\hat{\tau}_{IPW}$, but it gives larger standard deviation. The evidence for the positive educational effect is no longer significant since 95% confidence interval covers $\tau = 0$. We performed the likelihood ratio test to check the validity of $H_0^{Ign} : \gamma = 0$. The test statistic is $2(\hat{l}_1 - \hat{l}_0) = 0.0009$ and the corresponding p-value is 0.976. We do not find a statistical evidence for the departure from H_0^{Ign}. Note that it is also possible to estimate the treatment effect using the maximized likelihood methods under $H_0^{Ign} : \gamma = 0$. The maximum likelihood procedure under $H_0^{Ign} : \gamma = 0$ yields an estimate $\hat{\tau}_0 = 1.760$ (SD=0.393), which is very close to the results of $\hat{\tau}_{IPW}$.

Given the acceptance of $H_0^{Ign} : \gamma = 0$, we are much confident that $\hat{\tau}$ and $\hat{\tau}_{IPW}$ are consistent estimators of the true parameter. If the efficiency and robustness are of concern, $\hat{\tau}_{IPW}$ may be the best choice. In conclusion, we reject $H_0^{Trt} : \tau = 0$ and accept $H_0^{Ign} : \gamma = 0$. The simultaneous assessment of the two assumptions, $H_0^{Trt} : \tau = 0$ and $H_0^{Ign} : \gamma = 0$, is an important feature of the proposed approach.

We obtained four different estimates for the treatment effect and their values were somewhat similar. Since these estimates should be close to each other under the conditions of Theorem I, the current dataset may be a good example to show the case that Theorem I holds. For some datasets, the differences among the four point estimates could be large and the statistical test of the strong ignorability could be rejected. In such cases, it is not

trivial to know which point estimates are valid. Theorem I implies that, large differences between the estimators could be due to either the violation of the strong ignorability or misspecifications of the imposed parametric forms. Rejection of the strong ignorability could also occur due to the same reasons.

6 Conclusion and discussion

In this article, we have proposed a parametric approach for assessing the treatment effect and applied it to investigate the English educational effect in Japanese elementary schools. We have emphasized the fact that the assumption of strong ignorability is not statistically testable nonparametrically. The proposed method gives an alternative way of estimating the treatment effect without the assumption of strong ignorability but with fully parametric assumptions as well as the assumption of constant effect. It also provides a tool for testing the assumption of strong ignorability based on the likelihood ratio statistic. As a result, the two important conditions, namely no treatment effect and strong ignorability, can be empirically tested by the proposed method.

We also compared our method with the double-robust method of Robins et al. (1994). The double-robust approach in Robins et al. is valid when either the mean-regression model on the response or the propensity score model is correctly specified. Our approach requires both of them. It further requires the parametric specification of the error distribution as well as the assumption of constant effect. Instead, the proposed method does not requires the assumption of strong ignorability. Therefore, the proposed method is most useful when researchers have enough information about the distributional assumptions of their models, possibly from previous experiences. In addition, we suggest conducting rigorous model diagnostics procedures before applying the proposed method. In many social sciences, the distributional forms may be known *apriori* from the previous studies but researchers are unsure about the strong ignorability, especially when there is not enought covariats available due to financial or ethical reasons as in our English test score study.

The correctness of the model assumptions is critical in the present parametric ap-

proach. The most funcamental assumption is the constant effect stated in Assumption I. We must admit that this assumption is fairly strong. However, the assumption makes the counterfactual models identifiable by reducing the demension in the model. Also, Assumption I implies the equality of variances, which may be a more acceptable assumption for practitionars who routinly use analysis of variance for their experiments. For the special case of the normal distribution models, Assumption I may be checked empirically from data as we demonstrated in Section 5.2. The assumption of the parametric form in the error distribution is also restrictive. The normal distribution assumption in the case study is in many ways suited to analysis of test score data, where the potential outcomes can be approximated by the normal distributions. In many cases, however, the normal distribution may not correctly represent the population model if, for example, the English test score were highly skewed. Transformation of the response variables sometimes allieviate the problem of nonnormality.

Appendix: score and observed Fisher information matrix

Let $\psi = \sigma^2$. The log-likelihood function for θ can be written as $\sum_i l_i(\theta)$, where

$$\begin{aligned} l_i(\theta) = &-\frac{1}{2}\log(\psi) - \frac{1}{2\psi}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i)^2 \\ &+ T_i\{\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i - \gamma\tau T_i\} - \log\{1 + e^{(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i - \gamma\tau T_i}\}. \end{aligned}$$

We write

$$\mu_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau) = \frac{e^{\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i - \gamma\tau T_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i - \gamma\tau T_i}},$$

$$v_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau) = \mu_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau)\{1 - \mu_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau)\}.$$

Then, the score functions for $l_i(\theta)$ can be written as

$$\frac{\partial l_i(\theta)}{\partial \tau} = \frac{T_i}{\psi}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i) - \gamma T_i\{T_i - \mu_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau)\},$$

$$\frac{\partial l_i(\theta)}{\partial (\alpha_0, \boldsymbol{\alpha}')'} = \frac{1}{\psi}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i),$$

$$\frac{\partial l_i(\theta)}{\partial \psi} = -\frac{1}{2\psi} + \frac{1}{2\psi^2}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i)^2,$$

$$\frac{\partial l_i(\theta)}{\partial (\beta_0, \boldsymbol{\beta}', \gamma)'} = \begin{pmatrix} 1 \\ \boldsymbol{X}_i \\ Y_i - \tau T_i \end{pmatrix}\{T_i - \mu_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau)\}.$$

Also, the observed Fisher information matrix consists of

$$-\frac{\partial^2 l_i(\theta)}{\partial \tau^2} = T_i\left[\frac{1}{\psi} + \gamma^2 v_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau)\right],$$

$$-\frac{\partial^2 l_i(\theta)}{\partial \tau \partial (\alpha_0, \boldsymbol{\alpha}')'} = \frac{T_i}{\psi}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix},$$

$$-\frac{\partial^2 l_i(\theta)}{\partial \tau \partial \psi} = \frac{T_i}{\psi^2}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i),$$

$$-\frac{\partial^2 l_i(\theta)}{\partial \tau \partial (\beta_0, \boldsymbol{\beta}', \gamma)'} = -\gamma T_i \begin{pmatrix} 1 \\ \boldsymbol{X}_i \\ Y_i - \tau T_i \end{pmatrix} v_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau),$$

$$-\frac{\partial^2 l_i(\theta)}{\partial (\alpha_0, \boldsymbol{\alpha}')' \partial (\alpha_0, \boldsymbol{\alpha}')} = \frac{1}{\psi}\begin{pmatrix} 1 & \boldsymbol{X}_i' \\ \boldsymbol{X}_i & \boldsymbol{X}_i\boldsymbol{X}_i' \end{pmatrix},$$

$$-\frac{\partial^2 l_i(\theta)}{\partial (\alpha_0, \boldsymbol{\alpha}')' \partial \psi} = \frac{1}{\psi^2}\begin{pmatrix} 1 \\ \boldsymbol{X}_i \end{pmatrix}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i),$$

$$-\frac{\partial^2 l_i(\theta)}{\partial (\alpha_0, \boldsymbol{\alpha}')' \partial (\beta_0, \boldsymbol{\beta}', \gamma)} = O,$$

$$-\frac{\partial^2 l_i(\theta)}{\partial \psi^2} = -\frac{1}{2\psi^2} + \frac{1}{\psi^3}(Y_i - \tau T_i - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i)^2,$$

$$-\frac{\partial^2 l_i(\theta)}{\partial \psi \partial (\beta_0, \boldsymbol{\beta}', \gamma)} = \mathbf{0}',$$

$$
\begin{aligned}
&- \frac{\partial^2 l_i(\theta)}{\partial(\beta_0, \boldsymbol{\beta}', \gamma)' \partial(\beta_0, \boldsymbol{\beta}', \gamma)} \\
&= \begin{pmatrix} 1 & \boldsymbol{X}_i' & Y_i - \tau T_i \\ \boldsymbol{X}_i & \boldsymbol{X}_i \boldsymbol{X}_i' & (Y_i - \tau T_i)\boldsymbol{X}_i \\ Y_i - \tau T_i & (Y_i - \tau T_i)\boldsymbol{X}_i' & (Y_i - \tau T_i)^2 \end{pmatrix} v_i(\beta_0, \boldsymbol{\beta}', \gamma, \tau).
\end{aligned}
$$

References

[1] Cochran, W. G. (1965), "The Planning of Observational Studies of Human Population (with discussion)," *Journal of the Royal Statistical Society, Series B*, 128, 234-235.

[2] Coleman, J. S., Hoffer, T., Kilgore, S. (1982), *High School Achievement*, New York: Basic.

[3] Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society, Series B*, 41, 1-31.

[4] Dehejia, R. H., and Wahba, S. (1998), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

[5] Foutz, R. V. (1977), "On the Unique Consistent Solution to the Likelihood Equations," *Journal of the American Statistical Association*, 72, 147-148.

[6] Goldberger, A. S. and Cain, G. G. (1982), "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer and Kilgore Report," *Sociology of Education*, 55, 103-122.

[7] Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.

[8] Holland, P. W. (1986), "Statistics and Causal Inference (with discussion)," *Journal of the American Statistical Association*, 81, 945-970.

[9] Katsuyama, H., Nishigaki, C. and Wang, J. (2006), "A Study on the Effect of English Teaching in Public Elementary Schools," *KATE Bulletin*, 20, 113-124.

[10] Katsuyama, H., Nishigaki, C. and Wang, J. (2008), "The Effectiveness of English Teaching in Japanese Elementary Schools: Measured by Proficiency Tests Administered to Seventh-year Students, *RELC Journal*, 39, 359-380.

[11] McCullagh, N. T. and Nelder, J. A. (1991), *Generalized Linear Models, 2nd ed.*, London: Chapman & Hall/CRC.

[12] Morgan, S. L. (2001), "Counterfactuals, Causal Effect Heterogeneity, and the Catholic School Effect on Learning," *Sociology of Education*, 74, 341-374.

[13] Otsu, (2004), *Is English In Elementary School Necessary?* Keio University Press, Inc.

[14] Otsu, (2005), *English Teaching in Elementary School Is Not Necessary.* Keio University Press, Inc.

[15] Pearl, J. (2001), "Causal Inference in the Health Sciences: A Conceptual Introduction," *Health Services and Outcomes Research Methodology*, 2, 189-220. *Journal of the American Statistical Association*, 79, 41-48.

[16] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some of Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 94, 846-866.

[17] Rosenbaum, P. (2002), *Observational Studies*, New York: Springer-Verlag.

[18] Rosenbaum, P. and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

[19] Van Der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

[20] Winship, C. and Morgan, S. L. (1999), "The Estimation of Causal Effects from Observational Data," *Annual Review of Sociology*, 25, 659-706.

Table 1: Performances of the maximum likelihood estimators for (τ, γ) based on 1,000 replications when $\alpha_0 = 1$, $\boldsymbol{\alpha}' = (1, 1)$.

n	(τ, γ)	Bias for ($\hat{\tau}_1$, $\hat{\gamma}$)	95%Cov for ($\hat{\tau}_1$, $\hat{\gamma}$)	smallest eigen value of $ni_n(\hat{\theta})$
800	(0, -1.5)	(-0.079, 0.053)	(0.930, 0.946)	1.474 (0.0194)
	(1.5, -1.5)	(-0.077, 0.052)	(0.927, 0.946)	1.465 (0.0149)
	(3, -1.5)	(-0.079 0.057)	(0.935 0.963)	1.466 (0.0089)
	(0, 1.5)	(0.017, 0.026)	(0.952, 0.956)	2.127 (0.2825)
	(1.5, 1.5)	(0.034, -0.007)	(0.947, 0.962)	2.117 (0.3200)
	(3, 1.5)	(0.035, 0.004)	(0.947, 0.953)	2.111 (0.1385)
600	(0, -1.5)	(-0.088, 0.037)	(0.912, 0.942)	1.087 (0.0052)
	(1.5, -1.5)	(-0.090, 0.044)	(0.923, 0.947)	1.076 (0.0083)
	(3, -1.5)	(-0.114, 0.075)	(0.914, 0.941)	1.079 (0.0067)
	(0, 1.5)	(0.037, 0.002)	(0.939, 0.949)	1.559 (0.1059)
	(1.5, 1.5)	(0.038, 0.003)	(0.939, 0.965)	1.547 (0.1372)
	(3, 1.5)	(0.053, -0.014)	(0.939, 0.953)	1.534 (0.0820)
400	(0, -1.5)	(-0.144, 0.070)	(0.877, 0.915)	0.683 (0.0031)
	(1.5, -1.5)	(-0.097, 0.034)	(0.897, 0.934)	0.694 (0.0047)
	(3, -1.5)	(-0.134, 0.064)	(0.908, 0.942)	0.703 (0.0101)
	(0, 1.5)	(-0.062, -0.019)	(0.908, 0.937)	0.988 (0.0753)
	(1.5, 1.5)	(0.082, -0.018)	(0.927, 0.934)	0.974 (0.0370)
	(3, 1.5)	(0.058, 0.024)	(0.901, 0.918)	0.986 (0.0820)

NOTE: Bias is the average of $(\hat{\tau}_1, \hat{\gamma}) - (\tau, \gamma)$. 95% confidence intervals (denoted as 95%Cov) are based on the Wald-type normal approximations using the standerd error calculated from the observed Fisher information matrix. Averages and minimums (in parenthesis) of the smallest eigen values of the observed Fisher information matrix ($ni_n(\hat{\theta})$) are reported in the rightmost column.

Table 2: Bias and standard deviations (SD) for jointly estimating (τ, γ) under H_1 based on 1,000 replications.

n	(τ, γ)	Bias for $\hat{\tau}_1$ (SD for $\hat{\tau}_1$)	Bias for $\hat{\tau}_0$ (SD for $\hat{\tau}_0$)
600	(0, -1.5)	-0.088 (0.493)	-1.216 (0.119)
	(1.5, -1.5)	-0.090 (0.500)	-1.218 (0.124)
	(3, -1.5)	-0.114 (0.525)	-1.217 (0.121)
	(0, 0)	0.010 (0.687)	-0.001 (0.084)
	(1.5, 0)	-0.012 (0.697)	-0.004 (0.087)
	(3, 0)	0.038 (0.701)	-0.001 (0.086)
	(0, 1.5)	0.037 (0.333)	1.133 (0.109)
	(1.5, 1.5)	0.038 (0.331)	1.132 (0.110)
	(3, 1.5)	0.053 (0.358)	1.135 (0.105)
400	(0, -1.5)	-0.144 (0.646)	-1.217 (0.152)
	(1.5, -1.5)	-0.097 (0.583)	-1.209 (0.148)
	(3, -1.5)	-0.134 (0.605)	-1.209 (0.148)
	(0, 0)	0.049 (0.750)	0.000 (0.103)
	(1.5, 0)	0.007 (0.708)	-0.001 (0.102)
	(3, 0)	0.020 (0.703)	0.003 (0.106)
	(0, 1.5)	0.062 (0.452)	1.141 (0.133)
	(1.5, 1.5)	0.082 (0.445)	1.140 (0.134)
	(3, 1.5)	0.058 (0.446)	1.135 (0.135)

NOTE: $\hat{\tau}_1$ is the MLE without $H_0^{Ign} : \gamma = 0$ while $\hat{\tau}_0$ is the MLE under $H_0^{Ign} : \gamma = 0$.

Table 3: Four point estimates for the English educational effect τ , their standard deviations and 95% confidence intervals.

	Estimate	SD	95% conf. interval
$\hat{\tau}$	1.525	0.411	(0.713, 2.325)
$\hat{\tau}_{IPW}$	1.770	0.393	(1.015, 2.524)
$\hat{\tau}_0$	1.760	0.393	(1.008, 2.534)
$\hat{\tau}_1$	1.921	1.388	(-2.652,2.963)

NOTE: $\hat{\tau}$ is the stratification estimator, $\hat{\tau}_{IPW}$ is the IPW estimator of Robins et al., $\hat{\tau}_1$ is the MLE under $H_0^{Ign} : \gamma = 0$ and $\hat{\tau}_1$ is the MLE without H_0^{Ign}. Standard deviations (SD) and 95% confidence intervals are based on the sample variance and percentiles for $B = 5,000$ Booststrap replications. The seed of Bootstraping samples are the same for the four estimators.

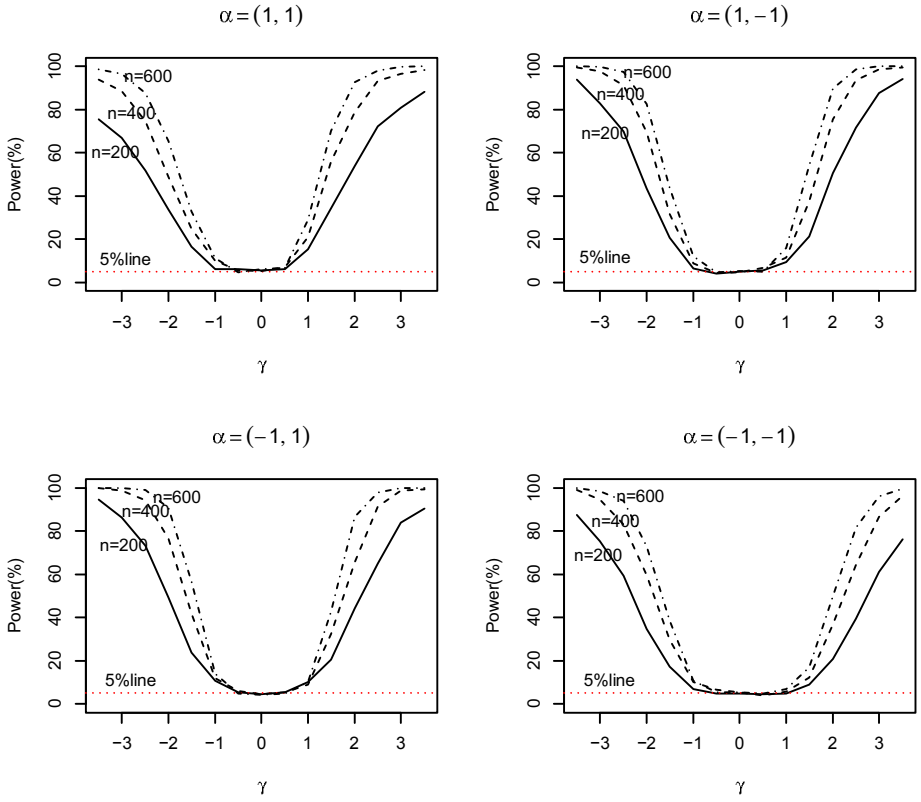

Figure 1: Power curves for testing $H_0^{Ign} : \gamma = 0$ with sample sizes $n = 200$ (solid Line), $n = 400$ (dashed line) and $n = 600$ (dashed and dotted line) based on 1,000 replications. Potential outcomes $Y(0)$ are generated from the normal distributions with mean $1 + \boldsymbol{\alpha}'X$ and variance 1. The assignment variable T follows a Bernoulli random distribution with the success probability p modeled by the logit model, $\text{logit}(p) = -1 + X_1 + X_2$.

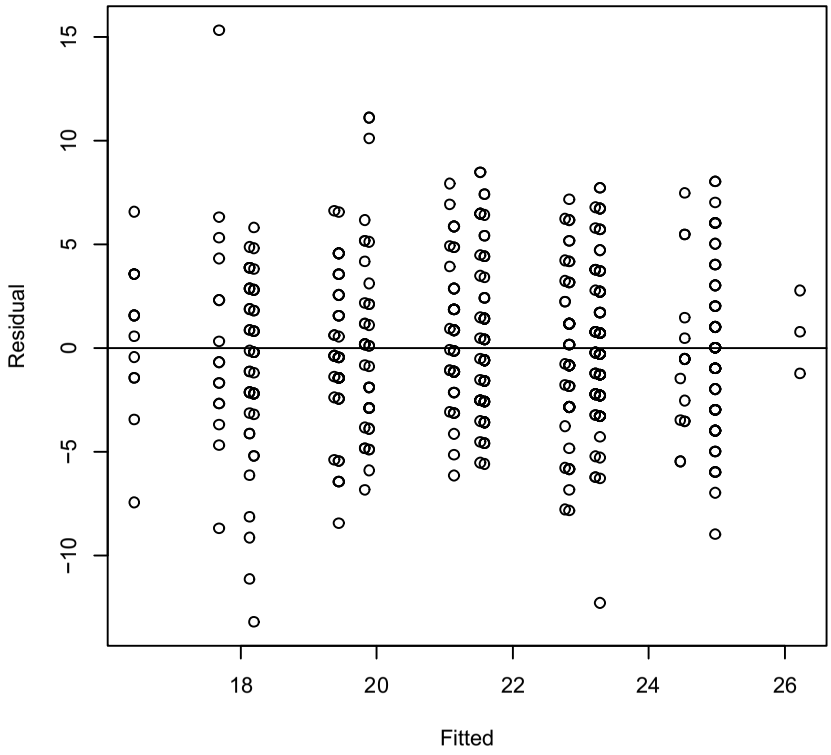

Figure 2: The residual plot for checking the linear assumption for $Y_i = \alpha_0 + \tau T_i + \boldsymbol{\alpha}' \boldsymbol{X}_i + \epsilon_i$. Variables Y_i and $\boldsymbol{X}_i$ are the test score and student background information respectively for a student from Japanese elementary School A ($T_i = 1$) or School B ($T_i = 0$). The residuals $Y_i - \hat{\alpha}_0 + \hat{\tau} T_i + \hat{\boldsymbol{\alpha}}' \boldsymbol{X}_i$ are plotted against the fitted linear predictors $\hat{\alpha}_0 + \hat{\tau} T_i + \hat{\boldsymbol{\alpha}}' \boldsymbol{X}_i$.

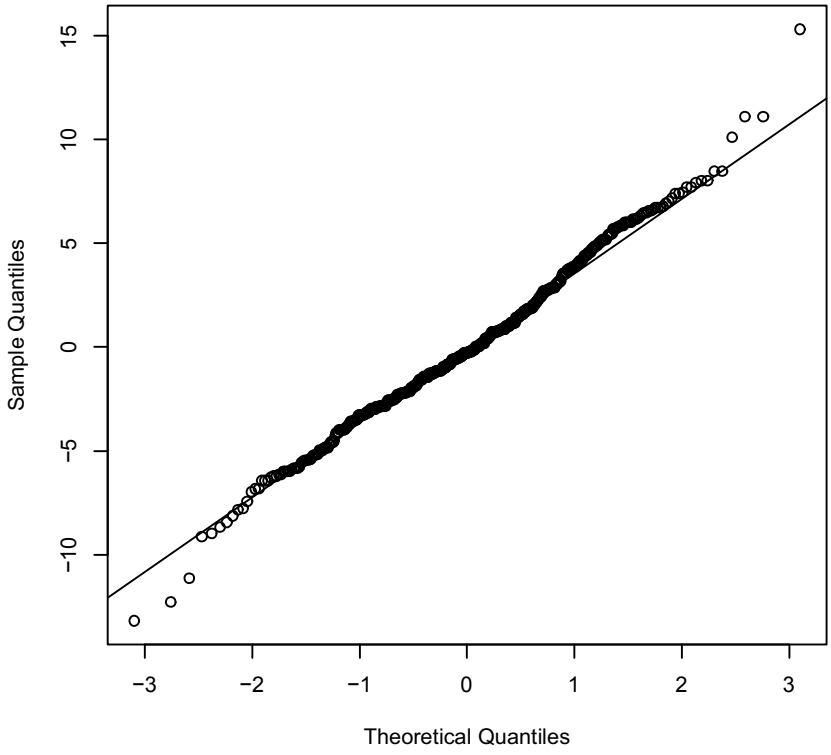

Figure 3: Q-Q plot for checking the normality assumption for the error distribution in $Y_i = \alpha_0 + \tau T_i + \boldsymbol{\alpha}' \boldsymbol{X}_i + \epsilon_i$. Variables Y_i and $\boldsymbol{X}_i$ are the test score and student backgrounds respectively for a student from Japanese elementary School A ($T_i = 1$) or School B ($T_i = 0$). The straight line passes through the first and third quartiles.