Assessing the Assumption of Strongly Ignorable Treatment Assignment Under Assumed Causal Models

Takeshi Emura, Jingfang Wang and Hitomi Katsuyama*

April 23, 2008

Abstract

The assumption of strong ignorability is imposed by most inference methods for estimating the treatment effect in observational studies. We propose a likelihood ratio test for checking this important assumption. Simulations are conducted to investigate the performance of the proposed method. We demonstrate the proposed method using data from an observational study for measuring English educational effect on Japanese elementary school students.

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

*Takeshi Emura is Assistant Professor, Division of Biostatistics, School of Pharmaceutical Sciences, Kitasato University, 5-9-1, Sirokane, Minato-ku, Japan, Jingfang Wang is Associate Professor, Faculty of Sciences, Chiba University, 1-33, Yayoi-cho, Inageku, Chiba-shi, Japan, and Hitomi Katsuyama is Lecturer, Department of English, Kawamura Gakuen Women's University, 1133 Sageto Abiko-shi, Chiba, Japan

1 INTRODUCTION

Observational studies are often used to study the effect of a treatment 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 setting. In Japan, the efficacy of early introduction of English training program is of substantial interest to both the general public and the government (Katsuyama, Nishigaki and Wang 2006). In Section 5.2, we consider a study on the effect of English educational program in School A in Chiba prefecture, Japan, based on data collected by the authors. 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. Specifically, 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 a common statistical practice to check the model assumptions on which estimates are calculated. Most model checking procedures can be roughly classified either by significance test or graphical diagnosis. The former approach is useful for quantifying the model discrepancy. Since a hypothesis test can potentially reject a model assumption for a large sample size even when the data structure only slighty differs from the model. For this reason, many applied researchers prefer graphical diagnostic approach to see the adequecy of model assumptions. It is not unusual that such model checking procedures require more elaborate techniques than the estimation for the treatment effect itself. Similar problems arise in survival analysis where survival time is subject to censoring by another random variable (Klein and Moeschberger 2003). Most techniques on survival analysis rely on the independence assumption between the survival and censoring times. However, it is well known that the independence assumption is not testable by data nonparametrically. This is known as the identifiability dilemma, and we will discuss this non-identifiability issue under the counterfactual models of causality in Section 2.2.

The paper is organized as follows. Section 2 introduces the counterfactual models of causality and states the assumption of strong ignorability. Section 3 reviews existing methods for assessing the assumption of strong ignorability. Section 4 proposes a method for testing the assumption of strong ignorability. Section 5 presents Monte Carlo simulation studies and data analysis results on the effect of English educational program carried out by the authors. Section 6 concludes the paper.

2 COUNTERFACTUAL MODELING

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. The difference $\Delta_i = Y_i(1) - Y_i(0)$ indicates that unit i gains the increase/decrease in the response variable by Δ_i due to the treatment. The average treatment effect describes the expected gain due to the treatment and it is defined as $\tau = E\{Y_i(1)\} - E\{Y_i(0)\}$, where $E(\cdot)$ denotes expectation in the population. 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. All information available to us is the set of observations $(Y_i, T_i, \boldsymbol{X}_i)$; $i = 1, \dots n$. Throughout the paper, we assume that the set of n observations are independent and identical replica of $(Y, T, \boldsymbol{X})$ whose distribution is determined by the probability distribution on $(Y(0), Y(1), T, \boldsymbol{X})$, where the subscript i is omitted. A covariate vector $\boldsymbol{X}$ is a potential confounder that may relate to the distributions of both $(Y(0), Y(1))$ and T. 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*

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

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

Here, the notation $A \perp B \mid C$ represents independence between variables A and B given an event 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 $\boldsymbol{X} = \boldsymbol{x}$. If the strong ignorability holds, the following identity can be used to construct an unbiased estimator of the average treatment effect:

$$\begin{aligned}\tau &= E[E\{Y(1)|\boldsymbol{X}\} - E\{Y(0)|\boldsymbol{X}\}] \\ &= E[E\{Y|T=1,\boldsymbol{X}\} - E\{Y|T=0,\boldsymbol{X}\}].\end{aligned}$$

If the covariate vector $\boldsymbol{X}$ is discrete, the method of moment leads to the following non-parametric estimator

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

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. 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 implimented on $P(\boldsymbol{X})$ rather than on $\boldsymbol{X}$. Dehejia and Wahba (1999) gives case studies for implementing propensity score method. An application for studying more elaborated causal mechanisms based on propensity score is proposed by Hong and Raudenbush (2006). All methodologies using propensity score, as well as the above two papers, rely entirely upon the assumption of strong ignorability. Although testing the assumption of strong ignorability is a statistical problem of considerable importance, there are few literature on the formal assessment of the assumption.

2.2 The Non-Identifiability Dilemma

One important issue in applying the estimator (1) and the propensity score method to a given data set is how to check the assumption of strong ignorability. We extract the core statement from Definition I:

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

Unfortunately, it turns out that H_0 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* $\boldsymbol{X}$,

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

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

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

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

In Model A, unit requires to enjoy the causal treatment effect Δ if 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 respense 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 $(Y, T|\boldsymbol{X})$:

$$\frac{1}{2} f(Y - \Delta T|\boldsymbol{X}). \tag{3}$$

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. This problem occurs since the family of distributions on $(Y(0), Y(1), T, \boldsymbol{X})$ is too large. More specifically, the dimension of the model on $(Y(0), Y(1), T, \boldsymbol{X})$ is higher than that of the observed model on

$(Y, T, \boldsymbol{X})$. Some restricting assumptions may help reduce the dimension of the family of distributions on $(Y(0), Y(1), T, \boldsymbol{X})$. For example, the additive treatment effecct assumption $Y(1) = Y(0) + \Delta$ for some unknown parameter Δ (see, e.g., Lehmann 1975 and Rosenbaum 2002) reduces a bivariate model on $(Y(0), Y(1))$ to one-dimensional distribution. If the additive treatment effect model is assumed, one can exclude Model A as a candidate model and identify Model B. The idea of demension reduction plays a foundamental role in the likelihood construction in Section 4. We suspect that the lack of literature on testing the assumption of strong ignorability may be related to the non-identifiability aspect of the counterfactual models.

3 EXISTING STRATEGIES FOR ASSESSING H_0

Before presenting our proposed testing method, we review some existing methodologies to assess the assumption of strong ignorability.

3.1 A Sensitivity Analysis

The robustness of the quantitative conclusions against the violation of the strongly ignorable assumption may be assessed by a sensitivity analysis (Rosenbaum 2002, chapter 4). This is a clever technique for bounding the range of estimated treatment effect when there exists an unobserved covariate. The sensitivity analysis requires to set a logit model for an unobserved covariate u:

$$\text{logit}\{\Pr(T = 1|\boldsymbol{X}, U = u)\} = \lambda(\boldsymbol{X}) + \gamma u, \quad 0 \leq u \leq 1. \tag{4}$$

Here, $\lambda(\cdot)$ is a function of $\boldsymbol{X}$, and γ is an unknown parameter. If $\gamma = 0$ and $0 < \lambda(\boldsymbol{x}) < 1$ for all $\boldsymbol{x}$, the assumption of strong ignorability is satisfied since the potential confounder u has no distributional effect on T given $\boldsymbol{X}$. For $\gamma \neq 0$, u is unobserved confounding variable that might affect $(Y(0), Y(1))$ and do affect T. For a given data set, a sensitivity analysis provides a bound for the estimates for the treatment effect when u moves over the interval $[0, 1]$ for each subject. Here, it is convenient to define the discrepancy measure from the strong ignorability by $\Gamma = e^{\gamma}$ that has an interpretation as the odds ratio (Rosenbaum 2002, pp. 108-109). Algorithms for sensitivity analysis for a given Γ depend on the type of estimators and methods of bounding used (Rosenbaum 2002, chapter 4). Note that the sensitivity analysis does not test H_0 itself.

3.2 Nonparametric Test for H_0

Rosenbaum (1984) provides an idea of testing H_0 in a nonparametric fashion. Now we briefly discuss a part of his idea. To make H_0 testable, he consider imposing the inequality constraint $Y(0) \leq Y(1)$. This assumption may be reasonable under many situations where there is a uniformly positive or negative treatment effect. Following the law of basic probability calculus, one can obtain the following stochastic order relation under H_0:

$$\Pr(Y \leq t | T = 0, \boldsymbol{X} = \boldsymbol{x}) \geq \Pr(Y \leq t | T = 1, \boldsymbol{X} = \boldsymbol{x}), \tag{5}$$

for all $\boldsymbol{x}$ and t. The inequality may not be true if H_0 does not hold. Rosenbaum proposes to check the inequality (5) for the purpose of testing H_0. If the covariate vector $\boldsymbol{X}$ is of low dimension and there are sufficiently many samples at each observed $\boldsymbol{X}$, one may check the following inequality:

$$\frac{\sum_i \mathrm{I}(Y_i \leq t, T_i = 0, \boldsymbol{X}_i = \boldsymbol{x})}{\sum_i \mathrm{I}(T_i = 0, \boldsymbol{X}_i = \boldsymbol{x})} \geq \frac{\sum_i \mathrm{I}(Y_i \leq t, T_i = 1, \boldsymbol{X}_i = \boldsymbol{x})}{\sum_i \mathrm{I}(T_i = 1, \boldsymbol{X}_i = \boldsymbol{x})}. \tag{6}$$

Although this idea is useful in some cases, there are also difficulties that may prevent us from using the criterion (6). First, in some observational studies, the covariate vector $\boldsymbol{X}$ may be of high dimension or continuous so that comparison based on (6) may be impractical. Secondary, this method implicitly assumes that, under the alternative hypothesis, say H_1, there exist some $\boldsymbol{x}$ and t such that

$$\Pr(Y \leq t | T = 0, \boldsymbol{X} = \boldsymbol{x}) < \Pr(Y \leq t | T = 1, \boldsymbol{X} = \boldsymbol{x}). \tag{7}$$

Unfortunately, it seems to be difficult to interpret the inequality (7) as the departure from H_0. In other words, the test based on (6) may not have reasonable power under H_1.

In this paper, we will not use comparisons based on equation (5). Instead, we will propose a test for H_0 based on an appropriate likelihood function, which allows a flexible form of $\boldsymbol{X}$ and is equiped with a natural alternative.

4 LIKELIHOOD RATIO TEST

The non-identifiability dilemma elaborated in Section 2.2 explains a technical difficulty of testing H_0. Nevertheless, by imposing appropriate assumptions on the counterfacual models, we show that H_0 will become testable. In this section, we present a procedure to test H_0, which employs the popular likelihood ratio test.

4.1 A Likelihood Ratio Test for H_0

The test proposed in this paper requires a parametric specification on the counterfactual repsponse $(Y(0), Y(1))$. Notice that the average treatment effect $\tau = E\{Y(1)\} - E\{Y(0)\}$ is the focus of many evaluation studies, where the effect is quantified as the difference between the mean responses. The non-identifiability dilemma discussed in Section 2.2 undermines any attempt to test H_0 nonparametrically; observed data cannot tell whether the model satisfies H_0 or not without any model assumption. To impose conditions that can make H_0 identifiable, we will employ the generalized linear model (GLM) framework.
Assumption I *Additive treatment effect model on* $(Y(0), Y(1))$:

$$Y(1) = Y(0) + \Delta.$$

Assumption II *Linear additive model on the mean response:*

$$g\{E[Y(0)|\boldsymbol{X}]\} = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X},$$

where, $g(\cdot)$ is a given link function and $(\alpha_0, \boldsymbol{\alpha}')$ are unknown parameters
Assumption III *Exponential-family model for $Y(0)$ given $\boldsymbol{X}$*:

$$f(y|\boldsymbol{X}) = \exp\{(y\theta - b(\theta))/\phi + c(y;\phi)\}$$

for some specified functions $b(\cdot)$ and $c(\cdot)$, where θ is a canonical parameter and ϕ is a nuisance parameter.

Assumption I states that the treatment uniformly changes the response $Y(0)$ by a magnitude Δ. The assumption of uniform treatment effect is common in nonparametric settings (Lehmann 1975). This assumption is used, for instance, to compute the well-known Hodges-Lehmann estimator (Hodges and Lehmann 1963). Assumptions II and III are the usual specifications in the GLM framework. A simple normal model satisfying Assumptions I-III will be studied in Section 4.2.

Conditional on the covariate vector $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, we define the full-data likelihood function

$$\prod_i L(Y_i(0), Y_i(1), T_i | \boldsymbol{X}_i) \tag{8}$$

Under Assumptions I-III, the full-data likelihood in (8) can be simplified as

$$\prod_i f(Y_i(0)|\boldsymbol{X}_i) \Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0))^{T_i} \Pr(T_i = 0|\boldsymbol{X}_i, Y_i(0))^{1-T_i} \tag{9}$$

The demension reduction due to Assumption I is very important here since the joint likelihood on $(Y(0), Y(1), T|\boldsymbol{X})$ reduces to the joint likelihood on $(Y(0), T|\boldsymbol{X})$. This model structure also allows us to work on a simpler conditional likelihood on the binary outcome T, conditional only on $Y(0)$, rather than on $(Y(0), Y(1))$. The assumption of strong ignorability implies $\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0)) = \Pr(T_i = 1|\boldsymbol{X}_i)$, that is, the assignment probability for any unit i is independent of the response $Y_i(0)$. The likelihood function (9) can be explicitely written under appropriate parametrization of $\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0))$. For its popularity in applications and mathematical tractability, we recommend the logit model on T:

$$\text{logit}\{\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0))\} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i(0). \tag{10}$$

Under the logit model (10), the null hypothesis of strong ignorability can be written as $H_0 : \gamma = 0$. Unfortunately, $Y_i(0)$ in the likelihood (9) is not directly observable for every i. We replace $Y_i(0)$ by $Y_i^*(0) = Y_i - \tau T_i$, where the parameter τ is to be estimated. This imputed outcome would be equivalent to the true $Y_i(0)$ if τ were the true treatment effect.

The resulting likelihood is given by

$$\begin{aligned} &L(\tau, \alpha_0, \boldsymbol{\alpha}', \sigma^2, \beta_0, \boldsymbol{\beta}', \gamma) \\ &= \prod_i f(Y_i^*(0) \mid \boldsymbol{X}_i) \Pr(T_i = 1 | \boldsymbol{X}_i, Y_i^*(0))^{T_i} \Pr(T_i = 0 | \boldsymbol{X}_i, Y_i^*(0))^{1-T_i}. \end{aligned} \tag{11}$$

This likelihood function is to be maximized over $(\tau, \alpha_0, \boldsymbol{\alpha}', \phi, \beta_0, \boldsymbol{\beta}', \gamma)$ under $H_0 : \gamma = 0$, and under $H_0 \cup H_1$, where $H_1 : \gamma \neq 0$. Let $\hat{l}_0$ and $\hat{l}_1$ be the maximized log-likelihood of (11) under $H_0 : \gamma = 0$ and $H_0 \cup H_1$ respectively. The likelihood ratio statistic

$$2(\hat{l}_1 - \hat{l}_0) \tag{12}$$

can be used as the test statistic, which have an approximate chi-squared distribution with one degree of freedom.

4.2 An Example for Normal Distribution Model

If the response $Y(0)$ follows a normal distribution, Assumption III can be specified as $\theta = \mu$, $c(y; \phi) = -y^2/(2\phi) - (1/2)\log(2\pi\phi)$, and $b(\theta) = \theta^2/2$ as in the usual GLM framework. Under Assumptions I-III, the log-likelihood can be written as

$$\begin{aligned} l(\tau, \alpha_0, \boldsymbol{\alpha}', \sigma^2, \beta_0, \boldsymbol{\beta}', \gamma) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (Y_i^*(0) - \alpha_0 - \boldsymbol{\alpha}'\boldsymbol{X}_i)^2 \\ &+ \sum_i \left[T_i\{\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i^*(0)\} - \log\{1 + \exp(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i^*(0))\}\right], \end{aligned} \tag{13}$$

where $Y_i^*(0) = Y_i - \tau T_i$ is the imputed outcome for $Y_i(0)$. The maximization for (13) under $H_0 : \gamma = 0$ can be obtained by separately applying the maximization over $(\tau, \alpha_0, \boldsymbol{\alpha}', \sigma^2)$ and $(\beta_0, \boldsymbol{\beta}')$ with the usual likelihood procedures. Maximization under $H_0 \cup H_1$ is not available in closed form, but the numerical maximization is still easy to conduct using available software, such as R. Specifically, for a fixed value of τ, we use $Y_i^*(0) = Y_i - \tau T_i$ in (13) as imputed observed values. Then in equation (13), parameters $(\alpha_0, \boldsymbol{\alpha}', \sigma^2)$ are maximized as the least square estimates for $(Y_1^*(0), \ldots, Y_n^*(0))$ and parameters $(\beta_0, \boldsymbol{\beta}', \gamma)$ are maximized as the logistic regression estimates on $(T_1, \ldots, T_n)$. With the maximized likelihood for a given τ, we obtain the profile likelihood $\hat{l}(\tau)$. The global maximum is attained by the maximum point of $\hat{l}(\tau)$ over all possible value of τ. These processes may be done using the combination of `glm` and `optimize` commands in R.

If the assumed normal model is incorrect, the likelihood ratio test may not have a valid type I error rate. We propose to perform model checking for Assumptions I-III. Fortunately, for normal distribution models, it is easy to assess the Assumptions I-III by standard regression procedures. Such methods are illustrated in Section 5.4 through real data analysis.

4.3 Properties of the Proposed Test

Properties for the proposed test directly follow from the general theory of the likelihood ratio test. The likelihood ratio test is asymptotically optimal for testing H_0 when the one-sided alternative hypothesis $H_1 : \gamma > 0$ or $H_1 : \gamma < 0$ is specified under Assumptions I-III (Van Der Vaart 1999). On the other hand, the test may have incorrect type I error rate when some of Assumptions I-III fail to hold. A detailed treatment of the behavior of the proposed test under model misspecification is beyond the scope of this paper. However, in Section 5.2, we do examine the behavior of the proposed test for a selected misspecified model by a series of simulations.

Given the data, the likelihood ratio statistic quantifies the increase of the fit due the inclusion of the parameter γ. Specifically, if the model

$$\text{logit}\{\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0))\} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \gamma Y_i(0) \tag{14}$$

is more suitable in describing the data structure than the null model

$$\text{logit}\{\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0))\} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i, \tag{15}$$

then the data will tend to reject H_0. In general the rejection of H_0 is due to the large value of the true γ, but it is still subject to the combined effect among parameters $\alpha_0, \boldsymbol{\alpha}', \beta_0, \boldsymbol{\beta}'$, and γ. To see this feature, we suppose that $\gamma \neq 0$ and that the normal distribution model is true as assumed in Section 4.2 with specific constraints,

$$\begin{aligned} \alpha_0 + \gamma\beta_0 &= 0, \\ \boldsymbol{\alpha} + \gamma\boldsymbol{\beta} &= \boldsymbol{0}. \end{aligned} \tag{16}$$

It follows that $\Pr(T_i = 1|\boldsymbol{X}_i, Y_i(0)) = 1/2$ no matter how large γ is, since the effect of $\boldsymbol{X}_i$ is offset by the effect of $Y_i(0)$. In such situation the two models (14) and (15) have almost the same ability to fit the data and decrease the difference between two competing likelihoods substantially. Although this kind of pathological parameter configurations occur rarely in practice, we will consider similar cases in simulation studies in Section 5.1.

A test based on a confidence interval for γ is asymptotically equivalent to the likelihood ratio test. Varieties of methods for estimating standard errors are available, such as methods based on the expected or observed information matrix or a bootstrap resampling method. However, it is often desirable to consider a test based on the likelihood ratio rather than the confidence interval for γ for accuracy in small sample sizes (McCullagh and Nelder 1989, p.471). Alternatively, one can obtain approximate confidence sets for γ^*'s so that the test $H_0 : \gamma = \gamma^*$ is accepted by the likelihood ratio test at certain significance level. On occasion, it may be helpful to refer the value of $\hat{\gamma}$ to see the direction of H_1 for further investigation.

4.4 Extension to Multivariate Responses

The present approach can be extended to consider multivariate counterfactual responses. A multivariate formulation of counterfactual models can be used when unit has k different responses. Assume that random vectors $\boldsymbol{Y}(0) = (Y_1(0), \ldots, Y_k(0))'$ and $\boldsymbol{Y}(1) = (Y_1(1), \ldots, Y_k(1))'$ be counterfactual outcomes for k different measurements of unit. The response $\boldsymbol{Y}(t)$ is observed if and only if $T = t$ for $t = 0$ or 1. Also, assume that $p \times k$ covariate matrix $\boldsymbol{X} = (X_1, \ldots, X_k)$ are obtained for each unit. Based on the similar argument with Section 2.2, the model $(\boldsymbol{Y}(0), \boldsymbol{Y}(1), T, \boldsymbol{X})$ is not identifiable from the observable data on $(\boldsymbol{Y}, T, \boldsymbol{X})$, where $\boldsymbol{Y} = \boldsymbol{Y}(1)T + \boldsymbol{Y}(0)(1-T)$. We wish to draw the inference on the multivariate treatment effects $\boldsymbol{\tau} = E\{\boldsymbol{Y}(1)\} - \{\boldsymbol{Y}(0)\}$, based on the observation $(\boldsymbol{Y}_i, T_i, \boldsymbol{X}_i)$ for $i = 1, \ldots n$, n random replications of $(\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{X})$. The version of strong ignorability for estimating the multivariate treatment effects $\boldsymbol{\tau}$ can be stated as follows:

Definition II *Treatment assignments are strongly ignorable given the observed covariate matrix* $\boldsymbol{X}$ *if, for each component* l*,*

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

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

Techniques of multivariate analysis may allow one to model the conditional likelihood for $\boldsymbol{Y}(0)$ given $\boldsymbol{X}$ and conditional likelihood for T given $(\boldsymbol{X}, \boldsymbol{Y}(0))$ under the assumption of the additive effect: $\boldsymbol{Y}(1) - \boldsymbol{Y}(0) = \boldsymbol{\tau}$. A similar likelihood ratio test could be constructed to assess the assumption of strong ignorability. Alternatively, we can use a simpler approach that requires only the marginal parametric specification. The hypothesis of interest can be stated as follows:

$$\bar{H}_0 : H_{0l} \text{ for all } l = 1, \ldots, k, \tag{17}$$

where $H_{0l} : (Y_l(0), Y_l(1)) \perp T \mid \boldsymbol{X}_l$ for $l = 1, \ldots, k$. Note that H_{0l} can be checked by the likelihood ratio test using data only from the component l and the method based on Section 4.1. Specifically, the hypothesis $\bar{H}_0$ is rejected at level α if at least one of the observed p-values from the marginal tests is below α/k. The test based on the Bonferroni inequality becomes conservative due to the lack of capturing the joint distribution of the observed k p-values. However, the test is robust against misspecified modeling of the joint structure on the multivariate distributions of $\boldsymbol{Y}(0)$ given $\boldsymbol{X}$ and distribution for T given $(\boldsymbol{X}, \boldsymbol{Y}(0))$. Perhaps the k p-values from the marginal likelihood ratio tests are themselves more informative than the test result for $\bar{H}_0$.

5 NUMERICAL ANALYSYS

5.1 Simulation Studies

We conducted simulation studies to examine the performance of the proposed test. 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 $\boldsymbol{\alpha}' = (-1,-1)$. The assignment probability is modeled through the logistic model

$$\text{logit}\{\Pr(T = 1|X_1, X_2, Y(0))\} = -1 + X_1 + X_2 + \gamma Y(0), \tag{18}$$

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 computed the likelihood ratio statistic (12) under the normal distribution model. $H_0 : \gamma = 0$ is rejected if the likelihood ratio exceed the upper 5% limit of the chi-squared distribution with one degree of freedom.

We compared the power of the test on $\gamma \in (-3.5, 3.5)$ based 5% level from 1000 replications, for a chosen value of $\tau = 1.5$. Figure 1 shows the power curves with sample sizes 200, 400 and 600 under the four configurations. The powers at $\gamma = 0$ are very close to 5%, indicating that the test has accurate type I error rates. Thus the large sample approximation of the critical point is quite satisfactory for all the sample sizes considered here. The power trajectories in all configurations are what is expected. The power that reject $H_0 : \gamma = 0$ increases as γ deviates from zero. The power curves for larger sample sizes uniformly dominate those for smaller sample sizes. The cases for the other values of τ exhibits virtually the same tendency as Figure 1, and we do not report these results here. For $\boldsymbol{\alpha}' = (-1,-1)$, the statistical power to detect $H_1 : \gamma = 1$ are very small since the test fail to capture the likelihood difference; The parameter configuration in this special case satisfies the constraint (16) and decrease the power as explained in Section 4.3. Excluding such pathological situation, the test seems to have proper power properties under the normal distribution model.

5.2 Robustness Considerations

To investigate the robustness for the proposed test under model misspecification, we fit the normal distribution model under t-distribution. In robust statistics, the use of t-distributions is one of the popular ways for describing the departure from normal distributions (Lange, Little and Taylor 1989). Variables for $Y(0)$ were generated from the t-distributions with mean $\alpha_0 + \boldsymbol{\alpha}'\boldsymbol{X}$, scale parameter 1 and degree of freedom ν ($\nu > 2$). The assignment probability follows a Bernoulli distribution fitted by the logistic model $\text{logit}(p) = -1 + X_1 + X_2$, where $\boldsymbol{X}' = (X_1, X_2)$ and X_1 and X_2 are generated as in Section 5.1.

Figure 2 shows the type I error rates under $H_0 : \gamma = 0$ but under misspecified t-distribution with the kurtosis $\kappa \in [0, 1.5]$, whose degrees of freedom ranging on $\nu \in [8, \infty]$. As expected, the type I error rates deviate from 5 % when the kurtosis gets larger. Clearly, larger sample sizes give more serious inflation in type I error. The simulation studies suggest that the rejection of H_0 based on the likelihood ratio test may possibly due to a wrong model in Assumption III even when H_0 actually holds.

5.3 Test Score Data from Japanese Elementary School

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 this section, we introduce the survey data in Katsuyama et al. (2006), and in the next section, we demonstrate the proposed test using the data. Their primary interest is to measure the effect of an English educational program applied to a Japanese elementary school, School A in Chiba prefecture. As the control group, they choose School B located in the same school district but did not introduce any English educational program. Katsuyama et al. (2006) provides descriptive summaries of their data set.

English test scores and some background information for students were collected at School A from Decenber 15th to 17th, 2003 and at School B from February 24th and 27th, 2004, respectively. Both data sets were in 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 an observed pretreatment covariates 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 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 in School B and $Y(1)$ denote the potential score for students in School A.

5.4 Data Analysis Using Test Score Data

A primary concern in the study of English educational effect based on Schools A and B is the possibility that, even after adjustment for the observed covariate $\boldsymbol{X}$, we still cannot compare the two schools. In other words, more background information may be necessary for adjustments. The assumption of strong ignorability states that the adjustment by $\boldsymbol{X}$ is sufficient for comparing these two schools.

Before applying the proposed likelihood analysis for testing the strong ignorability, one needs to impose Assumptions I-III. Since it is a common practice to approximate test scores by normal distribution, we fit the normal distribution model discussed in Section 4.2 to the test scores from the students. First, we check Assumptions I and II simultaneously. Assumptions I and II imply the relations

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

for all i. Here the error term ϵ_i has zero mean. Combining the preceding pair of displayed equations, we have

$$Y_i = \alpha_0 + \tau T_i + \boldsymbol{\alpha}'\boldsymbol{X}_i + \epsilon_i. \tag{19}$$

The linearity assumption in (19) can be checked by the residual plot, a popular tool for regression analyses. The residual plot displayed in Figure 3 shows no systematic departure from the linear assumption. Next, we check Assumption III. Note that Assumption III is equivalent to the normality assumption for ϵ_i under the linear model. We display the Q-Q plot in Figure 4 to check the normality of the residuals. The plot forms a approximately straight line except a few outlying observations in the tails. In general, we find no evidence for rejecting Assumptions I-III.

Now we test the assumption H_0 using the likelihood ratio test based on Assumptions I-III. The likelihood ratio statistic for this data set is $2(\hat{l}_1 - \hat{l}_0) = 0.00094$ and the corresponding p-value is 0.976. We do not find a statistical evidence for the departure from H_0. Assuming the strong ignorability, the $\boldsymbol{X}$-adjusted nonparametric estimator in (1) provides a valid point estimates for the program effect $\hat{\tau} = 1.525$ with 95% confidence interval (0.713,2.597). The interval is away from zero, and it provide the significant evidence for the positive educational effect for School A. Alternatively, it is possible to estimate the treatment effect using the maximized likelihood methods. We obtain estimates $\hat{\tau}_0 = 1.760$ under $H_0 : \gamma = 0$ and $\hat{\tau}_1 = 1.895$ under $H_0 \cup H_1$ respectively. Note that the distance between $\hat{\tau}$ and $\hat{\tau}_0$ are close to each other. The difference between the nonparametric and model-based estimates is expected to be small when the model is valid (White 1981). In addition to the p-value from the likelihood ratio statistics, the small difference suggests that the assumed normal disbribution model with the strong ignorability is a good approximation for the data set

6 CONCLUSION

In this paper, we have proposed a hypothesis test for assessing the assumption of strong ignorability. The logic of non-identifiability dilemma revealed that the assumption of strong ignorability is not testable without appropriate distributional assumptions. By imposing appropriate parametric assumptions (Assumptions I-III and logistic model) on the counterfactual models, we have shown that the strong ignorability can be empirically checked by the likelihood ratio test. Permitting flexible parametric families based on the GLM, we have emphasized that the likelihood ratio test is very general and applicable to many different types of data. We have derived a simple algorithm when the counterfactual models follows a normal distribution. Also, extension for multivariate counterfactual models is briefly discussed. Simulation studies indicate that the proposed test has regular power properties if the Assumptons I-III are true. On the other hand, if the parametric assumption is wrong, the test can fail to provide correct type I error rates.

To identify the assumption of strong ignorability, Assumption I (additive treatment effect model) plays a key role in the likelihood construction. However, it is not clear that Assumption I is valid in many application studies, though it is easy to generate such types of data in simulations. Although some empirical procedures for checking Assumption I and II simultaneously is provided in Section 5, we do not develop a saisfactory method for checking Assumption I. There is a need for further research on the applicability of Assumption I in real data.

We showed how the assumption of strong ignorability can be empirically assessed under the GLM type framework with a simple example for the normal distribution case. The development of algorithms for the other types of models, such as exponential, gamma and Poisson-type distribution is mandated before the proposed method can be used more generally. We leave this as a topic for further investigation.

Researchers may perform the likelihood ratio test to check H_0 before making a statistical inference on the treatment effect. If the test reject H_0, the data may indicate either the violation of the strong ignorability, or perhaps wrong distributional assumptions in Assumptions I-III. The small p-value of the test results may give us insight on the mechanism of the treatment assignment that may deserve further investigation. Specifically, the existence of unobserved confounding variables may be suspected for such a case.

Finally, we note that the proposed method and sensitivity analysis play different roles in the assessment for the assumption of strong ignorability. The porposed test checks the validity of the assumption of strong ignorability while the sensitivity analysis investigates the robustness of the treatment effect under the violation from the strong ignorability. Combined use of these techniquues may enhance the integrity of a statistical inference for the average treatment effect from observational data.

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] 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.

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

[7] Hodges, J. and Lehmann, E. (1963), "Estimates of Location Based on Rank Tests," *Annals of Mathematical Statistics*, 34, 598-611.

[8] Hong, G. and Raudenbush, S. W. (2006), "Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data," *Journal of the American Statistical Association*, 101, 901-910.

[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] Klein, J. P. and Moeschberger, M. L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer.

[11] Lange K. L., Little, R. J. A. and Taylor, J. M. G (1989), "Robust Statistical Modeling Using the t-Distribution," *Journal of the American Statistical Association*, 84, 881-896.

[12] Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

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

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

[15] Pearl, J. (2001), "Causal Inference in the Health Sciences: A Conceptual Introduction," *Health Services and Outcomes Research Methodology*, 2, 189-220.

[16] Rosenbaum, P. (1984), "From Association to Causation in Observational Studies: The Role of Strongly Ignorable Treatment Assignment," *Journal of the American Statistical Association*, 79, 41-48.

[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] White, H. (1981), "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76, 1065-1073.

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

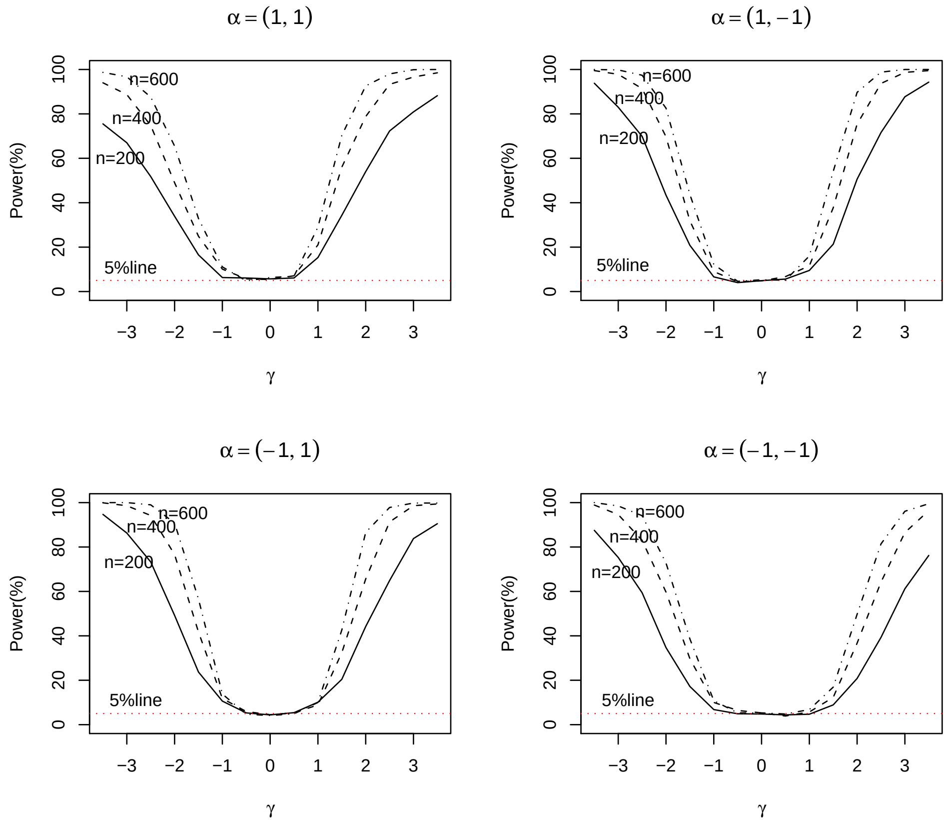

Figure 1: Power curves for sample sizes $n = 200$ (Solid Line), $n = 400$ (Dashed Line) and $n = 600$ (Dashed and Dotted Line) based on 1000 replicates. Potential outcomes $Y(0)$ are generated from the normal distributions with mean $1 + \boldsymbol{\alpha}'X$ and variance 1. The assignment variable T follows a Bernoulli random variable with the probability p modeled by the logit model $\text{logit}(p) = -1 + X_1 + X_2$.

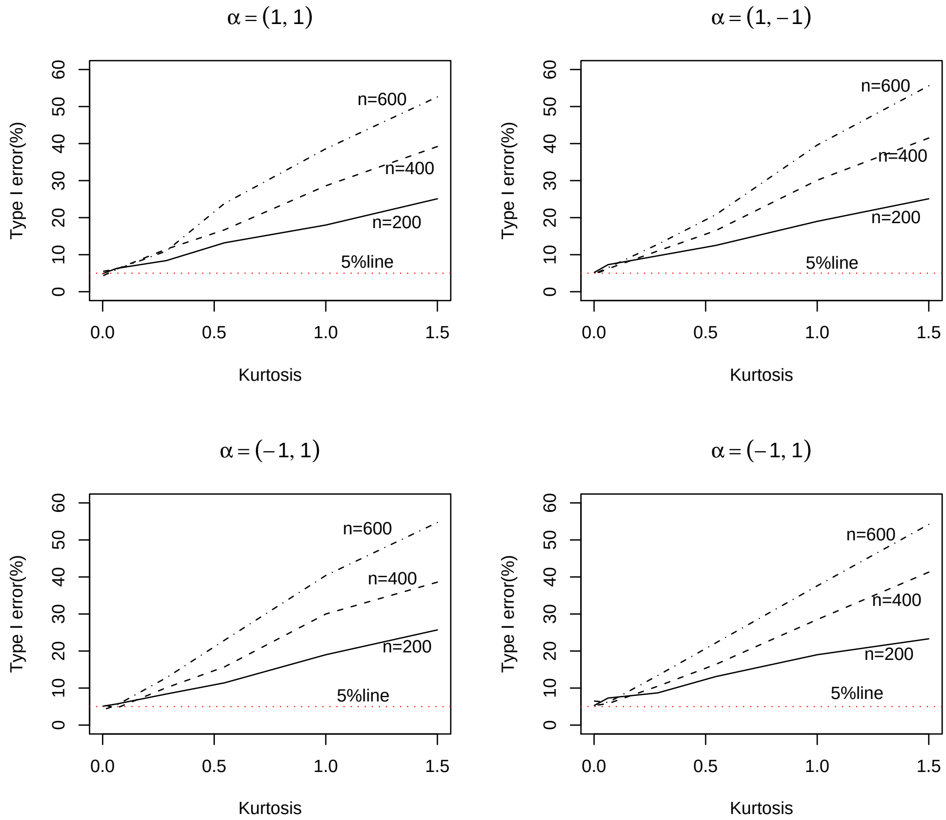

Figure 2: Type I error rates for sample sizes $n = 200$ (Solid Line), $n = 400$ (Dashed Line) and $n = 600$ (Dashed and Dotted Line) based on 1000 replicates. Potential outcomes $Y(0)$ are generated from the misspecified t-distributions with mean $1 + \boldsymbol{\alpha}'\boldsymbol{X}$, scale parameter 1 and the kurtosis $\nu \in [0, 1.5]$. The assignment variable T follows a Bernoulli distribution with the probability p modeled by the logit model $\text{logit}(p) = -1 + X_1 + X_2$.

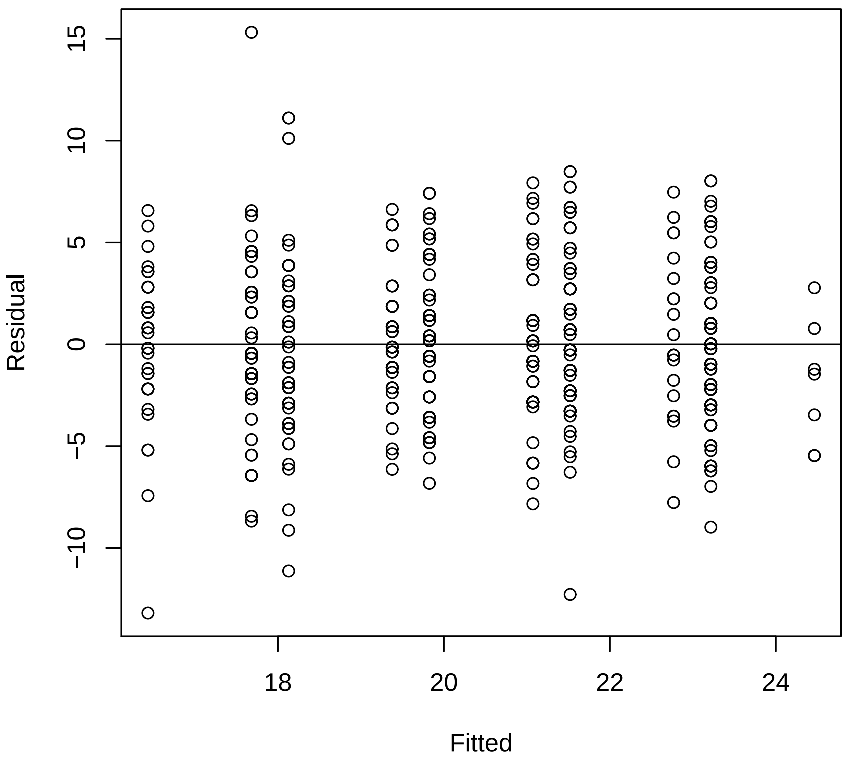

Figure 3: The residual plot for checking the linear assumption for $Y_i = \alpha_0 + \tau T_i + \boldsymbol{\alpha}' \boldsymbol{X}_i + \epsilon_i$. Variables Y_i and $\boldsymbol{X}_i$ are the test score and student backgrounds respectively for a student from Japanese elementary School A ($T_i = 1$) 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 $\hat{\alpha}_0 + \hat{\tau} T_i + \hat{\boldsymbol{\alpha}}' \boldsymbol{X}_i$.

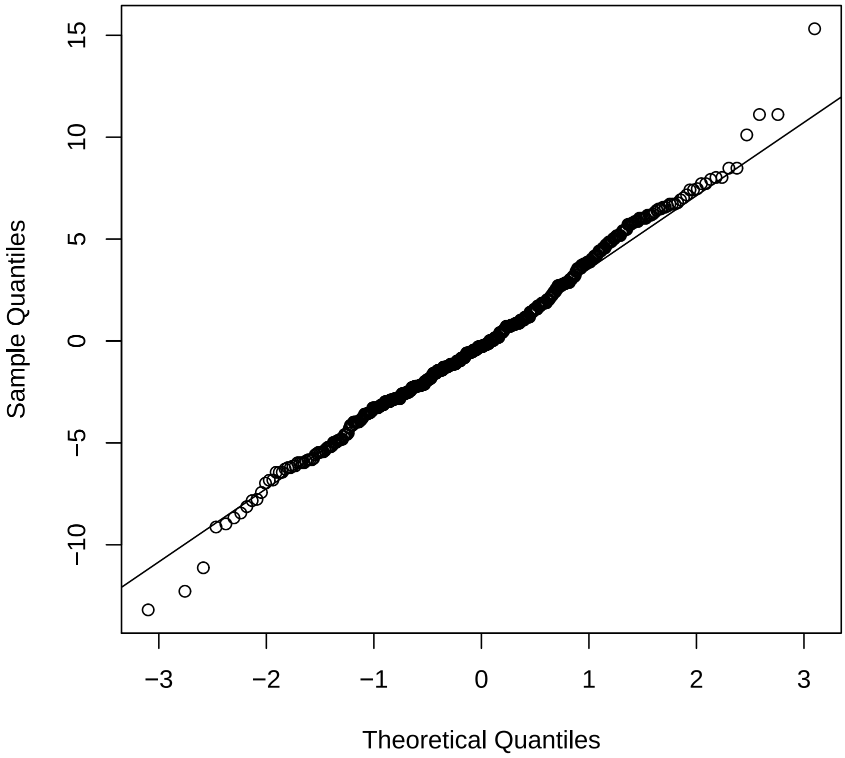

Figure 4: Q-Q plot for checking the normality assumption for the error distribution in $Y_i = \alpha_0 + \tau T_i + \boldsymbol{\alpha}' \boldsymbol{X}_i + \epsilon_i$. Variables Y_i and $\boldsymbol{X}_i$ are the test score and student backgrounds respectively for a student from Japanese elementary School A ($T_i = 1$) or School B ($T_i = 0$). The straight line passes through the first and third quartiles.