DISCUSSION PAPER SERIES

IZA DP No. 1873

Sensitivity of Propensity Score Methods to the Specifications

Zhong Zhao

December 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Sensitivity of Propensity Score Methods to the Specifications

Zhong Zhao
IZA Bonn

Discussion Paper No. 1873
December 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 1873
December 2005

ABSTRACT

Sensitivity of Propensity Score Methods to the Specifications*

Propensity score matching estimators have two advantages. One is that they overcome the curse of dimensionality of covariate matching, and the other is that they are nonparametric. However, the propensity score is usually unknown and needs to be estimated. If we estimate it nonparametrically, we are incurring the curse-of-dimensionality problem we are trying to avoid. If we estimate it parametrically, how sensitive the estimated treatment effects are to the specifications of the propensity score becomes an important question. In this paper, we study this issue. First, we use a Monte Carlo experimental method to investigate the sensitivity issue under the unconfoundedness assumption. We find that the estimates are not sensitive to the specifications. Next, we provide some theoretical justifications, using the insight from Rosenbaum and Rubin (1983) that any score finer than the propensity score is a balancing score. Then, we reconcile our finding with the finding in Smith and Todd (2005) that, if the unconfoundedness assumption fails, the matching results can be sensitive. However, failure of the unconfoundedness assumption will not necessarily result in sensitive estimates. Matching estimators can be speciously robust in the sense that the treatment effects are consistently overestimated or underestimated. Sensitivity checks applied in empirical studies are helpful in eliminating sensitive cases, but in general, it cannot help to solve the fundamental problem that the matching assumptions are inherently untestable. Last, our results suggest that including irrelevant variables in the propensity score will not bias the results, but overspecifying it (e.g., adding unnecessary nonlinear terms) probably will.

JEL Classification: C21, C14, C15, C16, C52

Keywords: sensitivity, propensity score, matching, causal model, Monte Carlo

Corresponding author:

Zhong Zhao
IZA
P.O. Box 7240
53072 Bonn
Germany
Email: zhao@iza.org

* I am grateful to Robert Moffitt and Geert Ridder for their valuable discussions, and to other seminar participants at Johns Hopkins, Washington University in St. Louis, and the Institute for the Study of Labor for their helpful comments. This research is partially supported by the Scientific Research Foundation for Returned Overseas Chinese Scholars, State Education Ministry, China. All errors are mine.

I. Introduction

Estimating treatment effects usually is plagued by the infamous selection bias problem (see Heckman 1979). Matching, which is a method for selecting comparison observations to match treated observations with similar covariates, has been becoming a popular procedure to correct selection bias under the assumption of unconfoundedness, which means that the selection bias is only due to observed variables. [1] [2] This assumption is also known as that of selection on observables or conditional independence. Imbens (2004) provides an excellent survey on estimating treatment effects under the unconfoundedness assumption.

Using covariate matching to correct the bias due to observables is intuitive, since the source of the bias is the difference of observables between the treated group and the comparison group. Matching on covariates by definition will remove this difference and hence the bias (see Rubin 1980).

The most attractive feature of matching, compared with regression-type estimators such as that of Barnow, Cain, and Golderger (1980), is its nonparametric nature. Matching neither imposes functional form restrictions such as linearity on the outcome equations nor assumes a homogeneous treatment effect across the population. Both assumptions are usually unjustified either by economic theory or by the data.

When there are many covariates, it is impractical to match directly on covariates because of the curse of dimensionality. Taking the study of the Comprehensive

[1] A covariate is defined as any variable such that its value is not affected by the treatment; e.g., sex is a covariate, but wage is not.

[2] Recent papers in this field by economists include Abadie and Imbens (2005, forthcoming), Angrist (1998), Dehejia and Wahba (1999, 2002), Hahn (1998), Heckman, Ichimura, and Todd (1997, 1998), Heckman, Ichimura, Smith, and Todd (1998), Imbens (2000), Lechner (2002), and Smith and Todd (2005). Also, see the symposium on matching estimators in the February 2004 issue of the *Review of Economics and Statistics*.

Employment and Training Act by Westat (1981) as an example, for controlling only 12 covariates, the covariate matching scheme of Westat led to more than 6 million cells. Since the number of observations is far less than 6 million, most of the cells are empty and it is very hard to find a good match on all 12 covariates.

Since the celebrated result of Rosenbaum and Rubin (1983), an attractive way to overcome the curse of dimensionality has been matching by propensity score, $p(x)$.[3] However, to implement propensity score matching methods in empirical studies, several issues need to be resolved. The first one is the unconfoundedness assumption. How sensitive the result is to this assumption is the topic of Imbens (2003). The second one is the common-support condition. Crump, Hotz, Imbens, and Mitnik (2004) study this issue. The third one is that the propensity score is usually unknown and needs to be estimated. Our paper focuses on this issue.

When the propensity score is unknown, if we estimate it nonparametrically, we are incurring the curse-of-dimensionality problem that we are trying to avoid. If we estimate it parametrically, the nonparametric advantage of matching estimators may be lost. How sensitive are the estimated treatment effects to the parametric specifications of the propensity score becomes an important question.

The findings in the literature are mixed. Some empirical studies, such as reanalyzing the National Supported Work Demonstration by Dehejia and Wahba (1999), evaluating antipoverty program in Argentina by Jalan and Ravallion (2003), investigating teenage out-of-wedlock childbearing by Levine and Painter (2003), and studying the

[3] The propensity score is the probability of being treated conditional on the covariate X, i.e., $p(x) = prob(T = 1 \mid X = x)$.

labor market outcomes of welfare reform by Heinrich, Mueser, and Troske (2005), suggest that the specification of the propensity score is not important.

However, an important paper by Smith and Todd (2005), which also studies the National Supported Work Demonstration as in Dehejia and Wahba (1999), argues otherwise.

In this paper, we investigate whether the propensity score matching results are sensitive to the specification of the propensity score or not. We restrict our attention to the binary treatment case and do not consider multiple treatment scenarios as in Imbens (2000) and Lechner (2002).

We investigate the sensitivity issue through Monte Carlo experiments. It is well known that when using the probit model to estimate a binary choice model with nonnormal error term, the estimation can be biased and inconsistent. The most vulnerable cases are bimodal and heteroskedastic error terms (Horowitz, 1993).

The major finding from our simulations is that the coefficients of the propensity score are indeed poorly estimated in misspecified models with bimodal and heteroskedastic error terms – which is consistent with Horowitz (1993) – but these poorly estimated propensity scores have little influence on the estimates of the treatment effects. In fact, the treatment effects estimated from the misspecified models are nearly as good as the ones from the correct models.

We provide two justifications for this insensitivity observed in the empirical literature and in our simulations. The first justification is based on the semiparametric and nonparametric literature on binary choice models. The second justification draws insight from Rosenbaum and Rubin (1983). They show that any function of the propensity score

is a balancing score, and that controlling for any balancing score is sufficient to remove selection bias caused by observables under matching assumptions.[4] Even if we estimate a misspecified propensity score, it is possible that the "wrong" propensity score still belongs to the class of balancing scores.

Our finding seems inconsistent with the finding in Smith and Todd (2005). Following the work of LaLonde (1986), Smith and Todd combine experimental data from the National Supported Work (NSW) Demonstration and survey data from the Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID) to evaluate the performance of propensity score matching estimators. They find that the results are sensitive to the propensity score specifications.

The data constructed in this way is very likely to violate the unconfoundedness assumption.[5] We conciliate our insensitive finding with the sensitive finding in Smith and Todd (2005) by noting that if the unconfoundedness condition fails, the matching results can be sensitive, which is shown in our Monte Carlo experiments.[6]

Sensitivity checks such as the one in Dehejia (2005) are helpful to eliminate sensitive cases.

Nonetheless, failure of the unconfoundedness assumption does not necessarily lead to sensitive results. We find in some cases that even if the unconfoundedness assumption fails, the results can be speciously robust, in the sense that the propensity score matching estimator constantly overestimates or underestimate the treatment effect.

[4] We define balancing score formally in section 2.
[5] In fact, constructing data sets and making them behave like nonexperimental data is exactly the purpose of Lalonde (1986).
[6] Imbens (2003) also analyzes sensitivity to the unconfoundedness assumption. We will compare our approach here with his in section 5.

The sensitivity check in general cannot help us to solve the fundamental problem that the matching assumptions are untestable.

We also consider the specification of the index function in the propensity score. Our result suggests that including irrelevant variables in the propensity score will not cause bias, but overspecifying it (e.g., adding unnecessary nonlinear terms) will probably bias the results.

The rest of the paper is as follows: Section 2 sets up the model using the potential outcome framework. Section 3 applies Monte Carlo experiments to investigate the consequences of misspecifying the propensity score under the unconfoundedness assumption. In that section, we also provide some justifications for the insensitivity observed in the empirical studies and in our Monte Carlo experiments. Section 4 discusses specification of the index function in the propensity score. Section 5 reconciles our finding with Smith and Todd's (2005). Section 6 concludes this paper.

II. Model Setup

In the potential outcome framework, such as in Rubin (1974), each individual has two potential responses (Y_{0i}, Y_{1i}) for a treatment, such as job training, education, or a welfare program. Y_{1i} is the outcome if individual i is treated, and Y_{0i} is the outcome if individual i is not treated. Let $T_i = 1$ indicate that individual i is treated and $T_i = 0$ indicate otherwise. With (Y_{0i}, Y_{1i}) we can define different treatment effects, such as those in Heckman and Vytlacil (2005), as follows:

$\Delta_i = Y_{1i} - Y_{0i}$ Treatment effect for individual i

$\Delta_{ATE} = E[\Delta_i]$ Average treatment effect for the population (ATE)

$\Delta_S = E[\Delta_i \mid i \in S]$ Average treatment effect for the subpopulation S

When $S = \{i : T_i = 1\}$, Δ_S is the treatment effect on the treated (TT), denoted as Δ_{TT}.

That the selection bias is only due to observables is formally characterized by the following two assumptions:

M-1: $(Y_0, Y_1) \perp\!\!\!\perp T \mid X$ Unconfoundedness assumption

M-2: $0 < prob(T = 1 \mid X) < 1$ Common-support assumption

where $\perp\!\!\!\perp$ is the notation for statistical independence as in Dawid (1979), and *prob* stands for probability.

Under M-1 and M-2,

$$\Delta_{TT} = E_{x|T=1}\{E[Y_1 \mid T = 1, X = x] - E[Y_0 \mid T = 1, X = x]\}$$

$$= E_{x|T=1}\{E[Y_1 \mid T = 1, X = x] - E[Y_0 \mid T = 0, X = x]\} \quad (1)$$

Unbiased estimates of $E[Y_1 \mid T = 1, X = x]$ and $E[Y_0 \mid T = 0, X = x]$ can be obtained from the data, and hence so can Δ_{TT}. This is also true for Δ_{ATE} and for other Δ_S.

From Rosenbaum and Rubin (1983), we have the following definition.

***Definition* 1** (Propensity score, Rosenbaum and Rubin 1983): A propensity score $p(x)$ is the conditional probability that an observation is in the treated group, conditioning on the observed covariates X, i.e., $p(x) = prob(T = 1 \mid X = x)$.

Rosenbaum and Rubin (1983) prove that M-1 and M-2 imply

P-1: $(Y_0, Y_1) \perp\!\!\!\perp T \mid p(X)$, and

P-2: $0 < prob(T = 1 \mid p(X)) < 1$.

It follows from P-1 and P-2 that

$$\Delta_{TT} = E_{p|T=1}\{E[Y_1 \mid T=1, p(X)=p] - E[Y_0 \mid T=1, p(X)=p]\}$$

$$= E_{p|T=1}\{E[Y_1 \mid T=1, p(X)=p] - E[Y_0 \mid T=0, p(X)=p]\} \qquad (2)$$

Unbiased estimates of $E[Y_1 \mid T=1, p(X)=p]$ and $E[Y_0 \mid T=0, p(X)=p]$ can be obtained if $p(X)$ is known. The advantage of formula (2) over formula (1) is that instead of controlling for a high-dimensional vector of X, formula (2) only needs to control for a scalar $p(X)$.

It is important to note that covariate X and propensity score $p(X)$ both belong to a class called balancing scores (Rosenbaum and Rubin 1983). Formally, we have:

***Definition* 2** (Balancing score, Rosenbaum and Rubin 1983): A balancing score $b(x)$ is a function of the observed covariate X such that the conditional distribution of X given $b(x)$ is the same for treated and comparison units, i.e., $X \perp\!\!\!\perp T \mid b(x)$.

Theorem (Rosenbaum and Rubin 1983): Let $b(x)$ be a function of X. Then $b(x)$ is a balancing score if and only if $b(x)$ is finer than the propensity score $p(x)$ in the sense that $p(x) = f\{b(x)\}$ for some function f.

As seen in Figure 1, covariate X is the finest balancing score, and the propensity score $p(X)$ is the coarsest one. In between, there are infinitely many balancing scores. The propensity score is not the only scalar balancing score. In theory, controlling for any balancing score is sufficient to correct the selection bias due to observables.

This result can be easily understood in the context of law of iterated expectations. Since $p(x)$ is a function of $b(x)$, knowing $b(x)$ implies knowing $p(x)$, so if outcome and treatment status are independent conditional on a smaller information set ($p(x)$), they will be also independent conditional on a larger information set ($b(x)$), i.e.,

$$E[Y_0 \mid T = 1, p(X) = p] = E[Y_0 \mid T = 0, p(X) = p]$$

$$\Rightarrow E[Y_0 \mid T = 1, b(X) = b] = E[Y_0 \mid T = 0, b(X) = b]$$

Heckman and Navarro-Lozano (2004) formally discuss this point. They define relevant information set and minimal relevant information set. These two sets bear some similarity to the balancing score and the coarsest balancing score in Rosenbaum and Rubin (1983), respectively. Nonetheless, they have important subtle differences. The minimal relevant set is the minimum amount of information that makes the unconfoundedness assumption hold, whereas the balancing score and coarse balancing score are not related to the unconfoundedness assumption. However, if a covariate X is a relevant information set, then any balancing score $b(x)$ is a relevant information set; furthermore, the propensity score $p(x)$ is a minimum relevant information set.

III. Specification of the Error Term in the Propensity Score

If the propensity score $p(x)$ is unknown, which is the case for most applications, then in order to apply formula (2), we need to estimate the propensity score. If we estimate it nonparametrically, we are involved in the curse-of-dimensionality problem we are trying to avoid. If we estimate it parametrically, sensitivity to the parametric specifications of the propensity score becomes an important issue.

It is well known that estimated coefficients from a probit or logit model can be far away from the true coefficients in a binary choice model if the error term in the true model follows a bimodal or heteroskedastic distribution (see Figure 2, reproduced from Horowitz 1993).

In this section, we investigate the question: how robust are the estimated treatment effects to the specification of the propensity score? Our approach is Monte Carlo experiment.[7] The idea is that we simulate the true propensity score using different error-term distributions, including bimodal and heteroskedastic distributions, as discussed in Horowitz (1993). Then we estimate the propensity score by probit, logit, and linear probability model (LPM), so some of the estimated propensity scores are from misspecified models. After we estimate the propensity score, we calculate the treatment effect by matching on the estimated score. Since we know the true treatment effect, we can assess the sensitivity to the specification of the propensity score.

Monte Carlo Experiment Setup. The Monte Carlo experiment is designed to investigate sensitivity to the propensity score specification based on the potential outcome model. We use linear specification in the outcome equations and in the propensity score:

$Y_1 = \alpha_{10} + \alpha_{11}X_1 + \alpha_{12}X_2 + \varepsilon_1 = X\alpha_1 + \varepsilon_1$	Outcome in treated state
$Y_0 = \alpha_{00} + \alpha_{01}X_1 + \alpha_{02}X_2 + \varepsilon_0 = X\alpha_0 + \varepsilon_0$	Outcome in untreated state
$T^* = \beta_0 + g\beta_1 X_1 + gr\beta_2 X_2 + \mu = X\beta + \mu$	Latent index function
$T = I(T^* > 0)$, where $I(\cdot)$ is the indicator function	Treatment indicator

We allow correlation between the error terms ε_0 and ε_1 in the outcome equations, but do not allow correlation between the error term in the selection equation and the error

[7] There is a small literature using Monte Carlo to evaluate matching estimators. Gu and Rosenbaum (1993) study different matching algorithms, Flörich (2004) compares matching and weighting estimators, and Zhao (2004) investigates different matching metrics and compares propensity score matching and covariate matching.

terms in the outcome equations, i.e., $\text{cov}(\mu, \varepsilon_0) = 0$ and $\text{cov}(\mu, \varepsilon_1) = 0$. So unconfoundedness is satisfied.

The error terms in the outcome equations, ε_0 and ε_1, follow a bivariate standard normal distribution.

Following Horowitz (1993), we consider six distributions of error term in propensity score:

(1) μ has the standard normal distribution, N(0,1).

(2) μ has a logistic distribution.

(3) μ is a 50-50 mixture of two normal distributions N(3,1) and N(−3,1).

(4) $\mu = |1 + 2(X\beta)|v$, where v has a logistic distribution.

(5) $\mu = 0.25[1 + 2(X\beta)^2 + (X\beta)^4]v$, where v has a logistic distribution.

(6) μ has a normal distribution, N(1, $1 + 0.2(X_1^2 + X_2^2)$).

Probit is the correct model for (1), and logit is the correct model for (2). (3) has a bimodal distribution. (4) and (5) are heteroskedastic error terms. We estimate the propensity score by probit, logit, and LPM.

Results. In Table 1, one *X* has a χ^2 distribution with one degree of freedom, and the other *X* is a mixture of two normal distributions.

Panel A reports estimates for a binary choice model. The results reported here are the ratios of coefficients of the *X*'s, since a binary choice model can be only identified up to scale. The coefficients of the propensity score are poorly estimated in the misspecified models when the error term is bimodal or heteroskedastic, which is consistent with Horowitz (1993).

Panel B reports the estimated treatment effects from matching with replacement. The results reported here are the ratios of estimated treatment effects to the true treatment effect. The true value is one. As shown in Panel B, these poor estimates have little influence on the estimates of the treatment effects. In fact, the treatment effects estimated from the misspecified models are comparable to the ones from the correct models. The choice of estimators (probit, logit, and LPM) for the propensity score also has little influence on the matching results.

When the sample size is increased to 2,000 (Table 2), the basic findings remain the same. We also consider the case that both *X*'s are normally distributed, and obtain similar results (see Table A1 and Table A2).

When matching without replacement (Panel C), we see an increase in bias but a decease in standard error. This is consistent with the theory as well as empirical evidence, such as those in Mueser, Troske, and Gorislavsky (2005).

Justifications. We provide two justifications for the insensitivity findings in the empirical studies and in our simulations.

When the propensity score is used as a device to match observations, any order-preserving transformation, such as a monotonic transformation, of the propensity score is sufficient to accomplish the task. It is well known (e.g., Goldberger 1980, Ruud 1983, 1986), that even if the distribution is misspecified, the coefficients of the propensity score can still be consistently estimated up to an unknown scale under mild conditions. Chung and Goldberger (1984) show that the estimates from least-squares linear regression are proportional to the estimates from the correctly specified probability model under weak

distributional assumptions. It turns out that this proportionality is all we need for the order-preserving property to hold.[8]

***Proposition*:** Let $G(\cdot)$ and $G^*(\cdot)$ be strictly monotonic increasing functions. Then $G(x_i\beta') > G(x_j\beta') \Leftrightarrow G^*(x_i\beta^{*\prime}) > G^*(x_j\beta^{*\prime})$ if $\beta = a\beta^*$, except for the intercept, where a is a possibly unknown positive scalar.

Another justification is based on the balancing-score concept in Rosenbaum and Rubin (1983). As defined in Section 2, any score finer than the propensity score is a balancing score, and controlling for any balancing score is sufficient to remove the selection bias under the unconfoundedness assumption.

As shown in Figure 2, for a bimodal error term, the estimated propensity score is an order-preserving transformation of the correct propensity score, so theoretically it should have little effect on the matching.

For a heteroskedastic error term, the score from a probit model is finer than the true score, in the sense that there is a many-to-one relation from the heteroskedastic propensity score to the probit propensity score in Figure 2. Matching on the misspecified score can still balance the covariate between treated and comparison groups.

It also worth noting that the heteroskedastic propensity score has smaller support. This should mitigate to some degree the common-support problem often encountered in empirical research.

We should emphasize that we do not claim that the propensity score estimated from a probit or logit model is always a balancing score under any circumstances.

[8] On the contrary, if propensity score is used in a weighting estimators, as in Hirano, Imbens, and Ridder (2003), then it needs to be the correct one, and an order-preserving transformation is not enough.

Instead, our goal is to provide some intuition and explanations for the insensitivity findings in the literature and in our Monte Carlo experiments.

IV. Specification of the Index Function in the Propensity Score

The other aspect of the propensity score specification is the functional form of the index term. We are considering two cases here. One is adding an irrelevant variable to the propensity score, and the other is adding a nonlinear term.

Including Irrelevant Variables. Suppose $(Y_0, Y_1) \perp\!\!\!\perp T \mid X$, and we have additional information on Z. Instead of matching on the propensity score $p(x)$, we match on the joint conditional probability of X and Z, i.e., match on $p(x,x) = prob(T=1 \mid X=x, Z=z)$. We have

$$\begin{aligned} p(x) &= prob(T=1 \mid x) \\ &= \frac{prob(T=1, x)}{f(x)} \\ &= \frac{\int prob(T=1, x, z)dz}{\int f(x,z)dz} \\ &= \frac{\int p(x,z) f(x,z)dz}{\int f(x,z)dz} \end{aligned}$$

where $f(x,z)$ is the joint density function of X and Z, and $f(x)$ is the marginal density function of X. So $p(x,z)$ is finer than $p(x)$, and we have $(Y_0, Y_1) \perp\!\!\!\perp T \mid p(X,Z)$. Matching on $p(x,z)$ will balance X and will remove bias.

How does the above argument turn out in small sample? To answer this question, we perform simulations for small samples of 1,000 observations. Table 3 is results from including irrelevant variables. From this table, it is seen that adding irrelevant variables

does not affect the bias. Comparing Table 3 with Table 1 (both have 1,000 observations), we see that, both in bias and in standard error, the results in these two tables are very close to each other.

Nonlinearity. Let us start with two specifications of the propensity score index function:

(1) $T^* = \beta_0 + \beta_1 X + \varepsilon$

(2) $T^* = h(X) + \varepsilon$

In (1) we assume that T^* and X have a linear relationship, and in (2) we assume that $h(\cdot)$ is a nonlinear function of X. Suppose $h(\cdot)$ is a monotonic function; then the propensity score estimated from one specification is an order-preserving transformation of the propensity score estimated from the other specification. In this situation, the nonlinearity is not important.

Figure 3 illustrates the situation when $h(\cdot)$ is not a monotonic function.

Assume (1) is wrong and (2) is right, but we estimate (1). In this scenario, we fail to match point A and point B. There will be an efficiency loss, since fewer observations are used, but that will not affect the unbiasedness. In the terminology of Rosenbaum and Rubin (1983), (1) is finer than (2), and (2) is the propensity score, so matching on either of them will result in unbiased estimates.

Now suppose (1) is right and (2) is wrong, but we estimate (2). Under our estimation, we match point A and point B together, which is a mismatch and could cause bias. In this case, (1) is finer than (2), but (1) is the propensity score. Any score that is not finer than (1) is not a balancing score, so (2) is not a balancing score, and matching on (2) will cause bias.

This example suggests that modeling nonlinearity (overspecifying the propensity score) could be counterproductive in some cases.

In practice, this point may not be important. If the true model does not have a nonlinear term, the estimate of the coefficient of the nonlinear term will likely be close to zero, and will have little effect on the matching result.

V. Reconciling Findings in Smith and Todd (2005)

The influential paper by Smith and Todd (2005) clearly documents that the estimated treatment effects are very sensitive to the specification of the propensity score. It is important to reconcile our result with theirs.

Smith and Todd (2005) follow the approach starting from LaLonde (1986). They combine the experimental data from the NSW data with the survey data from CPS and from PSID to evaluate the performance of the propensity score estimator. Data sets constructed in this way are likely to violate the unconfoundedness assumption, and that can lead to very sensitive estimates.

Imbens (2003) analyzes the sensitivity to the unconfoundedness assumption. However, his objective is to trace out the importance of unobserved variables in different data settings, such as experimental data, a restricted subset of the survey data, and the unrestricted survey data. He finds that results based on experimental data are very robust, and that in that case controlling for the unobserved variables is not important; he also finds that the results based on a properly selected subset of the survey data are more robust than those based on the whole data set.

Our analysis of the failure of the unconfoundedness assumption has a different goal than the one in Imbens (2003). Our main goal is to show that if the unconfoundedness assumption fails, we can observe the sensitive results in Smith and Todd (2005).

In Table 4, we allow the error term in the propensity score equation to be correlated with the error term in the outcome equations. In this setup, the unconfoundedness condition fails.

Panel A gives the estimation results for the binary choice model. They are similar to the results in the previous tables.

From Panel B and Panel C, it is clear that the results are sensitive to the specification of the propensity score when the unconfoundedness assumption is violated. The estimated treatment effect can be either larger or smaller than the true treatment effect, which is consistent with the sensitivity finding in Smith and Todd (2005).

The choice of probit, logit, or PLM also has a big effect now.

When the unconfoundedness assumption fails, the distribution of the error terms in the propensity score has a direct influence on the treatment outcome, so it is possible that different specifications will produce different estimated treatment effects.

Conceptually, the sensitivity findings also consistent with the important argument in Heckman and Hotz (1989) that different nonexperimental estimators impose different assumptions on unobservables; hence the estimates can be different.[9]

Table 4 show that the sensitive estimates are an indication of the failure of the unconfoundedness assumption. It is important to carry out sensitivity checks, as pointed

[9] In our case, we impose distribution assumptions on the propensity score.

out by Dehejia (2005). If the results are sensitive, one should use other estimators instead of propensity score matching methods.

Nonetheless, failure of the unconfoundedness assumption does not necessarily result in sensitivity. From Table 5 and Table 6, we find that even if the unconfoundedness assumption fails, for certain variance-covariance structures of the error terms, the results can be speciously robust in the sense that the propensity score matching estimator constantly overestimates or underestimates the treatment effect. [10]

Sensitivity checking is helpful for eliminating sensitive cases, but in general it cannot help us to solve the fundamental problem that the matching assumptions are inherently untestable.

VI. Conclusions

If the propensity score is unknown, which is the case for most applications, then in order to apply propensity score matching methods, we need to estimate it. If we estimate it nonparametrically, we incur the curse-of-dimensionality problem we are trying to avoid. If we estimate it parametrically, sensitivity to the parametric specifications of the propensity score becomes an important issue.

The major finding from our simulations is that though the coefficients of the propensity score are poorly estimated in the misspecified models in cases of bimodal and heteroskedastic error terms, these poor estimates have little influence on the estimated treatment effects if the matching assumptions are satisfied.

[10] In our Monte Carlo experiments we have experienced more robust cases than sensitive cases. However, our experience is limited, and it is dangerous to draw any general conclusion.

Our Monte Carlo results show that if the unconfoundedness condition fails, the results can be sensitive, which reconciles our finding with the findings in Smith and Todd (2005).

Nonetheless, failure of the unconfoundedness assumption does not necessarily lead to sensitive results. We find cases where even if the unconfoundedness assumption fails, the results can be speciously robust in the sense that the propensity score matching estimator constantly overestimates or underestimates the treatment effect under certain variance-covariance values for the error terms. A sensitivity check is helpful for eliminating sensitive cases, but in general, it cannot help us to solve the fundamental problem that the matching assumptions are untestable.

Our study also suggests that including irrelevant variables in the propensity score will not cause bias, but overspecifying it (e.g., adding unnecessary nonlinear terms) will probably bias the results.

References

Abadie, Albert and Guido W. Imbens (2002), "Large Sample Properties of Matching Estimators for Average Treatment Effects," forthcoming in Econometrica.

Abadie, Albert and Guido W. Imbens (2005), "On the Failure of the Bootstrap for Matching Estimators," unpublished manuscript.

Angrist, Joshua. D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," Econometrica, Vol. 66 (March 1998), 249–288.

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger (1980), "Issues in the Analysis of Selection Bias," Evaluation Studies Review Annual 5 (1980), edited by Stromsdorfer, E. and Farkas, G.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar Mitnik (2005), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," unpublished manuscript.

Chung, C. and A. S. Goldberger (1984), "Proportional Projections in Limited Dependent Variable Models," Econometrica 52 (1894), 531–534.

Dawid, A. Philip (1979), "Conditional Independence in Statistical Theory," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 41, No. 1 (1979), 1–31.

Dehejia, Rajeev H. and Sadek Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," Journal of the American Statistical Association 94 (December 1999), 1053–1062.

_____ (2002), "Propensity Score Matching Methods for Non-Experimental Causal

Studies," Review of Economics and Statistics 84 (February 2002), 151–175.

Dehejia, Rajeev H. (2005), "Practical Propensity Score Matching: A Reply to Smith and Todd," Journal of Econometrics 125 (March–April 2005), 355–364.

Goldberger, A. S. (1981), "Linear Regression After Selection," Journal of Econometrics 15 (1981), 357–366.

Gu, Xing Sam and Paul. R. Rosenbaum (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," Journal of Computational and Graphical Statistics 2 (1993).

Flörich, Markus (2004), "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," Review of Economic and Statistics 86 (February 2004), 77–90.

Hahn, Jinyong (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," Econometrica 66 (March 1998), 315–331.

Heckman, James J. (1979), "Sample Selection Bias as a Specification Error," Econometrica 47 (January 1979), 153–162.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd (1998), "Characterizing Selection Bias Using Experimental Dada," Econometrica 66 (September 1998), 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," Review of Economic Studies 64 (October 1997), 605–654.

_____(1998), "Matching as an Econometric Evaluation Estimator," Review of

Economic Studies 65 (April 1998), 261–294.

Heckman, James J. and Salvador Navarro-Lozano (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," Review of Economic and Statistics 86 (February 2004), 30–57.

Heckman, James J. and Edward Vytlacil (2005), "Structural Equations, Treatment Effects and Econometric Policy Evaluation," Econometrica 73 (May 2005), 669–738.

Heinrich, Carolyn J., Mueser, Peter R., Kenneth R. Troske (2005), "Welfare to Temporary Work: Implications for Labor Market Outcomes," Review of Economics and Statistics 87 (February, 2005), 154–173.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Econometrica 71 (July 2003), 1161–1189.

Horowitz, Joel L. (1993), "Semiparametric and Nonparametric Estimation of Quantal Response Models," Handbook of Statistics 11 (1993), ed. Maddala, G. S., Rao, C. R., and Vinod, H. D..

Imbens, Guido W. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions," Biometrika 87 (September 2000), 706–710.

Imbens, Guido W (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," American Economic Review Paper and Proceedings 93 (2003), 126–132.

Imbens, Guido W. (2004), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," Review of Economic and Statistics 86 (February 2004), 4–29.

Jalan, Jyotsna, and Martin Ravallion (2003), "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching," Journal of Business and Statistics 21 (January 2003), 19–30.

LaLonde, Robert J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," The American Economic Review 76 (September 1986), 604–620.

Lechner, Michael (2002), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," Review of Economics and Statistics 84 (May 2002), 205–220.

Levine, David I. and Gary Painter (2003), "The Schooling Costs Of Teenage Out-Of-Wedlock Childbearing: Analysis with a Within-School Propensity-Score-Matching Estimator," Review of Economics and Statistics 85 (November 2003), 884–900.

Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky (2005), "Using State Administrative Data to Measure Program Performance," unpublished manuscript.

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika 70 (April 1983), 41–55.

Rubin, Donald B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology 66 (1974), 688–701.

_____(1980), "Bias Reduction Using Mahalanobis-Metric Matching," Biometrics 36

(June 1980), 293–298.

Ruud, Paul A. (1983), "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models," Econometrica 51 (1983), 225–228.

Ruud, Paul A. (1986), "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," Journal of Econometrics 32 (1986), 157–187.

Smith, Jeffrey A. and Petra E. Todd (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" Journal of Econometrics 125 (March–April 2005), 305–353.

Westat (1981) "Continuous Longitudinal Manpower Survey Net Impact Report No. 1: Impact on 1977 Earnings of New FY 1976 CETA Enrollees in Selected Program Activities," Report prepared for US Department of Labor under Contract No. 23-23-74 (1981).

Zhao, Zhong (2004), "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," Review of Economic and Statistics 86 (February 2004), 91–107.

Figure 1: Balancing Score

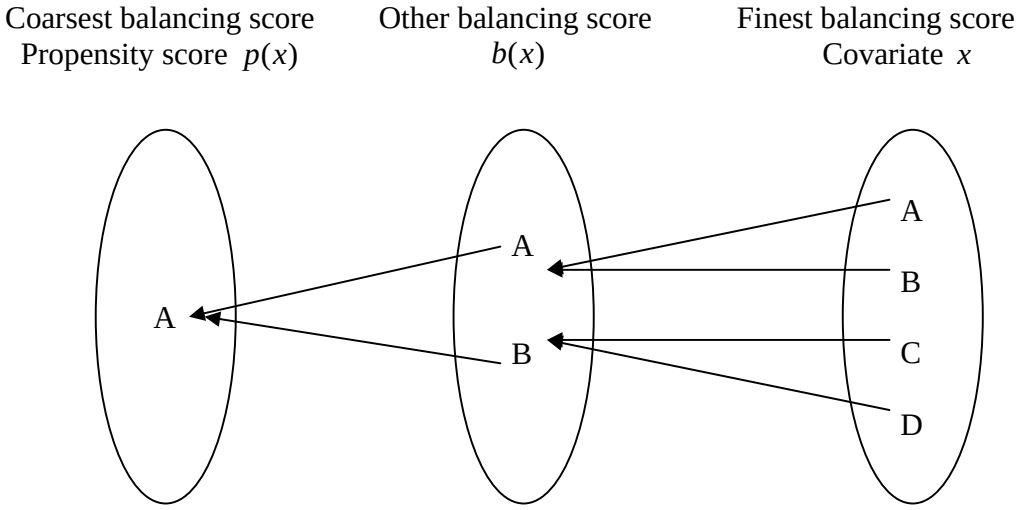
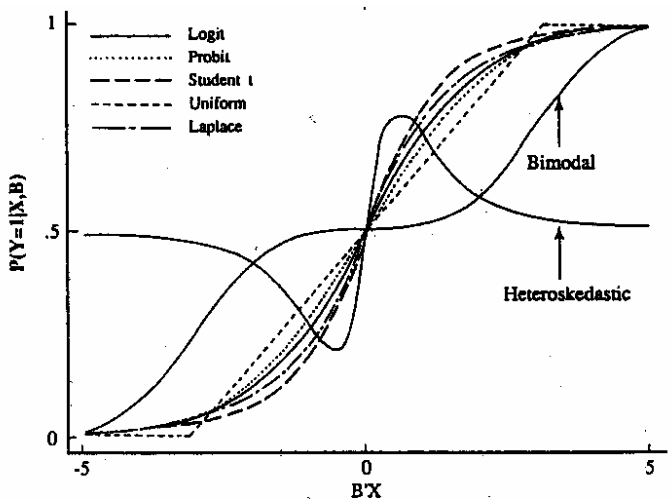


Figure 2: Binary Choice Model
with Various Distributions of Error Term



Source: Figure 1, Horowitz (1993)

Figure 3: Specification of Index Function

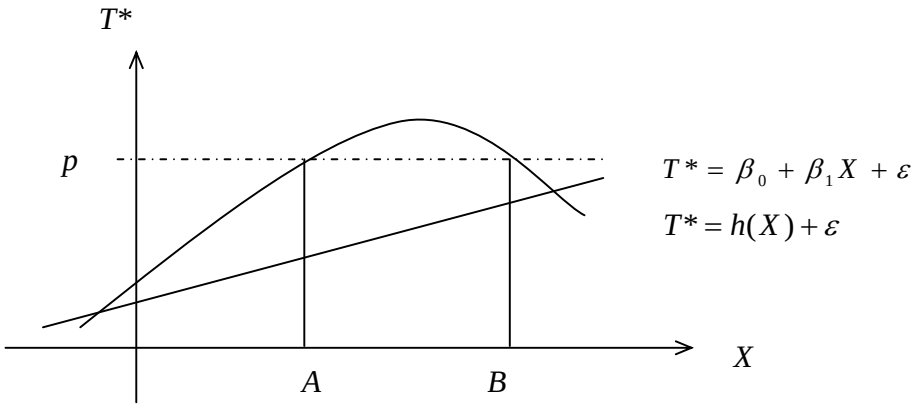


Table 1: Specification of the Error Term in Propensity Score Matching Model
(With non-normally distributed X)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.512	0.095	0.009	-0.486	0.092	0.009	-1.099	0.174	0.389
Logistic				-0.545	0.134	0.020	-0.519	0.129	0.017	-0.833	0.182	0.144
Bimodal				-0.513	0.084	0.007	-0.486	0.081	0.007	-1.421	0.207	0.891
Heter. Logistic				-1.566	3.936	16.630	-1.486	3.290	11.797	-1.563	3.336	12.262
Heter. Logistic				1.413	0.805	4.307	1.462	0.857	4.586	1.380	0.825	4.217
Heter. Normal	N/A	N/A	N/A	-6.759	14.264	242.648	-5.558	16.396	294.424	-6.183	13.821	223.317
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.803	0.111	0.657	1.002	0.107	0.012	1.005	0.106	0.011	0.949	0.107	0.014
Logistic	1.578	0.098	0.344	0.988	0.092	0.009	0.999	0.097	0.009	0.967	0.100	0.011
Bimodal	1.975	0.120	0.965	1.005	0.113	0.013	1.014	0.114	0.013	0.953	0.104	0.013
Heter. Logistic	1.209	0.093	0.052	0.988	0.080	0.006	0.988	0.080	0.006	0.991	0.078	0.006
Heter. Logistic	1.033	0.091	0.009	0.997	0.101	0.010	0.982	0.096	0.010	0.992	0.099	0.010
Heter. Normal	1.409	0.095	0.176	1.007	0.086	0.007	1.006	0.083	0.007	1.013	0.086	0.008
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.803	0.111	0.657	1.038	0.092	0.010	1.040	0.089	0.009	0.977	0.098	0.010
Logistic	1.578	0.098	0.344	1.013	0.081	0.007	1.023	0.083	0.007	0.982	0.084	0.007
Bimodal	1.975	0.120	0.965	1.050	0.101	0.013	1.062	0.101	0.014	0.994	0.092	0.009
Heter. Logistic	1.209	0.093	0.052	1.004	0.067	0.005	1.005	0.069	0.005	1.002	0.068	0.005
Heter. Logistic	1.033	0.091	0.009	0.986	0.081	0.007	0.977	0.073	0.006	0.981	0.078	0.007
Heter. Normal	1.409	0.095	0.176	1.020	0.070	0.005	1.019	0.070	0.005	1.020	0.070	0.005

Note: 1. Sample size: 1,000. Replication: 200.

2. One X has a Chi-squared distibution with 1 dgree of freedom and the other X is a mixture of two standard normal distributions.

Table 2: Specification of the Error Term in Propensity Score Matching Model
(With non-normally distributed X)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.508	0.071	0.005	-0.479	0.070	0.005	-1.093	0.138	0.370
Logistic				-0.529	0.102	0.011	-0.501	0.098	0.010	-0.815	0.148	0.121
Bimodal				-0.506	0.060	0.004	-0.478	0.059	0.004	-1.412	0.147	0.853
Heter. Logistic				-1.157	0.677	0.890	-1.132	0.669	0.847	-1.201	0.685	0.960
Heter. Logistic				1.262	0.560	3.417	1.301	0.592	3.596	1.222	0.566	3.286
Heter. Normal	N/A	N/A	N/A	-4.948	6.280	59.220	-4.571	6.620	60.405	-4.797	7.156	69.668
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.801	0.078	0.647	1.005	0.071	0.005	1.015	0.076	0.006	0.948	0.068	0.007
Logistic	1.574	0.064	0.333	0.993	0.066	0.004	1.003	0.057	0.003	0.956	0.061	0.006
Bimodal	1.968	0.084	0.943	1.000	0.083	0.007	1.019	0.084	0.007	0.959	0.080	0.008
Heter. Logistic	1.200	0.065	0.044	0.987	0.061	0.004	0.986	0.056	0.003	0.987	0.054	0.003
Heter. Logistic	1.022	0.068	0.005	0.989	0.069	0.005	0.981	0.070	0.005	0.986	0.071	0.005
Heter. Normal	1.403	0.066	0.167	1.003	0.053	0.003	0.998	0.057	0.003	1.003	0.052	0.003
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.801	0.078	0.647	1.034	0.063	0.005	1.041	0.060	0.005	0.974	0.060	0.004
Logistic	1.574	0.064	0.333	1.010	0.052	0.003	1.017	0.049	0.003	0.976	0.050	0.003
Bimodal	1.968	0.084	0.943	1.051	0.073	0.008	1.066	0.069	0.009	0.994	0.072	0.005
Heter. Logistic	1.200	0.065	0.044	0.997	0.048	0.002	0.998	0.047	0.002	0.995	0.047	0.002
Heter. Logistic	1.022	0.068	0.005	0.980	0.057	0.004	0.976	0.058	0.004	0.983	0.060	0.004
Heter. Normal	1.403	0.066	0.167	1.013	0.045	0.002	1.009	0.045	0.002	1.010	0.045	0.002

Note: 1. Sample size: 2,000. Replication: 200.

2. One X has a Chi-squared distibution with 1 dgree of freedom and the other X is a mixture of two standard normal distributions.

Table 3: Specification of the Error Term in Propensity Score Matching Model
(With non-normally distributed X; including irrelevant variables)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.497	0.099	0.010	-0.469	0.095	0.010	-1.064	0.190	0.354
Logistic				-0.522	0.144	0.021	-0.494	0.136	0.019	-0.791	0.201	0.125
Bimodal				-0.501	0.078	0.006	-0.474	0.076	0.006	-1.380	0.192	0.811
Heter. Logistic				-3.373	37.477	1412.770	-2.236	24.031	580.491	-2.830	30.933	962.305
Heter. Logistic				1.353	1.075	4.589	1.401	1.163	4.969	1.329	1.154	4.676
Heter. Normal	N/A	N/A	N/A	-5.312	52.315	2760.037	9.606	188.126	35493.403	-15.598	204.111	41889.371
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.806	0.108	0.661	1.009	0.105	0.011	1.023	0.104	0.011	0.964	0.093	0.010
Logistic	1.575	0.101	0.341	1.012	0.091	0.009	1.013	0.093	0.009	0.975	0.087	0.008
Bimodal	1.977	0.121	0.969	1.007	0.113	0.013	1.017	0.106	0.011	0.969	0.119	0.015
Heter. Logistic	1.201	0.098	0.050	0.993	0.084	0.007	1.001	0.084	0.007	0.993	0.089	0.008
Heter. Logistic	1.022	0.091	0.009	0.994	0.103	0.011	0.987	0.112	0.013	0.997	0.095	0.009
Heter. Normal	1.406	0.099	0.175	1.006	0.079	0.006	1.013	0.079	0.006	1.014	0.082	0.007
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.806	0.108	0.661	1.045	0.089	0.010	1.054	0.090	0.011	0.987	0.079	0.006
Logistic	1.575	0.101	0.341	1.027	0.077	0.007	1.030	0.077	0.007	0.993	0.071	0.005
Bimodal	1.977	0.121	0.969	1.056	0.098	0.013	1.064	0.098	0.014	1.009	0.107	0.011
Heter. Logistic	1.201	0.098	0.050	1.006	0.068	0.005	1.007	0.068	0.005	1.009	0.070	0.005
Heter. Logistic	1.022	0.091	0.009	0.985	0.081	0.007	0.981	0.082	0.007	0.993	0.072	0.005
Heter. Normal	1.406	0.099	0.175	1.014	0.070	0.005	1.017	0.071	0.005	1.015	0.071	0.005

Note: 1. Sample size: 1,000. Replication: 200.

2. One X has a Chi-squared distibution with 1 dgree of freedom and the other X is a mixture of two standard normal distributions.

Table 4: Specification of the Error Term in Propensity Score Matching Model
(With non-normally distributed X; unconfoundeness condition fails)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.497	0.091	0.008	-0.469	0.088	0.009	-1.063	0.175	0.348
Logistic				-0.529	0.144	0.022	-0.501	0.138	0.019	-0.805	0.200	0.133
Bimodal				-0.499	0.084	0.007	-0.471	0.081	0.007	-1.364	0.203	0.787
Heter. Logistic				-1.483	3.477	13.059	-1.478	3.765	15.134	-1.542	3.765	15.260
Heter. Logistic				1.464	1.195	5.283	1.521	1.334	5.866	1.442	1.312	5.492
Heter. Normal	N/A	N/A	N/A	-3.325	13.599	192.907	-2.512	15.453	242.839	-2.504	17.192	299.577
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	-1.763	0.993	8.618	3.116	1.338	6.269	3.174	1.243	6.269	1.051	1.187	1.411
Logistic	-1.666	1.008	8.124	3.160	1.183	6.066	3.304	1.213	6.778	2.336	1.183	3.185
Bimodal	-0.115	0.365	1.376	1.562	0.375	0.456	1.632	0.369	0.535	0.389	0.392	0.527
Heter. Logistic	2.008	1.346	2.828	3.095	1.371	6.269	3.230	1.302	6.667	3.044	1.327	5.941
Heter. Logistic	-44.065	803.834	648179.858	-22.446	386.133	149648.702	-3.823	173.827	30239.201	-20.590	380.160	144987.425
Heter. Normal	4.708	2.232	18.734	3.367	1.821	8.918	3.646	2.121	11.500	3.294	1.815	8.557
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	-1.763	0.993	8.618	3.285	1.225	6.721	3.398	1.157	7.089	1.167	0.941	0.914
Logistic	-1.666	1.008	8.124	3.219	0.981	5.885	3.282	0.993	6.193	2.405	1.043	3.063
Bimodal	-0.115	0.365	1.376	1.672	0.321	0.555	1.734	0.310	0.635	0.439	0.332	0.425
Heter. Logistic	2.008	1.346	2.828	3.183	1.138	6.062	3.261	1.180	6.502	3.147	1.151	5.935
Heter. Logistic	-44.065	803.834	648179.858	-11.036	241.333	58386.528	-23.562	413.923	171935.187	-15.759	316.803	100645.005
Heter. Normal	4.708	2.232	18.734	3.415	1.528	8.167	3.624	1.647	9.596	3.339	1.475	7.648

Note: 1. Sample size: 1,000. Replication: 200.

2. One X has a Chi-squared distibution with 1 dgree of freedom and the other X is a mixture of two standard normal distributions.

Table 5: Specification of the Error Term in Propensity Score Matching Model
(With non-normally distributed X; unconfoundeness condition fails)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.516	0.093	0.009	-0.488	0.089	0.008	-1.085	0.180	0.375
Logistic				-0.548	0.143	0.023	-0.522	0.138	0.020	-0.829	0.199	0.147
Bimodal				-0.503	0.074	0.006	-0.476	0.073	0.006	-1.364	0.197	0.786
Heter. Logistic				-1.258	1.105	1.797	-1.234	1.085	1.716	-1.296	1.102	1.849
Heter. Logistic				1.439	2.874	12.018	1.549	3.807	18.690	1.569	5.093	30.222
Heter. Normal	N/A	N/A	N/A	-5.340	13.392	202.768	-4.686	13.348	195.701	-4.650	15.772	265.992
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.416	0.070	0.178	0.846	0.065	0.028	0.855	0.063	0.025	0.802	0.068	0.044
Logistic	1.262	0.069	0.073	0.852	0.063	0.026	0.845	0.059	0.027	0.820	0.059	0.036
Bimodal	1.622	0.083	0.393	0.882	0.081	0.020	0.889	0.079	0.019	0.849	0.075	0.028
Heter. Logistic	0.993	0.065	0.004	0.841	0.054	0.028	0.842	0.057	0.028	0.844	0.060	0.028
Heter. Logistic	0.874	0.070	0.021	0.853	0.069	0.026	0.842	0.070	0.030	0.849	0.071	0.028
Heter. Normal	1.151	0.064	0.027	0.856	0.053	0.024	0.855	0.054	0.024	0.855	0.054	0.024
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	1.416	0.070	0.178	0.876	0.056	0.018	0.881	0.055	0.017	0.831	0.060	0.032
Logistic	1.262	0.069	0.073	0.865	0.051	0.021	0.865	0.052	0.021	0.839	0.051	0.029
Bimodal	1.622	0.083	0.393	0.922	0.070	0.011	0.929	0.070	0.010	0.881	0.068	0.019
Heter. Logistic	0.993	0.065	0.004	0.849	0.046	0.025	0.851	0.045	0.024	0.852	0.047	0.024
Heter. Logistic	0.874	0.070	0.021	0.844	0.053	0.027	0.838	0.055	0.029	0.843	0.056	0.028
Heter. Normal	1.151	0.064	0.027	0.866	0.046	0.020	0.865	0.046	0.020	0.866	0.047	0.020

Note: 1. Sample size:1,000. Replication: 200.
2. One X has a Chi-squared distibution with 1 dgree of freedom and the other X is a mixture of two standard normal distributions.

Table 6: Specification of the Error Term in Propensity Score Matching Model
(With non-normally distributed X; unconfoundeness condition fails)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.497	0.100	0.010	-0.469	0.097	0.010	-1.055	0.203	0.349
Logistic				-0.529	0.158	0.026	-0.502	0.149	0.022	-0.798	0.221	0.138
Bimodal				-0.497	0.085	0.007	-0.470	0.083	0.008	-1.362	0.219	0.790
Heter. Logistic				-17.181	223.509	50234.334	-2.385	14.952	227.102	-2.435	14.814	223.206
Heter. Logistic				1.327	1.534	5.691	1.392	1.867	7.067	1.322	1.905	6.951
Heter. Normal	N/A	N/A	N/A	-6.597	23.076	569.657	-4.835	14.245	221.722	-4.925	14.136	219.410
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	2.808	0.220	3.317	1.396	0.184	0.191	1.429	0.193	0.222	1.311	0.195	0.135
Logistic	2.349	0.177	1.850	1.370	0.171	0.166	1.370	0.172	0.167	1.311	0.159	0.122
Bimodal	2.626	0.198	2.682	1.245	0.181	0.093	1.254	0.175	0.095	1.151	0.174	0.053
Heter. Logistic	1.643	0.139	0.432	1.317	0.139	0.120	1.314	0.134	0.116	1.315	0.131	0.116
Heter. Logistic	1.344	0.162	0.144	1.297	0.168	0.116	1.297	0.164	0.115	1.304	0.174	0.123
Heter. Normal	1.955	0.148	0.934	1.340	0.129	0.133	1.334	0.126	0.128	1.333	0.139	0.130
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	2.808	0.220	3.317	1.466	0.167	0.246	1.494	0.173	0.274	1.356	0.171	0.156
Logistic	2.349	0.177	1.850	1.419	0.134	0.194	1.431	0.146	0.207	1.354	0.142	0.145
Bimodal	2.626	0.198	2.682	1.321	0.157	0.128	1.333	0.154	0.135	1.212	0.154	0.069
Heter. Logistic	1.643	0.139	0.432	1.340	0.112	0.128	1.341	0.112	0.129	1.341	0.114	0.130
Heter. Logistic	1.344	0.162	0.144	1.283	0.147	0.102	1.286	0.137	0.101	1.293	0.133	0.104
Heter. Normal	1.955	0.148	0.934	1.357	0.109	0.139	1.348	0.113	0.134	1.354	0.111	0.137

Note: 1. Sample size: 1,000. Replication: 200.

2. One X has a Chi-squared distibution with 1 dgree of freedom and the other X is a mixture of two standard normal distributions.

Appendix Table 1a: Specification of the Error Term in Propensity Score Matching Model
(With normally distributed X)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.505	0.098	0.010	-0.505	0.099	0.010	-0.508	0.100	0.010
Logistic				-0.509	0.148	0.022	-0.510	0.148	0.022	-0.511	0.148	0.022
Bimodal				-0.508	0.076	0.006	-0.508	0.078	0.006	-0.512	0.084	0.007
Heter. Logistic				-1.599	13.629	186.965	-2.688	29.050	848.720	-1.904	17.957	324.434
Heter. Logistic				10.597	161.288	26136.918	-0.253	16.643	277.037	-0.143	16.338	267.062
Heter. Normal	N/A	N/A	N/A	-0.517	0.150	0.023	-0.517	0.150	0.023	-0.518	0.150	0.023
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	3.108	0.425	4.624	1.035	0.332	0.111	1.035	0.342	0.118	1.048	0.330	0.111
Logistic	2.224	0.271	1.571	1.031	0.227	0.053	1.035	0.236	0.057	1.025	0.238	0.057
Bimodal	3.896	0.644	8.804	1.092	0.459	0.219	1.116	0.446	0.212	1.061	0.447	0.204
Heter. Logistic	1.410	0.203	0.209	1.022	0.214	0.046	1.026	0.204	0.042	1.013	0.212	0.045
Heter. Logistic	1.053	0.205	0.045	1.005	0.185	0.034	1.015	0.202	0.041	1.025	0.197	0.040
Heter. Normal	2.374	0.318	1.988	1.054	0.267	0.074	1.031	0.278	0.079	1.048	0.277	0.079
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	3.108	0.425	4.624	1.228	0.284	0.133	1.226	0.277	0.128	1.221	0.265	0.119
Logistic	2.224	0.271	1.571	1.116	0.191	0.050	1.114	0.202	0.054	1.104	0.207	0.054
Bimodal	3.896	0.644	8.804	1.375	0.355	0.267	1.385	0.351	0.272	1.357	0.368	0.263
Heter. Logistic	1.410	0.203	0.209	1.062	0.162	0.030	1.056	0.159	0.028	1.057	0.164	0.030
Heter. Logistic	1.053	0.205	0.045	1.052	0.178	0.034	1.055	0.167	0.031	1.064	0.165	0.031
Heter. Normal	2.374	0.318	1.988	1.113	0.217	0.060	1.112	0.226	0.063	1.109	0.232	0.066

Note: 1. Sample size: 1,000. Replication: 200.
2. Both X's have a standard normal distribution.

Appendix Table 2a: Specification of the Error Term in Propensity Score Matching Model
(With normally distributed X)

Panel A: Coefficients (Beta1/Beta2)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value				-0.500	N/A	N/A	-0.500	N/A	N/A	-0.500	N/A	N/A
Normal				-0.497	0.059	0.003	-0.496	0.059	0.004	-0.496	0.060	0.004
Logistic				-0.498	0.090	0.008	-0.498	0.090	0.008	-0.498	0.090	0.008
Bimodal				-0.497	0.047	0.002	-0.496	0.048	0.002	-0.497	0.052	0.003
Heter. Logistic				-0.505	0.268	0.072	-0.505	0.267	0.071	-0.505	0.267	0.071
Heter. Logistic				-0.443	15.783	249.091	0.602	22.425	504.089	0.704	23.935	574.326
Heter. Normal	N/A	N/A	N/A	-0.502	0.094	0.009	-0.501	0.093	0.009	-0.501	0.094	0.009
Panel B: Treatment Effects (Estimated Effect/True Effect) (Matching with Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	3.044	0.278	4.255	1.013	0.224	0.050	1.021	0.229	0.053	1.035	0.233	0.056
Logistic	2.210	0.177	1.495	1.011	0.175	0.031	1.010	0.168	0.028	1.017	0.169	0.029
Bimodal	3.829	0.443	8.199	1.067	0.334	0.116	1.062	0.336	0.117	1.054	0.329	0.111
Heter. Logistic	1.401	0.144	0.181	1.012	0.143	0.021	1.004	0.138	0.019	1.003	0.143	0.020
Heter. Logistic	1.047	0.147	0.024	1.032	0.155	0.025	1.022	0.152	0.024	1.019	0.147	0.022
Heter. Normal	2.353	0.192	1.867	1.005	0.188	0.035	0.996	0.208	0.043	1.018	0.207	0.043
Panel C: Treatment Effects (Estimated Effect/True Effect) (Matching without Replacement)												
	Simple Difference			Probit			Logit			LPM		
	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse	Mean	Std. Error	Mse
True Value	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A	1.000	N/A	N/A
Normal	3.044	0.278	4.255	1.177	0.194	0.069	1.175	0.194	0.068	1.191	0.190	0.072
Logistic	2.210	0.177	1.495	1.085	0.140	0.027	1.082	0.138	0.026	1.078	0.140	0.026
Bimodal	3.829	0.443	8.199	1.328	0.262	0.176	1.333	0.270	0.184	1.323	0.262	0.173
Heter. Logistic	1.401	0.144	0.181	1.039	0.114	0.015	1.044	0.116	0.015	1.045	0.116	0.015
Heter. Logistic	1.047	0.147	0.024	1.063	0.123	0.019	1.049	0.127	0.019	1.051	0.123	0.018
Heter. Normal	2.353	0.192	1.867	1.073	0.163	0.032	1.068	0.164	0.031	1.079	0.169	0.035

Note: 1. Sample size: 2,000. Replication: 200.
2. Both X's have a standard normal distribution.