USING MATCHING TO ESTIMATE TREATMENT EFFECTS: DATA REQUIREMENTS, MATCHING METRICS, AND MONTE CARLO EVIDENCE

Zhong Zhao*

Abstract—We compare propensity-score matching methods with covariate matching estimators. We first discuss the data requirements of propensity-score matching estimators and covariate matching estimators. Then we propose two new matching metrics incorporating the treatment outcome information and participation indicator information, and discuss the motivations of different metrics. Next we study the small-sample properties of propensity-score matching versus covariate matching estimators, and of different matching metrics, through Monte Carlo experiments. Through a series of simulations, we provide some guidance to practitioners on how to choose among different matching estimators and matching metrics.

I. Introduction

IDENTIFICATION and estimation of economic causal relationships has been a central theme in the history of econometrics (Heckman, 2000). This is especially true in the treatment effect literature (and the econometric program evaluation literature). However, using observational data to estimate treatment effects is plagued by the well-known selection bias problem, as mentioned by Heckman (1976, 1979, 1990), Barnow, Cain, and Golderger (1980), and Bjorklund and Moffitt (1987), among others. When the selection is due to unobservables, controlling for the resulting bias is a difficult and controversial issue. The discussion surrounding the instrumental variable approach by Angrist, Imbens, and Rubin (1996a, b), Angrist and Imbens (1999), Heckman (1996, 1997, 1999), and Moffitt (1996) is a good example. But when the selection is due to observables, the methods available for controlling the bias are relatively straightforward. A popular procedure is through matching, which is a method for selecting treated observations and comparison observations with similar covariates, by covariates X (for example, Rubin, 1980) or by propensity score p (for example, Rosenbaum and Rubin, 1983).[1,2]

Using covariate matching to correct the bias due to observables is intuitive, in that the source of the bias is the difference of observables in the treated group and comparison group. Matching on covariates, by definition, will remove this difference and hence the bias. When there are many covariates, however, it is impractical to match directly on covariates because of the curse of dimensionality. Taking the study of the Comprehensive 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. Because the number of observations is far less than 6 million, most of cells are empty and it was very difficult to find a good match on all 12 covariates. It is usually necessary to map the multiple covariates into a scalar through some metric, which measures the closeness of two observations. The most often used metric is the Mahalanobis metric (see, for example, Rubin, 1980).

Another way to overcome the curse of dimensionality is matching by propensity score. Rosenbaum and Rubin (1983) show that whereas covariate matching provides the finest balancing score, propensity-score methods provide the coarsest balancing score. A balancing score $b(x)$, as defined by Rosenbaum and Rubin (1983), is a function of the observed covariates such that the conditional distribution of x given $b(x)$ is the same for the treated and comparison groups. So conditioning on covariates and conditioning on the propensity score will both make the distribution of the covariates in the treated group the same as the distribution of the covariates in the comparison group.

The first issue this paper addresses concerns the finite-sample properties of covariate matching versus propensity-score matching. Most previous work has instead concerned large-sample properties. Hahn (1998) investigates the role of the propensity score in the estimation of the average treatment effect and shows that when using nonparametric series regression, adjusting for all covariates can achieve the efficiency bound, and adjusting for the propensity score alone cannot.[3] However, despite this asymptotic result, there is a sense among practitioners that propensity-score matching is better in that it avoids the problem of many small (or empty) cells with covariate matching. The analysis in this paper will investigate this intuition and show that it is not necessarily true in small samples.

Received for publication April 18, 2002. Revision accepted for publication August 28, 2003.

* China Center for Economic Research, Peking University.

This paper is a revision from chapter 2 of my PhD dissertation at Johns Hopkins. I am grateful to Robert Moffitt and Geert Ridder for their valuable advice and discussions, to Carl Christ for his detailed comments on the draft, and to Sid Chib, Bruce Hamilton, Jeff Smith, and other seminar participants at Johns Hopkins, Washington University in St. Louis, Manpower Demonstration Research Corporation, and Abt, Inc. for their helpful comments. Comments from an anonymous referee greatly improved the paper. All errors are mine.

[1] A covariate is defined as any variable whose value is not affected by the treatment; for example, sex is a covariate, but the treatment indicator is not. The propensity score is the probability of being treated conditional on the covariate X, that is, prob$(T = 1|X = x)$.

[2] Recent papers in this field by economists include Heckman, Ichimura, and Todd (1997, 1998), Heckman, Ichimura, Smith, and Todd (1998), Angrist (1998), Dehejia and Wahba (1999, 2002), Hahn (1998), Lechner (2002), Imbens (2000, 2003), Smith and Todd (2001, forthcoming), and Abadie and Imbens (2002).

[3] Other studies are as follows: Heckman, Ichimura, and Todd (1998) show that controlling for only the true propensity score can be less efficient than controlling for all covariates. Angrist and Hahn (1999) propose using a weighted combination of the covariate matching estimator and the propensity-score estimator to gain efficiency. Rubin and Thomas (1992) show that matching on the estimated propensity score is more efficient than on the true propensity score. Hirano, Imbens, and Ridder (2003) show that weighting with the inverse of a nonparametric estimate of the propensity score can achieve the efficiency bound.

The Review of Economics and Statistics, February 2004, 86(1): 91–107

The second issue addressed in this paper is the choice of matching metrics. In addition to the propensity-score matching metric and Mahalanobis metric, we will propose two new matching metrics incorporating the propensity information and the treatment outcome information.

The final task is to investigate the small-sample behavior of different matching estimators and of different matching metrics through a series of Monte Carlo experiments. In our simulations, we vary the number of covariates, the correlations between covariates and outcome, and the correlations between covariates and treatment indicator.

This paper is organized as follows. Section II sets up the model using the potential outcome framework, section III discusses different matching estimators and the related data issues, section IV proposes two new matching metrics and discusses the motivation behind them, section V provides Monte Carlo results for small-sample properties of various matching estimators and matching metrics, and section VI concludes this paper.

II. Model Setup

A fruitful framework for estimating treatment effects is the potential outcome framework, which dates back to Neyman (1923) and was popularized by Rubin (1974) and Holland (1986). This approach has also been used in economics, for example, by Roy (1951) and Quandt (1972). The setup in this section is standard in the treatment effect literature.[4]

In the potential-outcome framework, 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 $T_i = 1$ if individual i is treated, and $T_i = 0$ otherwise. With (Y_{0i}, Y_{1i}) we can define different treatment effects, such as those in Heckman and Vytlacil (1999), as follows:

$\Delta_i = Y_{1i} - Y_{0i}$	Treatment effect for individual i
$\Delta_{\text{ATE}} = E[\Delta_i]$	Average treatment effect (ATE) for the population
$\Delta_S = E[\Delta_i \mid i \in S]$	ATE for the subpopulation S

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

Unfortunately, for each i we can only observe $Y_i = Y_{1i}T_i + Y_{0i}(1 - T_i)$ and never both Y_{0i} and Y_{1i} simultaneously, so either Y_{0i} or Y_{1i} is missing for each i. It is clear that any attempt to estimate the treatment effect at the individual level will be hopeless without resorting to strong (and usually untestable) assumptions, such as the assumption of a homogeneous treatment effect across the population, regardless of whether the inference is drawn on experimental data or on observational data.

On the contrary, the ATE at the population (or subpopulation) level can be estimated without bias either by experimental data, or by observational data if the selection bias is only due to observables.

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: $0 < \text{prob}(T = 1 \mid X) < 1$	Common-support assumption

where $\perp\!\!\!\perp$ is the notation for statistical independence as in Dawid (1979). These assumptions in this form were introduced into the literature by Rosenbaum and Rubin (1983). In their terminology, treatment assignment is *strongly ignorable* when assumptions M-1 and M-2 are satisfied. Sometimes assumption M-1 is also referred as the "conditional independence assumption" or "selection on observables" in the literature.

Under M-1 and M-2,

$$\begin{aligned}\Delta_{\text{TT}} &= E_{x|T=1}\{E[Y_1 \mid T = 1, X = x] - E[Y_0 \mid T = 1, X = x]\} \\ &= E_{x|T=1}\{E[Y_1 \mid T = 1, X = x] \\ &\qquad - E[Y_0 \mid T = 0, X = x]\}.\end{aligned} \tag{1}$$

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's.

Following the celebrated result of Rosenbaum and Rubin (1983), define $p(X) = \text{prob}(T = 1 \mid X)$ to be the propensity score or the conditional probability of selection. Then M-1 and M-2 imply

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

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

The proof that M-1 and M-2 imply P-1 and P-2 is achieved by the so-called balancing property:

$$\begin{aligned}\text{prob}(X_i \mid T_i = 1, p(X_i) = p) &= \text{prob}(X_i \mid T_i = 0, p(X_i) = p) \\ &= \text{prob}(X_i \mid p).\end{aligned}$$

This identity is due to Rosenbaum and Rubin (1983). From P-1 and P-2 we have

$$\begin{aligned}\Delta_{\text{TT}} &= E_{p|T=1}\{E[Y_1 \mid T = 1, p(X) = p] \\ &\qquad - E[Y_0 \mid T = 1, p(X) = p]\} \\ &= E_{p|T=1}\{E[Y_1 \mid T = 1, p(X) = p] \\ &\qquad - E[Y_0 \mid T = 0, p(X) = p]\}.\end{aligned} \tag{2}$$

Unbiased estimates of $E[Y_1 \mid T = 1, p(X) = p]$ and $E[Y_0 \mid T = 0, p(X) = p]$ can also be obtained if $p(X)$ is

[4] For a different view, see Dawid (2000). The treatment effects discussed here—the average treatment effect for the population and the treatment effect on the treated—do not suffer from the metaphysic model problem raised in his paper.

known. The advantage of equation (2) over equation (1) is that instead of controlling for a high-dimensional vector of X, equation (2) only needs to control for a scalar p.

III. Covariate Matching, Propensity-Score Matching, and the Data Issues

The common approaches to control for the bias due to observable variables in the matching literature include matching on covariates or on the propensity score, subclassification by covariates or by the propensity score, and weighting by the propensity score.[5] We will focus on one-to-one matching estimators. One-to-one matching involves selecting a single observation from the comparison sample to match each observation in the treated sample by some metric. The choice to focus on matching estimators instead of on weighting estimators (the subclassification estimator is a special case of weighting estimators, as shown in footnote 5) is mainly driven by practical reasons. Both matching and weighting estimators are nonparametric in nature, but matching estimators can be used parametrically in the sense that the matching estimate is not sensitive to the parametric specification of the propensity score (Zhao, 2002), which is not true for weighting estimators. Furthermore, one-to-one matching estimators are widely used in empirical studies, and it is important to understand their properties.

In order to examine more closely how covariate matching and propensity-score matching work, write the two potential outcome equations and the selection equation as

$$Y_{1i} = f_1(X_i) + \varepsilon_{1i}, \tag{3}$$

$$Y_{0i} = f_0(X_i) + \varepsilon_{0i}, \tag{4}$$

$$T_i = I(T_i^* > 0), \tag{5}$$

$$T^* = h(X_i) + \mu_i, \tag{6}$$

where ε_{1i}, ε_{0i}, and μ_i are i.i.d. with zero conditional means (conditioning on X_i), and $I(\cdot)$ is the indicator function.

The basic ideas of covariate matching are

$$X_i = X_j \quad \Rightarrow \quad f_t(X_i) = f_t(X_j), \quad t = 0, 1, \tag{7}$$

and

$$d(X_i, X_j) < \varepsilon \quad \Rightarrow \quad d'(f_t(X_i), f_t(X_j)) < \delta, \quad t = 0, 1, \tag{8}$$

where d and d' are some metrics in the mathematical sense.

Equation (7) justifies exact matching. Equation (8) means that f_t is continuous at X. To see why the continuity of f_t is needed, we note that the true average treatment effect at x is $E[Y_1 - Y_0|x] = f_1(x) - f_0(x) + E[\varepsilon_1 - \varepsilon_0|x] = f_1(x) - f_0(x)$. Suppose that exact matching is infeasible, and we match on some x' within a small neighborhood of x, such that $d(x, x') < \delta$ but $x \neq x'$; then $E[Y_1|x] - E[Y_0|x'] = f_1(x) - f_0(x')$. Unless f_0 is a continuous function of X, $E[Y_1|x] - E[Y_0|x']$ will never converge to the true average treatment effect at x. The continuity of f_t justifies neighborhood matching when exact matching is infeasible. Because in most empirical studies only neighborhood matching is feasible (in fact, as long as there is a continuous covariate presented, it will exclude the exact matching), besides assumptions M-1 and M-2 we also need to assume that f_t is a continuous function of X.

Through covariate matching, observation i in the treated sample is matched with observation j in the comparison sample if $X_i = X_j = x$. Define

$$\begin{aligned}
\hat{\Delta}_i^C &= Y_{1i} - Y_{0j} \\
&= f_1(X_i) + \varepsilon_{1i} - f_0(X_j) - \varepsilon_{0j} \\
&= \{ f_1(x) + \varepsilon_{1i} - [f_0(x) + \varepsilon_{0i}]\} + \{\varepsilon_{0i} - \varepsilon_{0j}\} \\
&= \Delta_i + \{\varepsilon_{0i} - \varepsilon_{0j}\},
\end{aligned}$$

where Δ_i is the true treatment effect for individual i. The term $\varepsilon_{0i} - \varepsilon_{0j}$ almost surely does not equal 0, and it does not make sense to estimate the treatment effect for each individual as discussed earlier, but $\hat{\Delta}_i^C$ serves as the building block for estimating the treatment effect for some subpopulation. The treatment effect on the treated can be estimated by the following estimator:

$$\begin{aligned}
\hat{\Delta}_{\text{TT}}^C &= \frac{1}{n^C} \sum_{i=1}^{n^C} \hat{\Delta}_i^C \\
&= \Delta_{\text{TT}} + \frac{1}{n^C} \sum_{i=1}^{n^C} \varepsilon_{0i} - \frac{1}{n^C} \sum_{j=1}^{n^C} \varepsilon_{0j},
\end{aligned}$$

where n^C is the number of covariate matched pairs. Clearly $\hat{\Delta}_{\text{TT}}^C$ is an unbiased and consistent estimator of Δ_{TT}.

The closeness of the covariates of each matching pair plays a crucial role in the covariate matching and itself is enough to guarantee the reliability of the estimator under assumptions M-1 and M-2 and the continuity of f_t.

The theory behind propensity-score matching is quite different from that behind covariate matching. The basic ideas of propensity-score matching are:

1. $\text{prob}(X_i|T_i = 1, p(X_i) = p) = \text{prob}(X_i|T_i = 0, p(X_i) = p) = \text{prob}(X_i|p)$,
2. $d(p_k, p_l) < \varepsilon \Rightarrow d'(\text{prob}(X_i|p_k), \text{prob}(X_j|p_l)) < \delta$.

[5] Subclassification by covariates can be treated as a special case of weighting by propensity score, such as the one proposed in Hirano, Imbens, and Ridder (2003). Taking Δ_{ATE} as an example, it can be shown that its subclassification estimator is $\hat{\Delta}_{\text{ATE}} = (1/n) \sum_{i=1}^{n} \{(Y_iT_i)/\hat{P}_i - [Y_i(1 - T_i)]/(1 - \hat{P}_i)\}$, where $\hat{P}_i = \{\sum_i I(x_{c-1} < X_i < x_c)T_i\}/\{\sum_i I(x_{c-1} < X_i < x_c)\}$ is a naive propensity-score estimator, x_{c-1} and x_c are the border points for subclass c, and $I(\cdot)$ is the indicator function.

These two ideas are parallel to the two ideas of covariate matching. Idea 1 says that when the matching is exact at the propensity score p, then the distribution of X will be the same for the treated sample and the comparison sample at p. Idea 2 says that if exact matching is impossible and instead matching is on some neighborhood of p, then the distribution of X is still approximately the same for the treated sample and the comparison sample within the neighborhood of p.

To see the difference between covariate matching methods and propensity score matching methods, we observe the following:

Observation. In general, for individuals k and l with $p(X_k) = p(X_l)$ but $X_k \neq X_l$, one has $f_t(X_k) \neq f_t(X_l)$, $t = 0, 1$.

It is very possible that individuals with the same propensity score will have quite different treatment outcomes, that is, that p approximately the same, does not imply that the treatment outcome $f(X)$ is approximately the same. Because of the balancing property, $\text{prob}(X_i|T_i = 1, p(X_i) = p) = \text{prob}(X_i|T_i = 0, p(X_i) = p)$, this will not be a problem if the number of observations at each value of the propensity score is large. This can be easily seen if we compare propensity-score matching methods with a randomized experiment. $\text{prob}(X, \nu|\text{treated}) = \text{prob}(X, \nu|\text{control})$ is the foundation of a randomized experiment, where X is observable and ν is unobservable. The balancing property plays a similar role in propensity-score matching, but propensity-score matching methods differ from randomization in two important ways. First, a randomized experiment balances the distributions of both observables and unobservables between treated and control samples, but propensity-score matching only balances the observables. This is why the unconfoundedness assumption M-1 is needed. Second, a randomized experiment balances the distributions for the whole sample, but propensity-score matching balances the distributions at each individual propensity-score value. In other words, under M-1 and M-2, the matched sample at each propensity-score value p is equivalent to a randomized sample. The estimate of propensity-score matching can be thought of as a weighted average of the estimates from many miniature randomized experiments (at different p's). The overall quality of the estimation depends on the quality of each of these miniature experiments. A substantial sample size is needed to obtain a meaningful estimate from a randomized experiment, and this is translated into a sufficiently large sample size at each p for a meaningful propensity-score matching estimate. The following example provides some idea for this argument.

EXAMPLE (PROPENSITY-SCORE PAIR MATCHING): Suppose there is a p cell such that $\text{prob}(T = 1|X = x_1) = \text{prob}(T = 1|X = x_2) = p$ but $f_0(x_1) \neq f_0(x_2)$. The comparison sample is twice as large as the treated sample, X is balanced in the comparison and the treated samples such that $\text{prob}(X = x_1|T = 1) = \text{prob}(X = x_1|T = 0) = 0.5$ and $\text{prob}(X = x_2|T = 1) = \text{prob}(X = x_2|T = 0) = 0.5$.[6] When there are two treated and four comparisons (that is, treated $= \{x_1, x_2\}$ and comparison $= \{x_1, x_1, x_2, x_2\}$), selecting two comparisons to match the two treated by propensity score without replacement is equivalent to randomly picking up two elements from the comparison sample $\{x_1, x_1, x_2, x_2\}$, because they have the same propensity score. There are three possible cases: (1) {treated $= \{x_1, x_2\}$, comparison $= \{x_1, x_2\}$}, (2) {treated $= \{x_1, x_2\}$, comparison $= \{x_1, x_1\}$}, and (3) {treated $= \{x_1, x_2\}$, comparison $= \{x_2, x_2\}$}, with probabilities of 2/3, 1/6, and 1/6, respectively. Define $B_p^s = (1/m) \sum_{i=1}^{m} f_0(x_i) - (1/m) \sum_{j=1}^{m} f_0(x_j)$, where m is the number of matched pairs. In case (1), B_p^s is 0. In cases (2) and (3), the absolute value of B_p^s is $|f_0(x_1) - f_0(x_2)|/2$. The expectation of B_p^s is 0, and its standard error is $0.17|f_0(x_1) - f_0(x_2)|$. These two terms have standard statistical interpretations. The zero expectation of B_p^s means that the estimation is unbiased. The standard error captures the accuracy of the estimation. When the cell size is small, the standard error of B_p^s is very large, and it decreases with increasing cell size. For example, when $m(p) = 8$, the standard error reduces to $0.095|f_0(x_1) - f_0(x_2)|$.

When using propensity-score matching methods, observation i in the treated sample is matched with observation j in the comparison sample if $p(X_i) = p(X_j) = p$. Define

$$
\begin{aligned}
\hat{\Delta}_i^P &= Y_{1i} - Y_{0j} \\
&= f_1(X_i) + \varepsilon_{1i} - \{f_0(X_j) + \varepsilon_{0j}\} \\
&= \{f_1(X_i) + \varepsilon_{1i} - [f_0(X_i) + \varepsilon_{0i}]\} + \{f_0(X_i) - f_0(X_j) \\
&\quad + \varepsilon_{0i} - \varepsilon_{0j}\} \\
&= \Delta_i + \{f_0(X_i) - f_0(X_j) + \varepsilon_{0i} - \varepsilon_{0j}\}.
\end{aligned}
$$

Using $\hat{\Delta}_i^P$ as the building block, we can estimate the treatment effect on the treated by the following estimator, which is widely used in the matching literature:

$$
\begin{aligned}
\hat{\Delta}_{\mathrm{TT}}^P &= \frac{1}{n^p} \sum_{i=1}^{n^p} \hat{\Delta}_i^P \\
&= \frac{1}{n^p} \sum_{i=1}^{n^p} \Delta_i + \frac{1}{n^p} \sum_{i=1}^{n^p} \varepsilon_{0i} - \frac{1}{n^p} \sum_{j=1}^{n^p} \varepsilon_{0j} + \frac{1}{n^p} \sum_{i=1}^{n^p} f_0(x_i)
\end{aligned}
$$

[6] In this example, we assume that the covariates are already balanced between the treated sample and the comparison sample, so subclassification by propensity score seems superior to matching by propensity score. But in practice, when the p-cell size is small, it is very unlikely the covariates will be balanced between the treated sample and the comparison sample, so subclassification by propensity score is also problematic.

$$-\frac{1}{n^p}\sum_{j=1}^{n^p} f_0(x_j)$$

$$= \Delta_{\mathrm{TT}} + \frac{1}{n^p}\sum_{i=1}^{n^p} \varepsilon_{0i} - \frac{1}{n^p}\sum_{j=1}^{n^p} \varepsilon_{0j} + \frac{1}{n^p}\sum_{i=1}^{n^p} f_0(x_i)$$

$$-\frac{1}{n^p}\sum_{j=1}^{n^p} f_0(x_j),$$

where n^P is the number of propensity-score-matched pairs.

The perception that propensity-score estimators are less data-hungry is related to the fact that in general the number of p cells, r^p, is less than the number of X cells, r^C (for example, see Rosenbaum, 1995; also see the discussion in Angrist & Hahn, 1999) when the covariates are discrete.[7] But matching on the propensity score has the risk of matching individuals with quite different treatment outcomes together. To average this kind of mismatching out, propensity-score matching methods must rely on the balancing property. It seems that the combination of r^p, r^C, and the minimum number of observations at each propensity score ultimately determines the relative data requirements between propensity-score matching and covariate matching.

Nonetheless, the role of the propensity score in the estimation of the treatment effects is still not completely clear and is under active research. Hahn (1998) shows that the propensity score is ancillary for estimation of the average treatment effect, but not for estimation of the treatment effect on the treated. Hirano, Imbens, and Ridder (2003) show that weighting with the inverse of a nonparametric estimate of the propensity score leads to efficient estimates of various treatment effects, including the treatment effect on the treated. Heckman and Navarro-Lozano (2003) discuss the propensity score in selection, matching, and instrumental variables models.

IV. Matching Metrics for Matching Estimators

Many matching metrics have been proposed in the literature, including metrics based on covariates, such as the categorical distance, caliper distance, and quadratic distance of X; based on propensity scores, such as the absolute difference of p; and based on both X and p, such as a combination of a caliper on p and the Mahalanobis distance on X. See Cochran and Rubin (1973), Rosenbaum and Rubin (1985), and Rosenbaum (1995).

Though the metrics based on both X and p are more appealing than the metrics based only on X or only on p, the outcome variable Y contains useful information and should also be utilized.[8] We propose two metrics. The first is based on X and p, and the second is based on X and Y.

Let $p(X) = G(X\beta')$, and consider the following metric:

$$d_{Z1} = \sum_{k=1}^{K} |X_{ki} - X_{kj}| \cdot |\beta_k|.$$

This metric incorporates information on both X and p. In contrast with the widely used Mahalanobis metric, $d_M = (X_i - X_j)\,D^{-1}(X_i - X_j)'$,[9] both metrics are unit-free, which is essential when X is a vector. The Mahalanobis metric assigns weight to each coordinate of X in inverse proportion to the variance of that coordinate, but d_{Z1} weights each coordinate of X by its marginal effect on the propensity score.[10] The metric in Abadie and Imbens (2002), $d_{\mathrm{AI}} = (X_i - X_j)\ \mathrm{diag}(D^{-1})(X_i - X_j)'$, is similar to the Mahalanobis metric.

Another useful comparison is with the absolute difference of the propensity score, $|G(X_i\beta') - G(X_j\beta')|$, which is used in propensity-score matching. The potential risk of using this metric is discussed in Section III, and it is clear that using d_{Z1} can avoid the risk, namely,

$$|G(X_i\beta') - G(X_j\beta')| = 0 \quad \not\Rightarrow \quad f_0(X_i) = f_0(X_j),$$

but if $\beta' \neq 0$ then

$$\sum_{k=1}^{K} |X_{ki} - X_{kj}| \cdot |\beta_k| = 0 \quad \Rightarrow \quad f_0(X_i) = f_0(X_j).$$

In many cases, d_{Z1} is a finer metric than $|G(X_i\beta') - G(X_j\beta')|$ in the sense that $\sum_{k=1}^{K} |X_{ki} - X_{kj}| \cdot |\beta_k| \leq \varepsilon \Rightarrow |G(X_i\beta') - G(X_j\beta')| \leq \varepsilon$, because

$$\begin{aligned}\sum_{k=1}^{K} |X_{ki} - X_{kj}| \cdot |\beta_k| &\geq \sum_{k=1}^{K} |(X_{ki} - X_{kj}) \cdot \beta_k| \\ &= \sum_{k=1}^{K} |X_{ki} \cdot \beta_k - X_{kj} \cdot \beta_k| \\ &\geq |G(X_i\beta') - G(X_j\beta')|.\end{aligned}$$

The first inequality follows from the triangle inequality, and the last inequality will hold if the maximum of the density function is less than or equal to 0.5, which is satisfied by many distributions, including the standard normal distribution and the logistic distribution.

[7] But, as a referee points out, when the covariates are continuous, the measure of the set of the covariates with the same propensity score is usually 0.

[8] Dickinson, Johnson, and West (1986) used the information of both X and Y in their matching scheme.

[9] D is the variance-covariance matrix of X.

[10] Strictly speaking, β can be interpreted as the marginal effect only if it is estimated from the LPM. For other models, like probit and logit, β is not the marginal effect, but it is still proportional to the marginal effect.

A metric containing propensity-score information certainly has its own merit, but a more intuitive approach is to incorporate outcome information into the metric. After all, what one is doing in matching is using one observation's outcome to mimic the other observation's outcome. Assume (Y_{0i}, Y_{1i}) and X have linear relationships, such that

$$Y_{ti} = f_t(X_i) + \varepsilon_{ti} = X_i\alpha_t' + \varepsilon_{ti}, \qquad t = 0, 1.$$

Define the metric as

$$d_{Z2} = \sum_{k=1}^{K} |X_{ki} - X_{kj}| \cdot |\alpha_{kt}|.$$

This metric weights the coordinates of X by their marginal effects on the potential outcomes. It is a more natural measure of closeness of two observations in terms of their potential outcomes than are metrics based only on X or p or both.

In fact, d_{Z2} consists of two metrics: the one with $t = 0$, and the one with $t = 1$. When matching the comparison sample to the treated sample, we should use $t = 0$, and this is the case of estimating Δ_{TT}. When matching the treated sample to the comparison sample, we should use $t = 1$, and this is the case for the treatment effect on the untreated Δ_{UTT}. When estimating Δ_{ATE}, we need to use both.

Imbens (2003) discusses the metrics of d_{Z2}, and derives optimal metrics under linear specification of the outcome regression function and a normal distribution of the error term. He points out that when the outcome equation is misspecified, the matching result using d_{Z2} may be inconsistent.

The three metrics d_M, d_{AI}, and d_{Z2} share a common feature: all of them can be written as $(X_i - X_j)W(X_i - X_j)'$. The only difference is the weighting matrix W. Thus the problem is how to choose W to gain more efficiency, which is similar to the selection of optimal weighting matrix in GMM. Whereas in GMM the efficiency is related to the coefficients, in treatment-effect models the efficiency is related to the treatment effects. The standard Euclidean metric is weighted by an identity matrix, d_M is weighted by the inverse of the variance-covariance matrix of X, and d_{AI} is weighted by a diagonal matrix with the inverse of the variances of the X's as its elements. d_{Z2} is weighted by the coefficients from a linear regression. Of course the outcome equation can be nonlinear, but we can think of the coefficients from a linear specification as a simple way to capture the variance-covariance of X and the correlations between outcome and covariates. In this paper, we focus on the small-sample behavior of different matching metrics, and leave the large-sample theory of the selection of the weighting matrix for further research.

V. Monte Carlo Experiments[11]

There are relatively few studies on the small-sample properties of different matching estimators. Frölich (2000) compares one-to-one pair matching with local polynomial estimators through Monte Carlo experiments. He finds that the local linear estimator suggested by Seifert and Gasser (1996, 2000) is better than matching estimators. His simulations only include one covariate, but when the number of covariates increases, neither the bias of his suggested local linear estimator nor that of the matching estimators vanishes fast enough (Heckman, Ichimura, & Todd, 1998; Abadie & Imbens, 2002; Imbens, 2003). It is unclear whether the increase of the number of covariates will affect his results. Abadie and Imbens (2002) propose a bias-corrected matching estimator. Comparisons of the Seifert-Gasser estimator with the Abadie-Imbens estimator with a high dimension of covariates have not been done.

Angrist and Hahn (1999) argue that conventional asymptotic theory on matching estimators provides poor guidance for practitioners who face a finite sample, and they develop an alternative theory and propose a panel-style estimator which can provide finite-sample efficiency gains over covariate matching and propensity-score matching.

Our study is different from these. Our focus is on comparing the behavior of four matching estimators through Monte Carlo experiments with one-to-one matching. The four are matching by the propensity score, by the Mahalanobis metric, by the metric d_{Z1}, and by the metric d_{Z2}. They are referred to as propensity-score matching, covariate matching, covariate-and-propensity matching, and covariate-and-outcome matching, respectively. We investigate this issue in different settings. We vary the relative importance of covariates in the determination of the participation decision, the correlations between outcome and covariates and between treatment indicator and covariates, and the number of covariates. The Monte Carlo experiments are divided into two parts.

A. Monte Carlo Design I

This design investigates the effect of change of the degree of selection on observables on the performance of various estimators, and also the effect of change of the relative importance of covariates to the treatment participation decision on the behavior of each estimator. The setup is as follows:

$$Y_1 = \alpha_{10} + \alpha_{11}X_1 + \alpha_{12}X_2 + \varepsilon_1,$$

$$Y_0 = \alpha_{00} + \alpha_{01}X_1 + \alpha_{02}X_2 + \varepsilon_0,$$

$$T^* = \beta_0 + g\beta_1X_1 + gr\beta_2X_2 + \mu,$$

[11] The design of the Monte Carlo experiments has benefited greatly from an anonymous referee's suggestions and from discussions in Imbens (2003).

$$T = I(T^* > 0), \quad \text{where} \quad I(\cdot) \text{ is the indicator function.}$$

Both X_1 and X_2 and the error terms ε_1 and ε_0 are unit normally distributed. The errors ε_1 and ε_0 are dependent, but both are independent of μ. We fix α and β, but allow g and r to vary. The intercept term in the selection equation is chosen such that on average 20% of observations are in the treated sample. The sample sizes are 500, 1,000 and 2,000. On average, the treated sample sizes used in our simulations are 100, 200, and 400, and they correspond roughly to the sample sizes in many empirical studies. For examples, the number of treated observations in LaLonde (1986) is 297, and that in Dehejia and Wahba (1999) is 185.

First we fix $r = 1$ but change g. Decreasing g has two effects. First, as g decreases, there will be less selection on observables, and in the extreme case when $g = 0$, we have a randomized experiment. So as g decreases, all four estimators should improve. Second, as g decreases, the variance of T^* becomes smaller and the distribution of T^* becomes more concentrated. So on average the number of observations within a small neighborhood of a propensity-score value will increase. This will improve the performance of propensity-score matching but should not affect the other estimators. The performance of propensity-score matching should improve at a faster rate than that of other methods.

Second, we fix $g = 1$ and change r, that is, we change the ratio of $g\beta_1$ to $gr\beta_2$. Here r captures the relative importance of X_1 and X_2 to the treatment participation decision. For the same X_1 and X_2 but different treatments, r can be different. Taking single mothers and preprogram earnings as an example, it is likely that the former is more important for Aid to Families with Dependent Children but the latter is more important for a job training program. When r changes, the relative importance of X_1 and X_2 to the treatment participation decision changes. On the one hand, matching by the Mahalanobis metric and the metric d_{Z2} do not take this relative importance into account (see the discussion in section IV above), and the change of r should not affect the matching results. On the other hand, propensity-score matching and matching on the metric d_{Z1} place more weight on the more important X, and this means that it is possible that after matching, the more important X is balanced but the less important X is not. Here we say that an X is less important in the sense that it has less effect on the participation decision, but the same X may have a large effect on the treatment outcomes, which is the case in our Monte Carlo experiments, so as r decreases, we may see the standard error increase.

In the tables, the columns labeled “Mean” contain the ratio of the estimated treatment effect to the true treatment effect, so the true value is 1. The tables also present standard errors, biases, and mean squared errors from 200 replications.

Table 1 summarizes the results of matching with replacement when g changes but r is fixed. The propensity scores are estimated by probit, logit, and the linear probability model (LPM). For matching with replacement, overall matching by the propensity score has a small bias but a large standard error compared with other estimators. As g decreases, all estimators improve and propensity-score matching improves at a faster rate, as expected. Matching by d_{Z1} behaves the worst in most cases. The choice of estimators for the propensity score has little effect on the results.

When g is large ($g = 2$, the correlation between T^* and X_1 is 0.36, and the correlation between T^* and X_2 is -0.89), the bias from propensity-score matching is smaller than those from other estimators, but the standard error of propensity-score matching can be up to 3 times larger than the others. When the sample size is small, it is unclear which estimator to use. As the sample size increases, the bias of propensity-score matching changes just a bit, but the standard error falls a great deal. For other estimators, both bias and standard error fall significantly, especially matching by the Mahalanobis distance and by d_{Z2}. At sample size 2,000, matching by the Mahalanobis distance and matching by d_{Z2} have comparable bias to matching by propensity score, and they have much smaller standard errors. When g and the sample size are large, matching by the Mahalanobis distance and by d_{Z2} are more desirable than matching by the propensity score. When g is large but the sample size is small, there is a tradeoff between bias and the standard error. There is no clear winner.

When g is small ($g = 0.5$, the correlation between T^* and X_1 is 0.23, and the correlation between T^* and X_2 is -0.58), the standard error of propensity-score matching becomes significantly smaller, and it is comparable to the standard error of other estimators. It also has the smallest bias. This is a favorable situation for propensity-score matching, but the standard error of propensity-score matching methods is still larger than the standard error of covariate matching methods. This is not surprising. Even in a randomized experiment, adjusting X can still reduce sampling error and improve the accuracy of the estimation.

Estimates from matching without replacement (Appendix, table A1) usually show larger bias than the ones from matching with replacement. The biases and standard errors for the four estimators are nearly the same. Matching by d_{Z2} seems a bit better. There is no strong evidence to suggest which estimator should be chosen.

Table 2 summarizes the results of matching with replacement when r changes but g is fixed. An interesting observation is that when r falls, the bias falls but the standard error rises for all estimators. A possible explanation for the decrease of the bias is that with the decrease of r, the participation decision is determined more and more by only one X. The other X is becoming more evenly distributed in the treated and comparison samples, because it is becoming less correlated with treatment status. Controlling for one X should be easier than controlling for two X’s. Among covariate matching methods, matching by the Mahalanobis distance is the best choice. Propensity-score matching

TABLE 1.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITH REPLACEMENT: TWO COVARIATES, VARIABLE g, FIXED $r = 1$

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: $g = 2$												
Simple mean difference	5.2425	4.2425	1.3449	19.8076	4.9697	3.9697	0.9022	16.5725	5.0227	4.0227	0.5085	16.4407
OLS	1.0905	0.0905	0.4645	0.2240	1.0499	0.0499	0.3347	0.1145	1.0491	0.0491	0.2247	0.0529
Propensity-score matching:												
True	1.4971	0.4971	2.0548	4.4693	1.5935	0.5935	1.8064	3.6153	1.3986	0.3986	1.4457	2.2489
Probit	1.5174	0.5174	2.1673	4.9649	1.5263	0.5263	1.7733	3.4216	1.4049	0.4049	1.4050	2.1380
Logit	1.5236	0.5236	2.1757	5.0078	1.5231	0.5231	1.7833	3.4538	1.3936	0.3936	1.3889	2.0840
LPM	1.4863	0.4863	2.1333	4.7875	1.4726	0.4726	1.7500	3.2859	1.4514	0.4514	1.4255	2.2358
Covariate matching (Mahalanobis)	1.7541	0.7541	1.0518	1.6750	1.6469	0.6469	0.9197	1.2643	1.5422	0.5422	0.6844	0.7624
Covariate-and-Propensity:												
True	2.1466	1.1466	1.1147	2.5572	1.9663	0.9663	0.9737	1.8818	1.8289	0.8289	0.7836	1.3011
Probit	2.1655	1.1655	1.1156	2.6030	1.9775	0.9775	0.9684	1.8933	1.8229	0.8229	0.7878	1.2978
Logit	2.1682	1.1682	1.1211	2.6216	1.9799	0.9799	0.9674	1.8961	1.8207	0.8207	0.7867	1.2924
LPM	2.1664	1.1664	1.1086	2.5895	1.9654	0.9654	0.9629	1.8592	1.8198	0.8198	0.7916	1.2987
Covariate-and-outcome	1.8526	0.8526	0.7642	1.3109	1.6915	0.6915	0.6304	0.8756	1.5728	0.5728	0.4552	0.5353
Panel B: $g = 1$												
Simple mean difference	4.6132	3.6132	1.1012	14.2679	4.3950	3.3950	0.7495	12.0878	4.4214	3.4214	0.4573	11.9151
OLS	1.0914	0.0914	0.4237	0.1879	1.0682	0.0682	0.2807	0.0834	1.0455	0.0455	0.2113	0.0467
Propensity-score matching:												
True	1.1088	0.1088	0.9903	0.9925	1.1322	0.1322	0.7951	0.6497	1.0288	0.0288	0.5924	0.3518
Probit	1.2080	0.2080	0.9820	1.0076	1.1454	0.1454	0.7725	0.6179	1.0911	0.0911	0.5678	0.3307
Logit	1.1905	0.1905	1.0147	1.0659	1.1516	0.1516	0.7592	0.5994	1.0908	0.0908	0.5511	0.3120
LPM	1.2928	0.2928	0.9658	1.0185	1.1888	0.1888	0.7630	0.6178	1.0883	0.0883	0.5500	0.3103
Covariate matching (Mahalanobis)	1.3611	0.3611	0.6954	0.6140	1.2974	0.2974	0.5279	0.3671	1.1625	0.1625	0.3941	0.1817
Covariate-and-Propensity:												
True	1.5246	0.5246	0.7352	0.8157	1.4002	0.4002	0.5529	0.4659	1.2456	0.2456	0.4039	0.2235
Probit	1.5344	0.5344	0.7356	0.8267	1.3998	0.3998	0.5486	0.4608	1.2524	0.2524	0.4058	0.2284
Logit	1.5329	0.5329	0.7374	0.8277	1.3991	0.3991	0.5475	0.4590	1.2529	0.2529	0.4071	0.2297
LPM	1.5292	0.5292	0.7405	0.8284	1.3963	0.3963	0.5511	0.4608	1.2527	0.2527	0.4072	0.2297
Covariate-and-outcome	1.5087	0.5087	0.6748	0.7141	1.3965	0.3965	0.4519	0.3614	1.2061	0.2061	0.3214	0.1458
Panel C: $g = 0.5$												
Simple mean difference	3.4050	2.4050	0.7088	6.2864	3.2818	2.2818	0.4627	5.4207	3.3071	2.3071	0.3505	5.4456
OLS	1.0886	0.0886	0.3583	0.1362	1.0655	0.0655	0.2252	0.0550	1.0574	0.0574	0.1727	0.0331
Propensity-score matching:												
True	1.0654	0.0654	0.6295	0.4005	1.0515	0.0515	0.4252	0.1834	1.0025	0.0025	0.2829	0.0800
Probit	1.0947	0.0947	0.5397	0.3002	1.0944	0.0944	0.3960	0.1657	1.0100	0.0100	0.2665	0.0711
Logit	1.0738	0.0738	0.5817	0.3438	1.0847	0.0847	0.4040	0.1704	1.0340	0.0340	0.2648	0.0713
LPM	1.0514	0.0514	0.5970	0.3591	1.0898	0.0898	0.3969	0.1656	1.0062	0.0062	0.2757	0.0760
Covariate matching (Mahalanobis)	1.1543	0.1543	0.4229	0.2027	1.1030	0.1030	0.3534	0.1355	1.0501	0.0501	0.2325	0.0566
Covariate-and-Propensity:												
True	1.2249	0.2249	0.4249	0.2311	1.1265	0.1265	0.3382	0.1304	1.0801	0.0801	0.2178	0.0539
Probit	1.2190	0.2190	0.4334	0.2358	1.1168	0.1168	0.3365	0.1269	1.0851	0.0851	0.2193	0.0553
Logit	1.2216	0.2216	0.4323	0.2360	1.1180	0.1180	0.3345	0.1258	1.0844	0.0844	0.2192	0.0552
LPM	1.2219	0.2219	0.4379	0.2410	1.1212	0.1212	0.3340	0.1262	1.0852	0.0852	0.2202	0.0557
Covariate-and-outcome	1.2416	0.2416	0.4481	0.2592	1.1742	0.1742	0.3588	0.1591	1.0881	0.0881	0.2343	0.0627

Notes: 1. Replications: 200.
2. The numbers in the "Mean" columns are the ratios of the estimated treatment effects to the true treatment effects.
3. MSE is the abbreviation of mean squared error.

methods usually have a small bias but a large standard error compared to covariate matching methods. Measured by the mean squared error (MSE), the propensity-score matching estimators have the worst performance. In most cases matching without replacement (appendix, table A2) has a larger bias and a smaller standard error than matching with replacement. Unlike in the matching-with-replacement case, the propensity-score matching methods have smaller mean squared errors.

In our simulation studies, when matching without replacement, approximately 25% of the comparison sample is used to match with the treated sample (this 25% is consistent with our simulation design that on average 20% of the observations are in the treated sample), but when matching with replacement

TABLE 2.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITH REPLACEMENT: TWO COVARIATES, VARIABLE r, FIXED $g = 1$

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: $r = 0.5$												
Simple mean difference	4.4335	3.4335	1.5463	14.1800	4.2589	3.2589	0.9244	11.4749	4.2439	3.2439	0.4995	10.7724
OLS	1.1418	0.1418	0.5109	0.2811	1.1336	0.1336	0.3443	0.1364	1.1024	0.1024	0.2507	0.0733
Propensity-score matching:												
True	1.2140	0.2140	1.3922	1.9840	1.1829	0.1829	0.9545	0.9445	1.0662	0.0662	0.7224	0.5262
Probit	1.2099	0.2099	1.3295	1.8116	1.1944	0.1944	0.9075	0.8613	1.0702	0.0702	0.7040	0.5005
Logit	1.2038	0.2038	1.3429	1.8449	1.2197	0.2197	0.9176	0.8903	1.0547	0.0547	0.6909	0.4803
LPM	1.3027	0.3027	1.4147	2.0930	1.2277	0.2277	0.8959	0.8545	1.0709	0.0709	0.6870	0.4770
Covariate matching (Mahalanobis)	1.3798	0.3798	0.8985	0.9516	1.3007	0.3007	0.5747	0.4207	1.1547	0.1547	0.4602	0.2357
Covariate-and-Propensity:												
True	1.5521	0.5521	1.0235	1.3524	1.4277	0.4277	0.7165	0.6963	1.2554	0.2554	0.5278	0.3438
Probit	1.5674	0.5674	1.0685	1.4636	1.4284	0.4284	0.7367	0.7263	1.2522	0.2522	0.5352	0.3500
Logit	1.5528	0.5528	1.0637	1.4370	1.4309	0.4309	0.7409	0.7346	1.2504	0.2504	0.5379	0.3520
LPM	1.5663	0.5663	1.0477	1.4184	1.4283	0.4283	0.7442	0.7373	1.2506	0.2506	0.5368	0.3510
Covariate-and-outcome	1.6179	0.6179	0.8803	1.1567	1.4401	0.4401	0.4949	0.4386	1.2361	0.2361	0.3883	0.2065
Panel B: $r = 0.2$												
Simple mean difference	4.3049	3.3049	1.9452	14.7062	4.0019	3.0019	0.8792	9.7844	4.0314	3.0314	0.5519	9.4940
OLS	1.2150	0.2150	0.6415	0.4577	1.1578	0.1578	0.3519	0.1487	1.1267	0.1267	0.2708	0.0894
Propensity-score matching:												
True	1.1816	0.1816	1.7665	3.1535	1.1462	0.1462	1.0516	1.1272	1.0792	0.0792	0.8447	0.7198
Probit	1.3077	0.3077	1.8582	3.5476	1.1700	0.1700	1.0425	1.1157	1.0982	0.0982	0.8076	0.6619
Logit	1.2823	0.2823	1.8149	3.3736	1.1906	0.1906	1.0201	1.0769	1.0830	0.0830	0.8328	0.7004
LPM	1.2716	0.2716	1.6407	2.7657	1.1747	0.1747	1.0496	1.1322	1.0247	0.0247	0.8227	0.6774
Covariate matching (Mahalanobis)	1.4291	0.4291	1.1066	1.4087	1.2444	0.2444	0.6437	0.4741	1.1038	0.1038	0.5050	0.2658
Covariate-and-Propensity:												
True	1.5017	0.5017	1.4903	2.4727	1.3670	0.3670	0.8639	0.8810	1.2019	0.2019	0.7029	0.5348
Probit	1.4797	0.4797	1.4852	2.4359	1.3585	0.3585	0.8680	0.8819	1.1990	0.1990	0.7100	0.5437
Logit	1.4811	0.4811	1.4841	2.4340	1.3583	0.3583	0.8549	0.8592	1.2004	0.2004	0.7062	0.5389
LPM	1.4502	0.4502	1.5518	2.6108	1.3348	0.3348	0.8495	0.8337	1.1900	0.1900	0.7033	0.5307
Covariate-and-outcome	1.7202	0.7202	1.1006	1.7300	1.4521	0.4521	0.5564	0.5140	1.2329	0.2329	0.4334	0.2421
Panel C: $r = 0.1$												
Simple mean difference	4.2253	3.2253	2.0547	14.6244	3.9031	2.9031	0.8861	9.2132	3.9159	2.9159	0.5566	8.8123
OLS	1.2325	0.2325	0.6690	0.5016	1.1630	0.1630	0.3618	0.1575	1.1339	0.1339	0.2837	0.0984
Propensity-score matching:												
True	1.1855	0.1855	1.8103	3.3116	1.1538	0.1538	1.1552	1.3581	1.0339	0.0339	0.8875	0.7888
Probit	1.2063	0.2063	1.7527	3.1145	1.2790	0.2790	1.1116	1.3135	1.0351	0.0351	0.8808	0.7770
Logit	1.1423	0.1423	1.7342	3.0277	1.2488	0.2488	1.0720	1.2111	1.0326	0.0326	0.9152	0.8387
LPM	1.0560	0.0560	1.7424	3.0391	1.2461	0.2461	1.1247	1.3255	1.0172	0.0172	0.8425	0.7101
Covariate matching (Mahalanobis)	1.4117	0.4117	1.1462	1.4833	1.2273	0.2273	0.6531	0.4782	1.0756	0.0756	0.5027	0.2584
Covariate-and-Propensity:												
True	1.3944	0.3944	1.6257	2.7985	1.2990	0.2990	0.9500	0.9919	1.1442	0.1442	0.7701	0.6138
Probit	1.3471	0.3471	1.6172	2.7358	1.2944	0.2944	0.9446	0.9789	1.1428	0.1428	0.7795	0.6280
Logit	1.3530	0.3530	1.6133	2.7273	1.2874	0.2874	0.9499	0.9849	1.1381	0.1381	0.7787	0.6254
LPM	1.2928	0.2928	1.5959	2.6326	1.2818	0.2818	0.9374	0.9581	1.1414	0.1414	0.7674	0.6089
Covariate-and-outcome	1.7288	0.7288	1.1707	1.9017	1.4551	0.4551	0.5702	0.5322	1.2365	0.2365	0.4440	0.2531

Notes: 1. Replications: 200.
2. The numbers in the "Mean" columns are the ratios of the estimated treatment effects to the true treatment effects.
3. MSE is the abbreviation of mean squared error.

only around 10–13% of the comparison sample observations are used to match with the treated sample. Some observations in the comparison sample are used repetitively to match the treated sample. In an extreme case, the same comparison observation is used 137 times in a sample with 2,000 observations. This confirms the suspicion in the literature that matching with replacement runs the risk of excessively using a small part of comparison sample. Nonetheless, when the comparison sample is not much larger than the treated sample, matching with replacement seems more sensible.

Because we have set up the Monte Carlo experiment with linear outcome equations, conceivably a simple OLS for the outcome equation will be enough to control the confounded variables and also gain efficiency.

B. Monte Carlo Design II

This design is to investigate the effect of the number of covariates on the performance of different estimators, and to examine how changes of the correlations between covariates and outcome and of the correlations between covariates and treatment indicator affect the small-sample behavior of different estimators. In our simulations, we consider four covariates and eight covariates. In the four-covariate case, we also add two interaction terms and two quadratic terms to the outcome equation, and add one interaction term, two quadratic terms, and one cubic term to the participation equation. In the eight-covariate case, we also add three interaction terms and three quadratic terms to the outcome equation, and add three interaction terms, two quadratic terms, and one cubic term to the participation equation. We use the R^2 of the outcome equations (denoted as R_O^2 hereafter) and the R^2 of the participation equation (calculated from the LPM model, and denoted as R_P^2 hereafter) as the measures of correlations between covariates and outcome and of correlations between covariates and treatment indicator, respectively. We consider three scenarios: $(R_O^2, R_P^2) = (0.6, 0.2)$, $(0.4, 0.4)$, and (0.2, 0.6).

Table 3 summarizes the case for four covariates with replacement. For different R_O^2 and R_P^2, and for different observation numbers, the propensity-score matching has the smallest biases, but the biggest standard errors.[12] Measured by MSE, the propensity-score matching estimator is worse than other estimators, and the worst case for propensity-score matching is with high R_O^2 $(=0.6)$ and low R_P^2 $(=0.2)$. The performance of the propensity-score matching estimator is improved with a decrease of R_O^2 and increase of R_P^2. All other matching estimators perform basically the same. Mahalanobis-metric matching is relatively robust under different setups. For the matching estimator using the metric d_{Z2}, the nonlinear outcome equation misspecified as linear does not make much difference.

Table 4 shows the results for the case of eight covariates with replacement. Measured by MSE, Mahalanobis matching performs best under different settings. When the sample size is small ($n = 500$), the propensity score doesn't perform well. When the sample size is relatively large ($n = 1{,}000$ or $n = 2{,}000$), and when R_P^2 is low ($R_P^2 = 0.2$ or $R_P^2 = 0.4$), the propensity-score matching outperforms other matching estimators except Mahalanobis matching. Among all matching estimators, propensity-score matching has the smallest bias under every scenario. Unlike in table 3, the misspecified outcome equation has a sizable effect when using the d_{Z2} metric.

Appendix tables A.3 and A.4 summarize the results for matching without replacement. Similar to the results in Monte Carlo design I, matching without replacement has a larger bias and a smaller standard error than matching with replacement. For the case of four covariates, there is no clear winner among different matching estimators. For the case of eight covariates, propensity-score matching performs best in most settings, particularly those with large sample sizes ($n = 1{,}000$ or $n = 2{,}000$), or with high R_P^2 ($R_P^2 = 0.4$ or $R_P^2 = 0.6$).

Overall, three conclusions are worth noting. The first is that when the correlations between covariates and the participation indicator are high, propensity-score matching is a good choice. The second is that, despite the fact that propensity-score matching can overcome the small-cell (or empty-cell) problem when the sample size is small, it does not perform well compared with other matching estimators if the sample size is too small (in our case $n = 500$). One reason is that in small samples, the variance dominates the bias. The last is that Mahalanobis matching is relatively robust under different settings. This may be due to the equal percentage bias reduction of Mahalanobis metric (Rubin, 1976).

One problem in using matching to estimate treatment effects is how to estimate the standard errors. In empirical studies, researchers usually calculate standard errors through bootstrapping, as do Smith and Todd (forthcoming). Rubin and Thomas (1992, 1996) discuss this issue for ellipsoidal and normal distributions. Abadie and Imbens (2002) provide an estimator for the conditional variance. In our Monte Carlo experiments, we take advantage of the fact that we do know the true treatment effect. From each draw, we get an estimated treatment effect, and we can calculate the sample standard error directly from the true treatment effect and these estimated treatment effects. Of course this approach is impracticable for empirical studies, where we usually do not know the true treatment effect.

VI. Conclusions

Selection bias due only to observables is a strong assumption. But with a proper data set and if the selection-on-observables assumption is justifiable, matching methods are useful tools to estimate treatment effects. There is no clear winner among the different matching estimators considered here. Propensity-score matching methods rely on the balancing property. Monte Carlo experiments show that the different methods do not dominate each other in term of performance. The evidence from the Monte Carlo experiment suggests that incorporating outcome information into the matching metric might be a promising approach.

From our simulations, three results are worth noting: first, when the correlations between covariates and the participation indicator are high, propensity-score matching is a good choice; second, when the sample size is too small, propensity-score matching does not perform well compared with other matching estimators; and last, Mahalanobis matching is relatively robust under different settings.

[12] We use the true propensity score here. In a separate project, we will investigate the issue of estimated propensity score and study the sensitivity of the estimated treatment effects to the specifications of the propensity score.

TABLE 3.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITH REPLACEMENT: FOUR COVARIATES

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: R^2 of Outcome Equation = 0.6; R^2 of Participation Equation = 0.2												
Simple mean difference	1.489	0.4890	0.2689	0.3114	1.4368	0.4368	0.1724	0.2205	1.4731	0.4731	0.1331	0.2415
OLS with linear specification	1.0975	0.0975	0.1773	0.0409	1.0627	0.0627	0.117	0.0176	1.0853	0.0853	0.0861	0.0147
OLS with correct specification	1.0368	0.0368	0.1507	0.0241	1.0103	0.0103	0.0997	0.0100	1.0307	0.0307	0.0744	0.0065
OLS separately for control and treated	1.0156	0.0156	0.1523	0.0234	0.986	−0.0140	0.1011	0.0104	1.008	0.0080	0.0745	0.0056
Propensity-score matching (true)	1.0235	0.0235	0.3397	0.1159	0.9834	−0.0166	0.2379	0.0569	1.0172	0.0172	0.1659	0.0278
Covariate matching (Mahalanobis)	1.1305	0.1305	0.2125	0.0622	1.0696	0.0696	0.1435	0.0254	1.0677	0.0677	0.1073	0.0161
Covariate-and-Propensity (true)	1.1248	0.1248	0.2063	0.0581	1.0799	0.0799	0.1432	0.0269	1.0815	0.0815	0.1023	0.0171
Covariate-and-Outcome:												
True	1.1167	0.1167	0.2223	0.0630	1.0441	0.0441	0.1415	0.0220	1.0485	0.0485	0.1089	0.0142
Correctly specified	1.1224	0.1224	0.226	0.0661	1.0584	0.0584	0.1499	0.0259	1.0529	0.0529	0.1131	0.0156
Linear specification	1.1428	0.1428	0.241	0.0785	1.0683	0.0683	0.1406	0.0244	1.0696	0.0696	0.1029	0.0154
Panel B: R^2 of Outcome Equation = 0.4; R^2 of Participation Equation = 0.4												
Simple mean difference	1.7267	0.7267	0.4092	0.6955	1.6946	0.6946	0.2704	0.5556	1.6776	0.6776	0.1899	0.4952
OLS with linear specification	1.1247	0.1247	0.2948	0.1025	1.098	0.0980	0.2031	0.0509	1.1	0.1000	0.145	0.0310
OLS with correct specification	1.0581	0.0581	0.28	0.0818	1.0368	0.0368	0.1971	0.0402	1.0336	0.0336	0.133	0.0188
OLS separately for control and treated	1.0327	0.0327	0.2969	0.0892	1.0209	0.0209	0.2027	0.0415	1.0072	0.0072	0.1371	0.0188
Propensity-score matching (true)	1.0723	0.0723	0.552	0.3099	1.0281	0.0281	0.3499	0.1232	1.0331	0.0331	0.2795	0.0792
Covariate matching (Mahalanobis)	1.2279	0.2279	0.3902	0.2042	1.1382	0.1382	0.2842	0.0999	1.0995	0.0995	0.1812	0.0427
Covariate-and-Propensity (true)	1.2029	0.2029	0.3824	0.1874	1.1663	0.1663	0.2838	0.1082	1.1229	0.1229	0.1765	0.0463
Covariate-and-Outcome:												
True	1.2249	0.2249	0.3946	0.2063	1.145	0.1450	0.2902	0.1052	1.0942	0.0942	0.1787	0.0408
Correctly specified	1.2687	0.2687	0.4004	0.2325	1.1714	0.1714	0.3095	0.1252	1.1134	0.1134	0.2031	0.0541
Linear specification	1.2702	0.2702	0.4304	0.2583	1.1672	0.1672	0.2873	0.1105	1.1255	0.1255	0.182	0.0489
Panel C: R^2 of Outcome Equation = 0.2; R^2 of Participation Equation = 0.6												
Simple mean difference	2.024	1.0240	0.8073	1.7003	1.917	0.9170	0.4437	1.0378	1.8389	0.8389	0.2918	0.7889
OLS with linear specification	1.1817	0.1817	0.613	0.4088	1.1054	0.1054	0.3835	0.1582	1.0778	0.0778	0.2624	0.0749
OLS with correct specification	1.118	0.1180	0.6153	0.3925	1.0442	0.0442	0.3889	0.1532	1.0245	0.0245	0.2633	0.0699
OLS separately for control and treated	1.1003	0.1003	0.7047	0.5067	1.0325	0.0325	0.4232	0.1802	0.9978	−0.0022	0.2847	0.0811
Propensity-score matching (true)	1.3063	0.3063	1.1484	1.4126	1.0017	0.0017	0.7325	0.5366	0.9854	−0.0146	0.5364	0.2879
Covariate matching (Mahalanobis)	1.3938	0.3938	0.914	0.9905	1.1826	0.1826	0.534	0.3185	1.1579	0.1579	0.3613	0.1555
Covariate-and-Propensity (true)	1.3541	0.3541	0.9135	0.9599	1.2527	0.2527	0.5144	0.3285	1.2114	0.2114	0.366	0.1786
Covariate-and-Outcome:												
True	1.4383	0.4383	0.9587	1.1112	1.2232	0.2232	0.4781	0.2784	1.189	0.1890	0.3627	0.1673
Correctly specified	1.5843	0.5843	0.9769	1.2957	1.3583	0.3583	0.5713	0.4548	1.2186	0.2186	0.4076	0.2139
Linear specification	1.5309	0.5309	0.9043	1.0996	1.3216	0.3216	0.5792	0.4389	1.2314	0.2314	0.4083	0.2203

Notes: 1. Replications: 200.
2. The numbers in the "Mean" columns are the ratios of the estimated treatment effects to the true treatment effects.
3. MSE is the abbreviation of mean squared error.

TABLE 4.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITH REPLACEMENT: EIGHT COVARIATES

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: R^2 of Outcome Equation = 0.6; R^2 of Participation Equation = 0.2												
Simple mean difference	1.7091	0.7091	0.8577	1.2385	1.6227	0.6227	0.2633	0.4571	1.5945	0.5945	0.1988	0.3930
OLS with linear specification	0.7522	−0.2478	0.1936	0.0989	0.7547	−0.2453	0.1424	0.0804	0.7375	−0.2625	0.0928	0.0775
OLS with correct specification	0.6256	−0.3744	0.1679	0.1684	0.6287	−0.3713	0.1279	0.1542	0.6313	−0.3687	0.0829	0.1428
OLS separately for control and treated	1.0071	0.0071	0.1481	0.0220	0.9969	−0.0031	0.0929	0.0086	1.0006	0.0006	0.0716	0.0051
Propensity-score matching (true)	1.0963	0.0963	0.7118	0.5159	1.0045	0.0045	0.2832	0.0802	0.9862	−0.0138	0.2278	0.0521
Covariate matching (Mahalanobis)	1.1881	0.1881	0.4468	0.2350	1.1417	0.1417	0.1585	0.0452	1.1208	0.1208	0.1196	0.0289
Covariate-and-Propensity (true)	1.3267	0.3267	0.4385	0.2990	1.249	0.2490	0.2162	0.1087	1.2417	0.2417	0.1467	0.0799
Covariate-and-Outcome:												
True	1.339	0.3390	0.4332	0.3026	1.274	0.2740	0.2146	0.1211	1.2495	0.2495	0.1661	0.0898
Correctly specified	1.3334	0.3334	0.4341	0.2996	1.2721	0.2721	0.2123	0.1191	1.2466	0.2466	0.1596	0.0863
Linear specification	1.3666	0.3666	0.5205	0.4053	1.3315	0.3315	0.2367	0.1659	1.3154	0.3154	0.1782	0.1312
Panel B: R^2 of Outcome Equation = 0.4; R^2 of Participation Equation = 0.4												
Simple mean difference	1.6474	0.6474	0.4584	0.6293	1.6283	0.6283	0.2723	0.4689	1.5753	0.5753	0.1662	0.3586
OLS with linear specification	0.6985	−0.3015	0.2378	0.1475	0.7028	−0.2972	0.1665	0.1161	0.7116	−0.2884	0.1151	0.0964
OLS with correct specification	0.5322	−0.4678	0.2376	0.2753	0.5532	−0.4468	0.1598	0.2252	0.5538	−0.4462	0.1137	0.2120
OLS separately for control and treated	0.989	−0.0110	0.252	0.0636	1.0106	0.0106	0.1743	0.0305	0.9927	−0.0073	0.1054	0.0112
Propensity-score matching (true)	1.0056	0.0056	0.5429	0.2948	0.9929	−0.0071	0.3642	0.1327	0.9949	−0.0051	0.2387	0.0570
Covariate matching (Mahalanobis)	1.1097	0.1097	0.343	0.1297	1.1689	0.1689	0.2268	0.0800	1.1187	0.1187	0.1481	0.0360
Covariate-and-Propensity (true)	1.2848	0.2848	0.4192	0.2568	1.3028	0.3028	0.2677	0.1634	1.2402	0.2402	0.1799	0.0901
Covariate-and-Outcome:												
True	1.3007	0.3007	0.4269	0.2727	1.3087	0.3087	0.2651	0.1656	1.2398	0.2398	0.1677	0.0856
Correctly specified	1.3171	0.3171	0.4723	0.3236	1.3114	0.3114	0.2688	0.1692	1.2469	0.2469	0.1774	0.0924
Linear specification	1.3558	0.3558	0.4133	0.2974	1.3726	0.3726	0.2945	0.2256	1.3122	0.3122	0.1829	0.1309
Panel C: R^2 of Outcome Equation = 0.2; R^2 of Participation Equation = 0.6												
Simple mean difference	1.7055	0.7055	0.5765	0.8301	1.6455	0.6455	0.3279	0.5242	1.635	0.6350	0.21	0.4473
OLS with linear specification	0.7063	−0.2937	0.4048	0.2501	0.6718	−0.3282	0.2801	0.1862	0.7008	−0.2992	0.1646	0.1166
OLS with correct specification	0.4739	−0.5261	0.4294	0.4612	0.4531	−0.5469	0.2744	0.3744	0.4745	−0.5255	0.1633	0.3028
OLS separately for control and treated	1.0401	0.0401	0.4842	0.2361	1.0068	0.0068	0.3047	0.0929	1.0142	0.0142	0.1756	0.0310
Propensity-score matching (true)	1.1183	0.1183	0.8848	0.7969	1.0078	0.0078	0.6782	0.4600	1.0785	0.0785	0.4722	0.2291
Covariate matching (Mahalanobis)	1.2533	0.2533	0.5502	0.3669	1.1895	0.1895	0.3799	0.1802	1.1827	0.1827	0.2423	0.0921
Covariate-and-Propensity (true)	1.3943	0.3943	0.5976	0.5126	1.3506	0.3506	0.3765	0.2647	1.3238	0.3238	0.2449	0.1648
Covariate-and-Outcome:												
True	1.3954	0.3954	0.5677	0.4786	1.3622	0.3622	0.386	0.2802	1.3252	0.3252	0.2446	0.1656
Correctly specified	1.4232	0.4232	0.6272	0.5725	1.3618	0.3618	0.4013	0.2919	1.3239	0.3239	0.2395	0.1623
Linear specification	1.4198	0.4198	0.5998	0.5360	1.4446	0.4446	0.4155	0.3703	1.4025	0.4025	0.2404	0.2198

Notes: 1. Replications: 200.
2. The numbers in the "Mean" columns are the ratios of the estimated treatment effects to the true treatment effects.
3. MSE is the abbreviation of mean squared error.

REFERENCES

Abadie, Albert, and Guido W. Imbens, "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," Department of Economics, University of California at Berkeley, unpublished manuscript (2002).

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

Angrist, Joshua D., and Jinyong Hahn, "When to Control For Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER technical working paper no. 241 (1999).

Angrist, Joshua D., and Guido W. Imbens, "Comments on James J. Heckman, 'Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations'," *The Journal of Human Resources* 34 (Fall 1999), 821–827.

Angrist, Joshua D., Guido W. Imbens, and D. B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91 (June 1996a), 444–458.

——— "Identification of Causal Effects Using Instrumental Variables: Rejoinder," *Journal of the American Statistical Association* 91 (June 1996b), 468–472.

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

Bjorklund, Anders, and Robert Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," this REVIEW, 69 (February 1987), 42–49.

Cochran, W. G., and Donald B. Rubin, "Controlling Bias in Observational Studies: A Review," *Sankhyā, Series A* 35 (1973), 417–446.

Dawid, A. Philip, "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society. Series B (Methodological)* 41:1 (1979), 1–31.

——— "Causal Inference without Counterfactuals," *Journal of the American Statistical Association* 95 (June 2000), 407–424.

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

——— "Propensity Score Matching Methods for Non-experimental Causal Studies," this REVIEW, 84 (February 2002), 151–175.

Dickinson, Katherine P., Terry R. Johnson, and Richard W. West, "An Analysis of the Impact of CETA Programs on Participants Earnings," *Journal of Human Resources* 21 (Winter 1986), 64–91.

Frölich, Markus, "Treatment Evaluation: Matching versus Local Polynomial Regression," Department of Economics, University of St. Gallen, discussion paper 2000-17 (2000).

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

Heckman, James J., "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5 (Fall 1976), 475–492.

——— "Sample Selection Bias as a Specification Error," *Econometrica* 47 (January 1979), 153–162.

——— "Varieties of Selection Bias," *American Economic Review* 80 (May 1990), 313–318.

——— "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association* 91 (June 1996), 444–458.

——— "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32 (Summer 1997), 441–462.

——— "Instrumental Variables: Response to Angrist and Imbens," *The Journal of Human Resources* 34 (Fall 1999), 828–837.

——— "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *Quarterly Journal of Economics* 115 (February 2000), 45–97.

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

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

——— "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65 (April 1998), 261–294.

Heckman, James J., and Salvador Navarro-Lozano, "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models," NBER working paper no. 9497 (2003).

Heckman, James J., and Edward Vytlacil, "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the U.S.A.* 96 (February 1999), 4730–4734.

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

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association* 81 (December 1986), 945–970.

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

——— "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review," Department of Economics, University of California at Berkeley, unpublished manuscript (2003).

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

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

Moffitt, Robert A., "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association* 91 (June 1996), 462–463.

Neyman, Jerzy S., "On the Application of Probability Theory to Agriculture Experiments. Essay on Principles. Section 9," *Statistical Science* 5 (1990), 465–485 [translated from the Polish, *Roczniki Nauk Rolniczych* X (1923), 1–51].

Quandt, Richard E., "A New Approach to Estimating Switching Regressions," *Journal of American Statistical Association* 67 (June 1972), 306–310.

Rosenbaum, Paul R., *Observational Studies* (New York: Springer-Verlag, 1995).

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

——— "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *American Statistician* 39 (February 1985), 33–38.

Roy, Andrew D., "Some Thoughts on the Distribution of Earnings," *Oxford Economics Papers* 3 (1951), 135–146.

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

——— "Multivariate Matching Methods That Are Equal Percent Bias Reducing, I: Some Examples," *Biometrics* 32:1 (March 1976), 109–120.

——— "Bias Reduction Using Mahalanobis-Metric Matching," *Biometrics* 36 (June 1980), 293–298.

Rubin, Donald B., and Neal Thomas, "Affinely Invariant Matching with Ellipsoidal Distributions," *Annals of Statistics* 20 (June 1992), 1079–1093.

——— "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometric* 52 (March 1996), 249–264.

Seifert, Burkhardt, and Theo Gasser, "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association* 91 (March 1996), 267–275.

——— "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics* (June 2000), 338–360.

Smith, Jeffrey A., and Petra E. Todd, "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *The American Economic Review* 91 (May 2001), 112–118.

——— "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* (forthcoming).

Westat, "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, "Two Essays In Social Program Evaluation," PhD Dissertation, The Johns Hopkins University (2002).

APPENDIX

TABLE A1.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITHOUT REPLACEMENT: TWO COVARIATES, VARIABLE g, FIXED $r = 1$

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: $g = 2$												
Simple mean difference	5.2425	4.2425	1.3449	19.8076	4.9697	3.9697	0.9022	16.5725	5.0227	4.0227	0.5085	16.4407
OLS	1.0905	0.0905	0.4655	0.2240	1.0499	0.0499	0.3347	0.1145	1.0491	0.0491	0.2247	0.529
Propensity-score matching:												
True	2.8702	1.8702	0.7774	4.1020	2.7771	1.7771	0.5764	3.4903	2.7859	1.7859	0.3481	3.3106
Probit	2.8592	1.8592	0.7751	4.0574	2.7729	1.7729	0.5536	3.4496	2.7893	1.7893	0.3395	3.3169
Logit	2.8558	1.8558	0.7706	4.0378	2.7724	1.7724	0.5584	3.4532	2.7880	1.7880	0.3390	3.3119
LPM	2.8680	1.8680	0.7548	4.0591	2.7787	1.7787	0.5647	3.4827	2.7867	1.7867	0.3331	3.3033
Covariate matching (Mahalanobis)	2.8960	1.8960	0.7651	4.1802	2.7762	1.7762	0.5540	3.4618	2.7946	1.7946	0.3340	3.3321
Covariate-and-Propensity:												
True	3.1853	2.1853	0.8190	5.4463	2.9746	1.9746	0.5837	4.2398	2.9194	1.9194	0.3473	3.8047
Probit	3.1720	2.1720	0.8207	5.3911	2.9722	1.9722	0.5751	4.2203	2.9164	1.9164	0.3432	3.7904
Logit	3.1714	2.1714	0.8203	5.3879	2.9714	1.9714	0.5769	4.2192	2.9166	1.9166	0.3428	3.7909
LPM	3.1724	2.1724	0.8187	5.3896	2.9646	1.9646	0.5820	4.1984	2.9145	1.9145	0.3402	3.7810
Covariate-and-outcome	2.4621	1.4621	0.6653	2.5804	2.2860	1.2860	0.4741	1.8786	2.2713	1.2713	0.2883	1.6993
Panel B: $g = 1$												
Simple mean difference	4.6132	3.6132	1.1012	14.2679	4.3950	3.3950	0.7495	12.0878	4.4214	3.4214	0.4573	11.9151
OLS	1.0914	0.0914	0.4237	0.1879	1.0682	0.0682	0.2807	0.0834	1.0455	0.0455	0.2113	0.0467
Propensity-score matching:												
True	2.1991	1.1991	0.6351	1.8412	2.1304	1.1304	0.5110	1.5389	2.1205	1.1205	0.3196	1.3577
Probit	2.2216	1.2216	0.5908	1.8414	2.1253	1.1253	0.4775	1.4943	2.1191	1.1191	0.3012	1.3431
Logit	2.2192	1.2192	0.5943	1.8396	2.1369	1.1369	0.4668	1.5104	2.1265	1.1265	0.3045	1.3617
LPM	2.2125	1.2125	0.5835	1.8106	2.1468	1.1468	0.4637	1.5302	2.1178	1.1178	0.2985	1.3386
Covariate matching (Mahalanobis)	2.2429	1.2429	0.5837	1.8855	2.1462	1.1462	0.4673	1.5321	2.1293	1.1293	0.2935	1.3615
Covariate-and-Propensity:												
True	2.4799	1.4799	0.6260	2.5820	2.3276	1.3276	0.4996	2.0121	2.2504	1.2504	0.3053	1.6567
Probit	2.4807	1.4807	0.6210	2.5781	2.3317	1.3317	0.4852	2.0088	2.2491	1.2491	0.3039	1.6526
Logit	2.4757	1.4757	0.6190	2.5609	2.3310	1.3310	0.4839	2.0057	2.2489	1.2489	0.3045	1.6525
LPM	2.4788	1.4788	0.6274	2.5805	2.3233	1.3233	0.4840	1.9854	2.2492	1.2492	0.3022	1.6518
Covariate-and-outcome	1.9941	0.9941	0.5420	1.2820	1.8611	0.8611	0.3905	0.8940	1.8061	0.8061	0.2616	0.7182
Panel C: $g = 0.5$												
Simple mean difference	3.4050	2.4050	0.7088	6.2864	3.2818	2.2818	0.4627	5.4207	3.3071	2.3071	0.3505	5.4456
OLS	1.0886	0.0886	0.3583	0.1362	1.0655	0.0655	0.2252	0.0550	1.0574	0.0574	0.1727	0.0331
Propensity-score matching:												
True	1.3625	0.3625	0.4785	0.3604	1.3358	0.3358	0.3271	0.2198	1.2955	0.2955	0.2535	0.1516
Probit	1.3817	0.3817	0.4162	0.3189	1.3312	0.3312	0.3025	0.2012	1.3076	0.3076	0.2365	0.1506
Logit	1.3753	0.3753	0.4403	0.3347	1.3366	0.3366	0.2883	0.1964	1.3131	0.3131	0.2318	0.1518
LPM	1.3625	0.3625	0.4418	0.3266	1.3372	0.3372	0.3103	0.2100	1.3097	0.3097	0.2270	0.1474
Covariate matching (Mahalanobis)	1.4244	0.4244	0.4053	0.3444	1.3570	0.3570	0.3075	0.2220	1.3199	0.3199	0.2231	0.1521
Covariate-and-Propensity:												
True	1.5337	0.5337	0.4440	0.4820	1.4577	0.4577	0.3260	0.3158	1.4101	0.4101	0.2295	0.2209
Probit	1.5339	0.5339	0.4491	0.4867	1.4598	0.4598	0.3284	0.3193	1.4127	0.4127	0.2323	0.2243
Logit	1.5354	0.5354	0.4410	0.4811	1.4600	0.4600	0.3248	0.3171	1.4123	0.4123	0.2322	0.2239
LPM	1.5310	0.5310	0.4417	0.4771	1.4619	0.4619	0.3280	0.3209	1.4104	0.4104	0.2320	0.2223
Covariate-and-outcome	1.4331	0.4331	0.4017	0.3489	1.3290	0.3290	0.2891	0.1918	1.2636	0.2636	0.2029	0.1107

Notes: 1. Replications: 200.
2. The numbers in the "Mean" columns are the ratios of the estimated treatment effects to the true treatment effects.
3. MSE is the abbreviation of mean squared error.

TABLE A2.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITHOUT REPLACEMENT: TWO COVARIATES, VARIABLE r, FIXED $g = 1$

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: $r = 0.5$												
Simple mean difference	4.4335	3.4335	1.5463	14.1800	4.2589	3.2589	0.9244	11.4749	4.2439	3.2439	0.4995	10.7724
OLS	1.1418	0.1418	0.5109	0.2811	1.1336	0.1336	0.3443	0.1364	1.1024	0.1024	0.2507	0.0733
Propensity-score matching:												
True	2.0703	1.0703	0.8924	1.9419	2.0589	1.0589	0.5948	1.4751	2.0364	1.0364	0.3527	1.1985
Probit	2.1308	1.1308	0.8311	1.9694	2.0741	1.0741	0.5709	1.4796	2.0394	1.0394	0.3485	1.2018
Logit	2.1287	1.1287	0.8120	1.9333	2.0755	1.0755	0.5632	1.4739	2.0344	1.0344	0.3434	1.1879
LPM	2.1609	1.1609	0.8728	2.1095	2.0749	1.0749	0.5529	1.4611	2.0294	1.0294	0.3342	1.1714
Covariate matching (Mahalanobis)	2.1624	1.1624	0.8515	2.0762	2.0725	1.0725	0.5312	1.4324	2.0301	1.0301	0.3200	1.1635
Covariate-and-Propensity:												
True	2.5221	1.5221	0.9733	3.2641	2.3788	1.3788	0.6006	2.2618	2.2686	1.2686	0.3400	1.7249
Probit	2.5245	1.5245	0.9544	3.2350	2.3713	1.3713	0.5998	2.2402	2.2696	1.2696	0.3449	1.7308
Logit	2.5209	1.5209	0.9519	3.2193	2.3702	1.3702	0.5946	2.2310	2.2694	1.2694	0.3465	1.7314
LPM	2.5222	1.5222	0.9507	3.2209	2.3658	1.3658	0.6007	2.2263	2.2666	1.2666	0.3464	1.7243
Covariate-and-outcome	2.0702	1.0702	0.8029	1.7900	1.9182	0.9182	0.4881	1.0813	1.8297	0.8297	0.3010	0.7790
Panel B: $r = 0.2$												
Simple mean difference	4.3049	3.3049	1.9452	14.7062	4.0019	3.0019	0.8792	9.7844	4.0314	3.0314	0.5519	9.4940
OLS	1.2150	0.2150	0.6415	0.4577	1.1578	0.1578	0.3519	0.1487	1.1267	0.1267	0.2708	0.0894
Propensity-score matching:												
True	2.0618	1.0618	1.0706	2.2736	1.9317	0.9317	0.5797	1.2041	1.9472	0.9472	0.4059	1.0619
Probit	2.1158	1.1158	1.1182	2.4954	1.9415	0.9415	0.5538	1.1931	1.9568	0.9568	0.3772	1.0577
Logit	2.1180	1.1180	1.0915	2.4413	1.9624	0.9624	0.5454	1.2237	1.9500	0.9500	0.3747	1.0429
LPM	2.1235	1.1235	1.0987	2.4694	1.9591	0.9591	0.5456	1.2176	1.9414	0.9414	0.3668	1.0208
Covariate matching (Mahalanobis)	2.1320	1.1320	1.0298	2.3419	1.9607	0.9607	0.5315	1.2054	1.9416	0.9416	0.3510	1.0098
Covariate-and-Propensity:												
True	2.3956	1.3956	1.1963	3.3788	2.2043	1.2043	0.6200	1.8347	2.1828	1.1828	0.3889	1.5503
Probit	2.3652	1.3652	1.1742	3.2425	2.1817	1.1817	0.6011	1.7577	2.1680	1.1680	0.3859	1.5131
Logit	2.3548	1.3548	1.1674	3.1983	2.1802	1.1802	0.5968	1.7490	2.1682	1.1682	0.3855	1.5133
LPM	2.3425	1.3425	1.1625	3.1537	2.1807	1.1807	0.5957	1.7489	2.1618	1.1618	0.3829	1.4964
Covariate-and-outcome	2.2133	1.2133	1.0768	2.6316	1.9601	0.9601	0.5027	1.1745	1.8820	0.8820	0.3341	0.8895
Panel C: $r = 0.1$												
Simple mean difference	4.2253	3.2253	2.0547	14.6244	3.9031	2.9031	0.8861	9.2132	3.9159	2.9159	0.5566	8.8123
OLS	1.2325	0.2325	0.6690	0.5016	1.1630	0.1630	0.3618	0.1575	1.1339	0.1339	0.2837	0.0984
Propensity-score matching:												
True	2.0406	1.0406	1.1835	2.4835	1.8881	0.8881	0.6368	1.1942	1.9013	0.9013	0.4130	0.9829
Probit	2.0773	1.0773	1.1370	2.4533	1.9200	0.9200	0.5766	1.1789	1.9087	0.9087	0.3774	0.9682
Logit	2.0716	1.0716	1.0618	2.2757	1.9199	0.9199	0.5804	1.1831	1.9197	0.9197	0.3921	0.9996
LPM	2.0426	1.0426	1.0852	2.2647	1.9288	0.9288	0.5493	1.1644	1.8996	0.8996	0.3688	0.9453
Covariate matching (Mahalanobis)	2.0848	1.0848	1.0806	2.3445	1.9260	0.9260	0.5333	1.1419	1.9009	0.9009	0.3548	0.9375
Covariate-and-Propensity:												
True	2.2246	1.2246	1.1989	2.9370	2.0683	1.0683	0.6459	1.5585	2.0631	1.0631	0.4066	1.2955
Probit	2.2190	1.2190	1.2104	2.9510	2.0401	1.0401	0.6260	1.4737	2.0523	1.0523	0.4004	1.2677
Logit	2.2125	1.2125	1.2062	2.9251	2.0399	1.0399	0.6276	1.4753	2.0506	1.0506	0.4011	1.2646
LPM	2.1994	1.1994	1.2193	2.9253	2.0419	1.0419	0.6254	1.4767	2.0431	1.0431	0.3950	1.2441
Covariate-and-outcome	2.2481	1.2481	1.1130	2.7965	1.9786	0.9786	0.5177	1.2257	1.8981	0.8981	0.3439	0.9249

Notes: 1. Replications: 200.

2. The numbers in the “Mean” columns are the ratios of the estimated treatment effects to the true treatment effects.

3. MSE is the abbreviation of mean squared error.

TABLE A3.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITHOUT REPLACEMENT: FOUR COVARIATES

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: R^2 of Outcome Equation = 0.6; R^2 of Participation Equation = 0.2												
Simple mean difference	1.489	0.4890	0.2689	0.3114	1.4368	0.4368	0.1724	0.2205	1.4731	0.4731	0.1331	0.2415
OLS with linear specification	1.0975	0.0975	0.1773	0.0409	1.0627	0.0627	0.117	0.0176	1.0853	0.0853	0.0861	0.0147
OLS with correct specification	1.0368	0.0368	0.1507	0.0241	1.0103	0.0103	0.0997	0.0100	1.0307	0.0307	0.0744	0.0065
OLS separately for control and treated	1.0156	0.0156	0.1523	0.0234	0.986	−0.0140	0.1011	0.0104	1.008	0.0080	0.0745	0.0056
Propensity-score matching (true)	1.0684	0.0684	0.2964	0.0925	1.0363	0.0363	0.2035	0.0427	1.0572	0.0572	0.1395	0.0227
Covariate matching (Mahalanobis)	1.1534	0.1534	0.2085	0.0670	1.0873	0.0873	0.1335	0.0254	1.0972	0.0972	0.101	0.0196
Covariate-and-Propensity (true)	1.1458	0.1458	0.1874	0.0564	1.0913	0.0913	0.1358	0.0268	1.1013	0.1013	0.0976	0.0198
Covariate-and-Outcome:												
True	1.156	0.1560	0.2138	0.0700	1.0831	0.0831	0.1312	0.0241	1.0936	0.0936	0.1049	0.0198
Correctly specified	1.1655	0.1655	0.2174	0.0747	1.0975	0.0975	0.1391	0.0289	1.0994	0.0994	0.1069	0.0213
Linear specification	1.1791	0.1791	0.2208	0.0808	1.0971	0.0971	0.133	0.0271	1.0987	0.0987	0.1003	0.0198
Panel B: R^2 of Outcome Equation = 0.4; R^2 of Participation Equation = 0.4												
Simple mean difference	1.7267	0.7267	0.4092	0.6955	1.6946	0.6946	0.2704	0.5556	1.6776	0.6776	0.1899	0.4952
OLS with linear specification	1.1247	0.1247	0.2948	0.1025	1.098	0.0980	0.2031	0.0509	1.1	0.1000	0.145	0.0310
OLS with correct specification	1.0581	0.0581	0.28	0.0818	1.0368	0.0368	0.1971	0.0402	1.0336	0.0336	0.133	0.0188
OLS separately for control and treated	1.0327	0.0327	0.2969	0.0892	1.0209	0.0209	0.2027	0.0415	1.0072	0.0072	0.1371	0.0188
Propensity-score matching (true)	1.2203	0.2203	0.4618	0.2618	1.2019	0.2019	0.2828	0.1207	1.1815	0.1815	0.2121	0.0779
Covariate matching (Mahalanobis)	1.2835	0.2835	0.3661	0.2144	1.2017	0.2017	0.259	0.1078	1.1844	0.1844	0.1674	0.0620
Covariate-and-Propensity (true)	1.2572	0.2572	0.3492	0.1881	1.2036	0.2036	0.2398	0.0990	1.1889	0.1889	0.1626	0.0621
Covariate-and-Outcome:												
True	1.3075	0.3075	0.3642	0.2272	1.2442	0.2442	0.258	0.1262	1.2259	0.2259	0.1703	0.0800
Correctly specified	1.3507	0.3507	0.3689	0.2591	1.2684	0.2684	0.2718	0.1459	1.2295	0.2295	0.1844	0.0867
Linear specification	1.3492	0.3492	0.3818	0.2677	1.2455	0.2455	0.2441	0.1199	1.2079	0.2079	0.1625	0.0696
Panel C: R^2 of Outcome Equation = 0.2; R^2 of Participation Equation = 0.6												
Simple mean difference	2.024	1.0240	0.8073	1.7003	1.917	0.9170	0.4437	1.0378	1.8389	0.8389	0.2918	0.7889
OLS with linear specification	1.1817	0.1817	0.613	0.4088	1.1054	0.1054	0.3835	0.1582	1.0778	0.0778	0.2624	0.0749
OLS with correct specification	1.118	0.1180	0.6153	0.3925	1.0442	0.0442	0.3889	0.1532	1.0245	0.0245	0.2633	0.0699
OLS separately for control and treated	1.1003	0.1003	0.7047	0.5067	1.0325	0.0325	0.4232	0.1802	0.9978	−0.0022	0.2847	0.0811
Propensity-score matching (true)	1.5113	0.5113	0.8803	1.0364	1.3956	0.3956	0.5001	0.4066	1.3347	0.3347	0.3237	0.2168
Covariate matching (Mahalanobis)	1.4968	0.4968	0.7986	0.8846	1.3881	0.3881	0.482	0.3829	1.3357	0.3357	0.2928	0.1984
Covariate-and-Propensity (true)	1.4542	0.4542	0.7625	0.7877	1.3749	0.3749	0.4424	0.3363	1.3275	0.3275	0.305	0.2003
Covariate-and-Outcome:												
True	1.595	0.5950	0.7555	0.9248	1.4405	0.4405	0.4737	0.4184	1.4188	0.4188	0.3048	0.2683
Correctly specified	1.6785	0.6785	0.8414	1.1683	1.5049	0.5049	0.5059	0.5109	1.4301	0.4301	0.3398	0.3005
Linear specification	1.6668	0.6668	0.8116	1.1033	1.5144	0.5144	0.4893	0.5040	1.4254	0.4254	0.3328	0.2917

Notes: 1. Replications: 200.

2. The numbers in the "Mean" columns are the ratios of the estimated treatment effects to the true treatment effects.

3. MSE is the abbreviation of mean squared error.

TABLE A4.—MONTE CARLO RESULTS FOR VARIOUS MATCHING ESTIMATORS WITHOUT REPLACEMENT: EIGHT COVARIATES

Number of observations	$n = 500$				$n = 1000$				$n = 2000$			
Method	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE	Mean	Bias	Std. Error	MSE
Panel A: R^2 of Outcome Equation = 0.6; R^2 of Participation Equation = 0.2												
Simple mean difference	1.7091	0.7091	0.8577	1.2385	1.6227	0.6227	0.2633	0.4571	1.5945	0.5945	0.1988	0.3930
OLS with linear specification	0.7522	−0.2478	0.1936	0.0989	0.7547	−0.2453	0.1424	0.0804	0.7375	−0.2625	0.0928	0.0775
OLS with correct specification	0.6256	−0.3744	0.1679	0.1684	0.6287	−0.3713	0.1279	0.1542	0.6313	−0.3687	0.0829	0.1428
OLS separately for control and treated	1.0071	0.0071	0.1481	0.0220	0.9969	−0.0031	0.0929	0.0086	1.0006	0.0006	0.0716	0.0051
Propensity-score matching (true)	1.0753	0.0753	0.5029	0.2586	1.0109	0.0109	0.2053	0.0423	0.9923	−0.0077	0.1701	0.0290
Covariate matching (Mahalanobis)	1.2137	0.2137	0.4288	0.2295	1.1707	0.1707	0.1627	0.0556	1.15	0.1500	0.1209	0.0371
Covariate-and-Propensity (true)	1.3578	0.3578	0.5098	0.3879	1.2763	0.2763	0.2082	0.1197	1.2649	0.2649	0.1559	0.0945
Covariate-and-Outcome:												
True	1.3776	0.3776	0.552	0.4473	1.3108	0.3108	0.2034	0.1380	1.2744	0.2744	0.1642	0.1023
Correctly specified	1.3596	0.3596	0.4838	0.3634	1.3124	0.3124	0.2069	0.1404	1.2665	0.2665	0.1539	0.0947
Linear specification	1.3811	0.3811	0.4983	0.3935	1.3398	0.3398	0.2359	0.1711	1.325	0.3250	0.1753	0.1364
Panel B: R^2 of Outcome Equation = 0.4; R^2 of Participation Equation = 0.4												
Simple mean difference	1.6474	0.6474	0.4584	0.6293	1.6283	0.6283	0.2723	0.4689	1.5753	0.5753	0.1662	0.3586
OLS with linear specification	0.6985	−0.3015	0.2378	0.1475	0.7028	−0.2972	0.1665	0.1161	0.7116	−0.2884	0.1151	0.0964
OLS with correct specification	0.5322	−0.4678	0.2376	0.2753	0.5532	−0.4468	0.1598	0.2252	0.5538	−0.4462	0.1137	0.2120
OLS separately for control and treated	0.989	−0.0110	0.252	0.0636	1.0106	0.0106	0.1743	0.0305	0.9927	−0.0073	0.1054	0.0112
Propensity-score matching (true)	1.045	0.0450	0.3538	0.1272	1.0361	0.0361	0.2456	0.0616	1.0193	0.0193	0.1628	0.0269
Covariate matching (Mahalanobis)	1.1769	0.1769	0.3273	0.1384	1.2066	0.2066	0.2081	0.0860	1.1547	0.1547	0.1328	0.0416
Covariate-and-Propensity (true)	1.3225	0.3225	0.3946	0.2597	1.3308	0.3308	0.2378	0.1660	1.2757	0.2757	0.1529	0.0994
Covariate-and-Outcome:												
True	1.3311	0.3311	0.3827	0.2561	1.3503	0.3503	0.2498	0.1851	1.2744	0.2744	0.1508	0.0980
Correctly specified	1.3425	0.3425	0.4139	0.2886	1.3417	0.3417	0.245	0.1768	1.281	0.2810	0.157	0.1036
Linear specification	1.3701	0.3701	0.4077	0.3032	1.3846	0.3846	0.2645	0.2179	1.3416	0.3416	0.1625	0.1431
Panel C: R^2 of Outcome Equation = 0.2; R^2 of Participation Equation = 0.6												
Simple mean difference	1.7055	0.7055	0.5765	0.8301	1.6455	0.6455	0.3279	0.5242	1.635	0.6350	0.21	0.4473
OLS with linear specification	0.7063	−0.2937	0.4048	0.2501	0.6718	−0.3282	0.2801	0.1862	0.7008	−0.2992	0.1646	0.1166
OLS with correct specification	0.4739	−0.5261	0.4294	0.4612	0.4531	−0.5469	0.2744	0.3744	0.4745	−0.5255	0.1633	0.3028
OLS separately for control and treated	1.0401	0.0401	0.4842	0.2361	1.0068	0.0068	0.3047	0.0929	1.0142	0.0142	0.1756	0.0310
Propensity-score matching (true)	1.1457	0.1457	0.4943	0.2656	1.1212	0.1212	0.3169	0.1151	1.1042	0.1042	0.2017	0.0515
Covariate matching (Mahalanobis)	1.3057	0.3057	0.4828	0.3265	1.2457	0.2457	0.3226	0.1644	1.2265	0.2265	0.1809	0.0840
Covariate-and-Propensity (true)	1.433	0.4330	0.5376	0.4765	1.4004	0.4004	0.3362	0.2734	1.3608	0.3608	0.2008	0.1705
Covariate-and-Outcome:												
True	1.4352	0.4352	0.5306	0.4709	1.4177	0.4177	0.3332	0.2855	1.3705	0.3705	0.1965	0.1759
Correctly specified	1.4307	0.4307	0.539	0.4760	1.4038	0.4038	0.3308	0.2725	1.3705	0.3705	0.1989	0.1768
Linear specification	1.4158	0.4158	0.514	0.4371	1.4486	0.4486	0.3459	0.3209	1.4321	0.4321	0.1981	0.2260

Notes: 1. Replications: 200.
2. The numbers in the “Mean” columns are the ratios of the estimated treatment effects to the true treatment effects.
3. MSE is the abbreviation of mean squared error.