Bias-Corrected Matching Estimators for Average Treatment Effects

Alberto ABADIE
John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138 and NBER (*alberto_abadie@harvard.edu*)

Guido W. IMBENS
Department of Economics, Harvard University, Cambridge, MA 02138 and NBER (*imbens@harvard.edu*)

In Abadie and Imbens (2006), it was shown that simple nearest-neighbor matching estimators include a conditional bias term that converges to zero at a rate that may be slower than $N^{1/2}$. As a result, matching estimators are not $N^{1/2}$-consistent in general. In this article, we propose a bias correction that renders matching estimators $N^{1/2}$-consistent and asymptotically normal. To demonstrate the methods proposed in this article, we apply them to the National Supported Work (NSW) data, originally analyzed in Lalonde (1986). We also carry out a small simulation study based on the NSW example. In this simulation study, a simple implementation of the bias-corrected matching estimator performs well compared to both simple matching estimators and to regression estimators in terms of bias, root-mean-squared-error, and coverage rates. Software to compute the estimators proposed in this article is available on the authors' web pages (*http://www.economics.harvard.edu/faculty/imbens/software.html*) and documented in Abadie et al. (2003).

KEY WORDS: Selection on observables; Treatment effects.

1. INTRODUCTION

The purpose of this article is to investigate the properties of estimators that combine matching with a bias correction proposed in Rubin (1973) and Quade (1982), and derive the large sample properties of a nonparametric extension of the bias-corrected estimator. We show that a nonparametric implementation of the bias correction removes the conditional bias of matching asymptotically to a sufficient degree so that the resulting estimator is $N^{1/2}$-consistent, without affecting the asymptotic variance.

We apply simple matching estimators and the bias-corrected matching estimators studied in the current article to the National Supported Work (NSW) demonstration data, analyzed originally by Lalonde (1986) and subsequently by many others, including Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2005), and Imbens (2003). For the Lalonde dataset we show that conventional matching estimators without bias correction are sensitive to the choice for the number of matches, whereas a simple implementation of the bias correction using linear least squares is relatively robust to this choice. Moreover, in small simulation studies designed to mimic the data from the NSW application, we find that the simple linear least-squares based implementation of the bias-corrected matching estimator performs well compared to both matching estimators without bias correction, and to regression and weighting estimators, in terms of bias, root-mean-squared-error, and coverage rates for the associated confidence intervals.

Bias-corrected matching estimators combine some of the advantages and disadvantages of both matching and regression estimators. Compared to matching estimators without bias correction, they have the advantage of being $N^{1/2}$-consistent and asymptotically normal irrespective of the number of covariates. However, bias-corrected matching estimators may be more difficult to implement than matching estimators without bias correction if the bias correction is calculated using nonparametric smoothing techniques, and therefore, involves the choice of a smoothing parameter as a function of the sample size. Compared to estimators based on nonparametric regression adjustment without matching (e.g., Hahn 1998; Heckman et al. 1998; Imbens, Newey, and Ridder 2005; Chen, Hong, and Tarozzi 2008) or weighting estimators (Horvitz and Thompson 1952; Robins and Rotnitzky 1995; Hirano, Imbens, and Ridder 2003; Abadie 2005), bias-corrected matching estimators have the advantage of an additional layer of robustness because matching ensures consistency for any given value of the smoothing parameters without requiring accurate approximations to either the regression function or the propensity score. However, in contrast to some regression adjustment and weighting estimators, bias-corrected matching estimator have the disadvantage of not being fully efficient (Abadie and Imbens 2006).

2. MATCHING ESTIMATORS

2.1 Setting and Notation

Matching estimators are often used in evaluation research to estimate treatment effects in the absence of experimental data. As is by now common in this literature, we use Rubin's potential outcome framework (e.g., Rubin 1974). See Rosenbaum (1995) and Imbens and Wooldridge (2009) for surveys. For N units, indexed by $i = 1, \ldots, N$, let W_i be a binary variable that indicates exposure of individual i to treatment, so that $W_i = 1$ if individual i was exposed to treatment, and $W_i = 0$ otherwise. Let $N_0 = \sum_{i=1}^{N}(1 - W_i)$ and $N_1 = \sum_{i=1}^{N} W_i = N - N_0$ be the number of control and treated units, respectively. The variables $Y_i(0)$ and $Y_i(1)$ represent potential outcomes with and without

Journal of Business & Economic Statistics
January 2011, Vol. 29, No. 1
DOI: 10.1198/jbes.2009.07333

treatment, respectively, and therefore, $Y_i(1) - Y_i(0)$ is the treatment effect for unit i. Depending on the value of W_i, one of the two potential outcomes is realized and observed:

$$Y_i = \begin{cases} Y_i(0) & \text{if } W_i = 0 \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

In settings with essentially unrestricted heterogeneity in the effect of the treatment, the typical goal of evaluation research is to estimate an average treatment effect. Here, we focus on the unconditional (population) average

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)].$$

In addition, in the applied literature the focus is often on the average effect for the treated,

$$\tau_{\text{treated}} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1].$$

In the body of the article we will largely focus on τ. In Appendix B we present the corresponding results for τ_{treated}.

In general, a simple comparison of average outcomes between treated and control units does not identify the average effect of the treatment. The reason is that this comparison may be contaminated by the effect of other variables that are correlated with the treatment, W_i, as well as with the potential outcomes, $Y_i(1)$ and $Y_i(0)$. The presence of these confounders may create a correlation between W_i and Y_i even if the treatment has no causal effect on the outcome. Randomization of the treatment eliminates the correlation between any potential confounder and W_i. In the absence of randomization, the following set of assumptions has been found useful as a basis for identification and estimation of τ when all confounders are observed. These observed confounders for unit i will be denoted by X_i, a vector of dimension k, with jth element X_{ij}.

Assumption A.1. Let X be a random vector of dimension k of continuous covariates distributed on $\mathbb{R}^k$ with compact and convex support $\mathbb{X}$, with (a version of the) density bounded and bounded away from zero on its support.

Assumption A.2. For almost every $x \in \mathbb{X}$,

(i) (unconfoundedness) W is independent of $(Y_i(0), Y_i(1))$ conditional on $X_i = x$;

(ii) (overlap) $\eta < \Pr(W_i = 1|X_i = x) < 1 - \eta$, for some $\eta > 0$.

Assumption A.3. $\{(Y_i, W_i, X_i)\}_{i=1}^N$ are independent draws from the distribution of (Y, W, X).

Assumption A.4. Let $\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x]$ and $\sigma_w^2(x) = \mathbb{E}[(Y_i - \mu_w(x))^2|X_i = x]$. Then, (i) $\mu_w(x)$ and $\sigma_w^2(x)$ are Lipschitz in $\mathbb{X}$ for $w = 0, 1$, (ii) $\mathbb{E}[(Y_i(w))^4|X_i = x] \le C$ for some finite C, for almost all $x \in \mathbb{X}$, and (iii) $\sigma_w^2(x)$ is bounded away from zero.

Assumption A.1 requires that all variables in X have a continuous distribution. Notice, however, that discrete covariates with a finite number of support points can be easily accommodated in our analysis by conditioning on their values. Assumption A.2(i) states that, conditional on X_i, the treatment W_i is "as good as randomized," that is, it is independent of the potential outcomes, $Y_i(1)$ and $Y_i(0)$. That will be the case, in particular, if all potential confounders are included in X. Therefore, conditional on $X_i = x$, a simple comparison of average outcomes between treated and control units is equal to the average effect of the treatment given $X_i = x$. This assumption originates in the seminal article by Rosenbaum and Rubin (1983). Assumption A.2(ii) is the usual support condition invoked for matching estimators. Assumption A.2(i) and Assumption A.2(ii) combined are referred to as "strong ignorability." Assumption A.3 refers to the sampling process. Finally, Assumption A.4 collects regularity conditions that will be used later. Note that given Assumption A.2(i), $\mu_w(x) = \mathbb{E}[Y_i|X_i = x, W_i = w]$, and $\sigma_w^2(x) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i = x, W_i = w])^2|X_i = x, W_i = w]$. Abadie and Imbens (2006) discussed Assumptions A.1 through A.4 in greater detail. Identification conditions for matching estimators are also discussed in Hahn (1998), Dehejia and Wahba (1999), Lechner (2002), and Imbens (2004), among others.

As in Abadie and Imbens (2006), we consider matching "with replacement," allowing each unit to be used as a match more than once. For $x \in \mathbb{X}$, and for some positive definite symmetric matrix A, let $\|x\|_A = (x'Ax)^{1/2}$ be some vector norm. Typically the $k \times k$ matrix A is choosen to be the inverse of the sample covariance matrix of the covariates, corresponding to the Mahalanobis metric,

$$A_{\text{maha}} = \left(\frac{1}{N} \sum_{i=1}^N (X_i - \overline{X}) \cdot (X_i - \overline{X})' \right)^{-1},$$

$$\text{where } \overline{X} = \frac{1}{N} \sum_{i=1}^N X_i,$$

or the normalized Euclidean distance, the diagonal matrix with the inverse of the sample variances on the diagonal (e.g., Abadie and Imbens 2006):

$$A_{\text{ne}} = \text{diag}(A_{\text{maha}}^{-1})^{-1}.$$

Let $\ell_m(i)$ be the index of the mth match to unit i. That is, among the units in the opposite treatment group to unit i, unit $\ell_m(i)$ is the mth closest unit to unit i in terms of covariate values. Thus, $\ell_m(j)$ satisfies, (i) $W_{\ell_m(i)} = 1 - W_i$, and (ii)

$$\sum_{j:W_j = 1 - W_i} \mathbf{1}\left\{ \|X_j - X_i\|_A \le \left\|X_{\ell_m(i)} - X_i\right\|_A \right\} = m,$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, equal to 1 if the expression in brackets is true and zero otherwise. For notational simplicity, we ignore ties in the matching, which happen with probability zero if the covariates are continuous. Let $\mathcal{J}_M(i) = \{\ell_1(i), \ldots, \ell_M(i)\}$ denote the set of indices for the first M matches for unit i, for M such that $M \le N_0$ and $M \le N_1$. Finally, let $K_M(i)$ denote the number of times unit i is used as a match if we match each unit to the nearest M matches:

$$K_M(i) = \sum_{l=1}^N \mathbf{1}\{i \in \mathcal{J}_M(l)\}.$$

Under matching without replacement, $K_m(i) \in \{0, 1\}$, but in our setting of matching with replacement, $K_m(i)$ can also take on integer values larger than 1 if unit i is the closest match for multiple units.

2.2 Estimators

If we were to observe the potential outcomes $Y_i(0)$ and $Y_i(1)$ for all units, we will simply estimate τ as the average $\sum_{i=1}^N (Y_i(1) - Y_i(0))/N$. The idea behind matching estimators is to estimate, for each $i = 1, \ldots, N$, the missing potential outcomes. For each i we know one of the potential outcomes, namely $Y_i(0)$ if $W_i = 0$, and $Y_i(1)$ otherwise. Hence, if $W_i = 0$, then we choose $\hat{Y}_i(0) = Y_i(0) = Y_i$, and if $W_i = 1$, then we choose $\hat{Y}_i(1) = Y_i(1) = Y_i$. The remaining potential outcome for unit i is imputed using the average of the outcomes for its matches. This leads to

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 1 \end{cases} \qquad \text{and}$$

$$\hat{Y}_i(1) = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1. \end{cases}$$

Using this notation, we can write the matching estimators for τ based on M matches per unit, with replacement, as

$$\hat{\tau}_M^{\text{m}} = \frac{1}{N} \sum_{i=1}^N (\hat{Y}_i(1) - \hat{Y}_i(0)). \tag{1}$$

Using the definition of $K_M(i)$, we can also write this estimator as a weighted average of the outcomes,

$$\hat{\tau}_M^{\text{m}} = \frac{1}{N} \sum_{i=1}^N (2W_i - 1) \cdot \left(1 + \frac{K_M(i)}{M}\right) \cdot Y_i. \tag{2}$$

This representation is useful for deriving the variance of the matching estimator.

In empirical applications, matching estimators are often implemented with small values for M, as small as 1 even in reasonably large sample sizes. Therefore, in order to obtain an accurate approximation to the finite sample distribution of matching estimators in such settings, we focus asymptotic approximations as N increases for fixed M.

Before introducing the bias-corrected matching estimator, let us briefly discuss regression estimators. Let $\hat{\mu}_w(x)$ be a consistent estimator of $\mu_w(x)$. A regression imputation estimator uses $\hat{\mu}_0(X_i)$ and $\hat{\mu}_1(X_i)$ to impute the missing values of $Y_i(0)$ and $Y_i(1)$, respectively. That is, for

$$\overline{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \hat{\mu}_0(X_i) & \text{if } W_i = 1 \end{cases} \qquad \text{and}$$

$$\overline{Y}_i(1) = \begin{cases} \hat{\mu}_1(X_i) & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1, \end{cases}$$

the regression imputation estimator of τ is

$$\hat{\tau}^{\text{reg}} = \frac{1}{N} \sum_{i=1}^N (\overline{Y}_i(1) - \overline{Y}_i(0)).$$

As in Abadie and Imbens (2006), we classify as regression imputation estimators those for which $\hat{\mu}_w(x)$ is a consistent estimator of $\mu_w(x)$. Various forms of such estimators were proposed by Hahn (1998), Heckman et al. (1998), Chen, Hong, and Tarozzi (2008), and Imbens, Newey, and Ridder (2005). The matching estimators in Equation (1) are similar to the regression imputation estimators, as they can be interpreted as imputing $Y_i(0)$ and $Y_i(1)$ with a nearest-neighbor estimate of $\mu_0(X_i)$ and $\mu_1(X_i)$, respectively. However, because M is held fixed under the matching asymptotics, $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ do not estimate $\mu_0(X_i)$ and $\mu_1(X_i)$ consistently.

Finally, we consider a bias-corrected matching estimator where the difference within the matches is regression-adjusted for the difference in covariate values:

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} \big(Y_j + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_j)\big) & \text{if } W_i = 1 \end{cases}$$

and

$$\tilde{Y}_i(1) = \begin{cases} \dfrac{1}{M} \displaystyle\sum_{j \in \mathcal{J}_M(i)} \big(Y_j + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_j)\big) & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1, \end{cases}$$

with corresponding estimator

$$\hat{\tau}_M^{\text{bcm}} = \frac{1}{N} \sum_{i=1}^N (\tilde{Y}_i(1) - \tilde{Y}_i(0)). \tag{3}$$

Rubin (1979) and Quade (1982) discussed such estimators in the context of matching without replacement and with linear covariance adjustment.

To further illustrate the difference between the simple matching estimator, the regression estimator, and the bias-corrected matching estimator, consider unit i with $W_i = 0$. For this unit, $Y_i(0)$ is known, and only $Y_i(1)$ needs to be imputed. The simple matching estimator imputes the missing potential outcome $Y_i(1)$ as

$$\hat{Y}_i(1) = \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j(1).$$

The regression imputation estimator imputes this missing potential outcome as

$$\overline{Y}_i(1) = \hat{\mu}_1(X_i).$$

The bias-corrected matching estimator imputes the missing potential outcome as

$$\begin{aligned} \tilde{Y}_i(1) &= \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j(1) + \left(\hat{\mu}_1(X_i) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \hat{\mu}_1(X_j)\right) \\ &= \hat{Y}_i(1) + \left(\hat{\mu}_1(X_i) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \hat{\mu}_1(X_j)\right) \\ &= \overline{Y}_i(1) + \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} (Y_j(1) - \hat{\mu}_1(X_j)). \end{aligned}$$

The imputation for the bias-corrected matching estimator adjusts the imputation under the simple matching estimator by the difference in the estimated regression function at X_i and the estimated regression function at the matched values, X_j for $j \in \mathcal{J}_M(i)$. Obviously that will improve the estimator if the estimated regression function is a good approximation to the true regression function. Even if the estimated regression function is noisy, the adjustment will typically be small because $X_i - X_j$ for $j \in \mathcal{J}_M(i)$ should be small in large samples. At the same time,

compared to the regression estimator, the bias-corrected matching estimator adds $\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j(1) - \hat{\mu}_1(X_j)$. If the estimated regression function is equal to the true regression function, this is simply adding noise to the estimator, making it less precise without introducing bias. However, if the regression function is misspecified, the fact that under very weak assumptions the expectation of $\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j(1)$ converges to $\mu_1(X_i)$ implies that bias correction, relative to imputation estimators, is, in expectation, approximately equal to $\mu_1(X_i) - \hat{\mu}_1(X_i)$, which will eliminate any inconsistency in the regression imputation estimator. In other words, the bias-corrected matching estimator is robust against misspecification of the regression function.

2.3 Large Sample Properties of Matching Estimators

Before presenting some results on the large sample properties of the bias-corrected matching estimator, we first collect some results on the large sample properties of matching estimators derived in Abadie and Imbens (2006), which motivated the use of bias-corrected matching estimators.

First, we introduce some additional notation. Let $\mathbf{X}$ be the $N \times k$ matrix with ith row equal to X_i'. Similarly, let $\mathbf{W}$ be $N \times 1$ vector with ith element equal to W_i. Let

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mu_1(x) - \mu_0(x),$$

be the average effect of the treatment conditional on $X = x$, and

$$\overline{\tau(X)} = \frac{1}{N}\sum_{i=1}^{N}(\mu_1(X_i) - \mu_0(X_i)),$$

the average of that over the covariate distribution. For $i = 1, \ldots, N$, define

$$B^{\mathrm{m}}_{M,i} = \mathbb{E}\big[\hat{Y}_i(1) - \hat{Y}_i(0) - (Y_i(1) - Y_i(0))|\mathbf{X}, \mathbf{W}\big]$$
$$= \frac{2W_i - 1}{M}\sum_{j=1}^{M}\big(\mu_{1-W_i}(X_i) - \mu_{1-W_i}\big(X_{\ell_j(i)}\big)\big)$$

and

$$B^{\mathrm{m}}_M = \frac{1}{N}\sum_{i=1}^{N} B^{\mathrm{m}}_{M,i} = \mathbb{E}[\hat{\tau}^{\mathrm{m}}_M - \overline{\tau(X)}|\mathbf{X}, \mathbf{W}]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{2W_i - 1}{M}\sum_{j=1}^{M}\big(\mu_{1-W_i}(X_i) - \mu_{1-W_i}\big(X_{\ell_j(i)}\big)\big).$$

Now we can write the simple matching estimator minus the average treatment effect using simple algebra as

$$\hat{\tau}^{\mathrm{m}}_M - \tau = (\overline{\tau(X)} - \tau) + D^{\mathrm{m}}_M + B^{\mathrm{m}}_M,$$

where

$$D^{\mathrm{m}}_M = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(1 + \frac{K_M(i)}{M}\right)\cdot\big(Y_i - \mu_{W_i}(X_i)\big).$$

The first term, $\overline{\tau(X)} - \tau$, captures the variation in the conditional treatment effect. This term is a simple sample average and satisfies a central limit theorem. The second term D_N has expectation zero conditional on $\mathbf{X}$ and $\mathbf{W}$. This term also satisfies a central limit theorem (Abadie and Imbens 2006). The last term captures the bias conditional on the covariates. This term does not necessarily satisfy a central limit theorem, and our bias-correction approach is geared toward eliminating it.

Next, we turn to the variance. Note that $K_M(i)$ is nonstochastic conditional on $\mathbf{X}$ and $\mathbf{W}$. Therefore, Equation (2) implies that the variance of $\hat{\tau}^{\mathrm{m}}_M$ conditional on $\mathbf{X}$ and $\mathbf{W}$ is

$$\mathbb{V}(\hat{\tau}^{\mathrm{m}}_M|\mathbf{X}, \mathbf{W}) = \mathbb{V}(D^{\mathrm{m}}_M|\mathbf{X}, \mathbf{W}) = \frac{1}{N^2}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2\sigma^2_{W_i}(X_i).$$

Let $V^E = N\cdot\mathbb{V}(\hat{\tau}^{\mathrm{m}}_M|\mathbf{X}, \mathbf{W})$ be the corresponding normalized variance. In addition, let $V^{\tau(X)} = \mathbb{E}[(\tau(X) - \tau)^2]$. The following result is given in Abadie and Imbens (2006):

Theorem 1 (Asymptotic normality for the simple matching estimator). Suppose Assumptions A.1–A.4 hold. Then

$$\big(V^E + V^{\tau(X)}\big)^{-1/2}\sqrt{N}(\hat{\tau}^{\mathrm{m}}_M - B^{\mathrm{m}}_M - \tau) \xrightarrow{d} \mathcal{N}(0, 1).$$

Abadie and Imbens (2006) also proposed a consistent estimator for V^E and $V^{\tau(X)}$ under Assumptions A.1–A.4.

The result of Theorem 1 shows that, after subtracting the conditional bias terms B^{m}_M, the simple matching estimator is $N^{1/2}$-consistent and asymptotically normal. Moreover, Abadie and Imbens (2006) showed that the same result holds without subtracting the conditional bias terms if matching is done for only one covariate (e.g., matching on the true propensity score) because in that case $\sqrt{N}B^{\mathrm{m}}_M = o_p(1)$.

3. BIAS CORRECTED MATCHING

In this section we analyze the properties of the bias-corrected matching estimators, defined in Equation (3). In order to establish the asymptotic behavior of the bias-corrected estimator, we consider a nonparametric series estimator for the two regression functions, $\mu_0(x)$ and $\mu_1(x)$, with $K(N)$ terms in the series, where $K(N)$ increases with N. An important disadvantage of this estimator is that it will rely on selecting smoothing parameters as functions of the sample size, something that the simple matching estimator allows us to avoid. The advantage of the bias-corrected matching estimator is that it is root-N consistent for any dimension of the covariates, k. In both these properties the bias-corrected matching estimator is similar to the regression imputation estimator. However, it has the same large sample variance as the simple matching estimator, and therefore, it is, in general, not as efficient as the regression imputation or weighting estimators in large samples. Compared to the regression imputation estimator, the bias-corrected matching estimator is more robust in the sense that it is consistent for a fixed value of the smoothing parameters. Because choosing smoothing parameters as functions of the sample size is precisely what matching estimators allow us to avoid, in the empirical analysis and simulations of Sections 4 and 5 we investigate the performance of a simple implementation of the bias correction by linear least squares.

Here, we discuss the formal nonparametric implementation of the bias adjustment. Let $\lambda = (\lambda_1, \ldots, \lambda_k)$ be a multi-index of dimension k, that is, a k-dimensional vector of nonnegative integers, with $|\lambda| = \sum_{i=1}^{k}\lambda_i$, and let $x^\lambda = x_1^{\lambda_1}, \ldots, x_k^{\lambda_k}$. Consider a series $\{\lambda(K)\}_{K=1}^{\infty}$ containing all distinct such vectors such that $|\lambda(K)|$ is nondecreasing. Let $p_K(x) = x^{\lambda(K)}$, and let

$p^K(x) = (p_1(x), \ldots, p_K(x))'$. Following Newey (1995), the nonparametric series estimator of the regression function $\mu_w(x)$ is given by

$$\hat{\mu}_w(x) = p^{K(N)}(x)'\Bigg(\sum_{i:W_i=w} p^{K(N)}(X_i)p^{K(N)}(X_i)'\Bigg)^- \times \sum_{i:W_i=w} p^{K(N)}(X_i)Y_i,$$

where $(\cdot)^-$ denotes a generalized inverse. Given the estimated regression function, let $\hat{B}_M^{\mathrm{m}}$ be the estimator for the average bias term:

$$\hat{B}_M^{\mathrm{m}} = \frac{1}{N}\sum_{i=1}^{N}\frac{2W_i-1}{M}\sum_{j=1}^{M}\big(\hat{\mu}_{1-W_i}(X_i) - \hat{\mu}_{1-W_i}\big(X_{\ell_j(i)}\big)\big).$$

Then the bias corrected matching estimator is

$$\hat{\tau}_M^{\mathrm{bcm}} = \hat{\tau}_M^{\mathrm{m}} - \hat{B}_M^{\mathrm{m}}. \tag{4}$$

The following theorem shows that the bias correction removes the bias without affecting the asymptotic variance.

Theorem 2 (Bias-corrected matching estimator). Suppose that Assumptions A.1–A.4 hold. Assume also that (i) the support of X, $\mathbb{X} \subset \mathbb{R}^k$, is a Cartesian product of compact intervals; (ii) $K(N) = O(N^\nu)$, with $0 < \nu < \min(2/(4k+3), 2/(4k^2-k))$; and (iii) there is a constant C such that for each multi-index λ the λth partial derivative of $\mu_w(x)$ exists for $w = 0, 1$ and its norm is bounded by $C^{|\lambda|}$. Then,

$$\sqrt{N}(B_M^{\mathrm{m}} - \hat{B}_M^{\mathrm{m}}) \xrightarrow{p} 0 \qquad \text{and}$$

$$\big(V^E + V^{\tau(X)}\big)^{1/2}\sqrt{N}(\hat{\tau}_M^{\mathrm{bcm}} - \tau) \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. See Appendix A.

The first result implies that we can estimate the bias faster than $N^{1/2}$. This may seem surprising, given that even in parametric settings we can typically not estimate parameters faster than $N^{1/2}$. That logic applies to objects of the type $\mu_w(x) - \mu_w(z)$; for fixed x and w we cannot estimate $\mu_w(x) - \mu_w(z)$ faster than $N^{1/2}$. However, here we are estimating objects of the type $\mu_w(x) - \mu_w(z)$ where $z - x$ goes to zero, allowing us to obtain a faster rate for the difference $\mu_w(x) - \mu_w(z)$. In other words, B_M^{m} itself is $o_p(1)$ [in fact, it is $O_p(N^{-1/k})$, giving us additional room to estimate it at a rate faster than $N^{1/2}$]. The second result says that the bias-corrected matching estimator has the same normalized variance as the simple matching estimator.

4. AN APPLICATION TO THE EVALUATION OF A LABOR MARKET PROGRAM

In this section we apply the estimators studied in this article to data from the National Supported Work (NSW) demonstration, an evaluation of a subsidized work program first analyzed by Lalonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999), Imbens (2003), Smith and Todd (2005), and others. The specific sample we use here is the one employed by Dehejia and Wahba (1999) and is available on Rajeev Dehejia's website. The dataset we use here contains an experimental sample from a randomized evaluation of the NSW program, and also a nonexperimental sample from the Panel Study of Income Dynamics (PSID). Using the experimental data we obtain an unbiased estimate of the average effect of the program. We then compute nonexperimental matching estimators using the experimental participants and the nonexperimental comparison group from the PSID, and compare them to the experimental estimate. In line with previous studies using these data, we focus on the average effect for the treated, and therefore, only match the treated units.

Table 1 presents summary statistics for the three groups used in our analysis. The first two columns present the summary sta-

Table 1. Summary statistics

	Experimental data						Normalized dif.	
	Treated (185)		Controls (260)		PSID (2490)			
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Treat/ Contr	Treat/ PSID
Panel A: Pretreatment variables								
Age	25.8	(7.2)	25.1	(7.1)	34.9	(10.4)	0.08	−0.71
Education	10.3	(2.0)	10.1	(1.6)	12.1	(3.1)	0.10	−0.48
Black	0.84	(0.36)	0.83	(0.38)	0.25	(0.43)	0.03	1.05
Hispanic	0.06	(0.24)	0.11	(0.31)	0.03	(0.18)	−0.12	0.09
Married	0.19	(0.39)	0.15	(0.36)	0.87	(0.34)	0.07	−1.30
Earnings 13–24	2.10	(4.89)	2.11	(5.69)	19.43	(13.41)	−0.00	−1.21
Unemployed 13–24	0.71	(0.46)	0.75	(0.43)	0.09	(0.28)	−0.07	1.16
Earnings '75	1.53	(3.22)	1.27	(3.10)	19.06	(13.60)	0.06	−1.25
Unemployed '75	0.60	(0.49)	0.68	(0.47)	0.10	(0.30)	−0.13	0.87
Panel B: Outcomes								
Earnings '78	6.35	7.87	4.55	5.48	21.55	15.56	0.19	−0.87
Unemployed '78	0.24	0.43	0.35	0.48	0.11	0.32	−0.17	0.24

NOTE: Earnings data are in thousands of 1978 dollars. Earnings 13–24 and Unemployed 13–24 refers to earnings and unemployment during the period 13 to 24 months prior to randomization.

tistics for the experimental treatment group. The second pair of columns presents summary statistics for the experimental controls. The third pair of columns presents summary statistics for the nonexperimental comparison group constructed from the PSID. The last two columns present normalized differences between the covariate distributions, between the experimental treated and controls, and between the experimental treated and the PSID comparison group, respectively. These normalized differences are calculated as

$$\text{nor-dif} = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{(S_0^2 + S_1^2)/2}},$$

where $\overline{X}_w = \sum_{i:W_i=w} X_i/N_w$ and $S_w^2 = \sum_{i:W_i=w}(X_i - \overline{X}_w)^2/(N_w - 1)$. Note that this differs from the t-statistic for the test of the null hypothesis that $\mathbb{E}[X|W = 0] = \mathbb{E}[X|W = 1]$, which will be

$$t\text{-stat} = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S_0^2/N_0 + S_1^2/N_1}}.$$

The normalized difference provides a scale-free measure of the difference in the location of the two distributions, and is useful for assessing the degree of difficulty in adjusting for differences in covariates.

Panel A contains the results for pretreatment variables and Panel B for outcomes. Notice the large differences in background characteristics between the program participants and the PSID sample. This is what makes drawing causal inferences from comparisons between the PSID sample and the treatment group a tenuous task. From Panel B, we can obtain an unbiased estimate of the effect of the NSW program on earnings in 1978 by comparing the averages for the experimental treated and controls, $6.35 - 4.55 = 1.80$, with a standard error of 0.67 (earnings are measured in thousands of dollars). Using a normal approximation to the limiting distribution of the effect of the program on earnings in 1978, we obtain a 95% confidence interval, which is [0.49, 3.10].

Table 2 presents estimates of the causal effect of the NSW program on earnings using various matching, regression, and weighting estimators. Panel A reports estimates for the experimental data (treated and controls). Panel B reports estimates based on the experimental treated and the PSID comparison group. The first set of rows in each case reports matching estimates for M equal to 1, 4, 16, 64, and 2490 (the size of the PSID comparison group). The matching estimates include simple matching with no-bias adjustment and bias-adjusted matching, first by matching on the covariates and then by matching on the estimated propensity score. The covariate matching estimators use the matrix A_{ne} (the diagonal matrix with inverse sample variances on the diagonal) as the distance measure. Because we are focused on the average effect for the treated, the bias correction only requires an estimate of $\mu_0(X_i)$. We estimate this regression function using linear regression on all nine pretreatment covariates in Table 1, panel A, but do not include any higher order terms or interactions, with only the control units that are used as a match [the units j such that $W_j = 0$ and

Table 2. Experimental and nonexperimental estimates for the NSW data

	$M = 1$		$M = 4$		$M = 16$		$M = 64$		$M = 2490$	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Panel A:										
Experimental estimates										
Covariate matching	1.22	(0.84)	1.99	(0.74)	1.75	(0.74)	2.20	(0.70)	1.79	(0.67)
Bias-adjusted cov matching	1.16	(0.84)	1.84	(0.74)	1.54	(0.75)	1.74	(0.71)	1.72	(0.68)
Pscore matching	1.43	(0.81)	1.95	(0.69)	1.85	(0.69)	1.85	(0.68)	1.79	(0.67)
Bias-adjusted pscore matching	1.22	(0.81)	1.89	(0.71)	1.78	(0.70)	1.67	(0.69)	1.72	(0.68)
Regression estimates										
Mean difference	1.79	(0.67)								
Linear	1.72	(0.68)								
Quadratic	2.27	(0.80)								
Weighting on pscore	1.79	(0.67)								
Weighting and linear regression	1.69	(0.66)								
Panel B:										
Nonexperimental estimates										
Simple matching	2.07	(1.13)	1.62	(0.91)	0.47	(0.85)	−0.11	(0.75)	−15.20	(0.61)
Bias-adjusted matching	2.42	(1.13)	2.51	(0.90)	2.48	(0.83)	2.26	(0.71)	0.84	(0.63)
Pscore matching	2.32	(1.21)	2.06	(1.01)	0.79	(1.25)	−0.18	(0.92)	−1.55	(0.80)
Bias-adjusted pscore matching	3.10	(1.21)	2.61	(1.03)	2.37	(1.28)	2.32	(0.94)	2.00	(0.84)
Regression estimates										
Mean difference	−15.20	(0.66)								
Linear	0.84	(0.88)								
Quadratic	3.26	(1.04)								
Weighting on pscore	1.77	(0.67)								
Weighting and linear regression	1.65	(0.66)								

NOTE: The outcome is earnings in 1978 in thousands of dollars.

$j \in \mathcal{J}_M(i)$ for some i]. The confidence intervals are based on the variance estimator proposed in Abadie and Imbens (2006). This variance estimator is formally justified for the case of matching on the covariates. It does not cover the case of matching on the estimated propensity score. We implement it for the case of matching on the estimated propensity score by ignoring the estimation error in the propensity score. The next three rows of each panel report estimates based on differences in means, linear regression including terms for all covariates, and linear regression also including quadratic terms and a full set of interactions, respectively. The last two rows in each panel report estimates for weighting estimators. Both rows use weights for the treated units equal to 1, and weights for the control units equal to the propensity score divided by 1 minus the propensity score. Then we normalize the weights in both groups to add up to N_1. The first weighting estimator is based solely on weighting, the second one uses weighted regression, with the nine pretreatment variables included.

The experimental estimates in Panel A range from 1.16 (bias-corrected matching with one match) to 2.27 (quadratic regression). The nonexperimental estimates in Panel B have a much wider range, from −15.20 (simple difference) to 3.26 (quadratic regression). For the nonexperimental sample, using a single match, there is little difference between the simple matching estimator and its bias-corrected version, 2.07 and 2.42, respectively. However, simple matching without bias-correction produces radically different estimates when the number of matches changes; a troubling result for the empirical implementation of these estimators. With $M \geq 16$, the simple matching estimator produces results outside the experimental 95% confidence interval. In contrast, the bias-corrected matching estimator shows a much more robust behavior when the number of matches changes: only with $M = 2490$ (that is, when all units in the comparison group are matched to each treated) the bias-corrected estimate deteriorates to 0.84, still inside the experimental 95% confidence interval.

To see how well the simple matching estimator performs in terms of balancing the covariates, Table 3 reports average differences within the matched pairs. First, all the covariates are normalized to have zero mean and unit variance. The first two columns report the averages of the normalized covariates for the PSID comparison group and the experimental treated. Before matching, the averages for some of the variables are more than one standard deviation apart, e.g., the earnings and employment variables. The next pair of columns reports the within-matched-pairs average difference and the standard deviation of this within-pair difference. For all the indicator variables the matching is exact. The other, more continuously distributed variables are not matched exactly, but the quality of the matches appears very high: the average difference within the pairs is very small compared to the average difference between treated and comparison units before the matching, and it is also small compared to the standard deviations of these differences. If we increase the number of matches the quality of the matches goes down, with even the indicator variables no longer matched exactly, but in most cases the average difference is still very small until we get to 16 or more matches. As expected, match quality deteriorates when the number of matches increases. This explains why, as shown in Table 2, the bias correction matters more for larger M. The last row reports matching differences for logistic estimates of the propensity score. Although here we match on the covariates directly, rather than on the propensity score, the matching still greatly reduces differences in the propensity score. With a single match ($M = 1$) the average difference in the propensity score is only 0.21, whereas without matching the difference between treated and comparison units is 8.16, almost 40 times higher.

5. A MONTE CARLO STUDY

In this section, we discuss some simulations designed to assess the performance of the various matching estimators. To get a realistic sense of the performance of the various estimators, we simulated datasets that aim to resemble actual datasets. For other Monte Carlo studies of matching type estimators see Zhao (2002), Frölich (2004), and Busso, DiNardo, and McCrary (2009). An additional Monte Carlo study based on data collected by Imbens, Rubin, and Sacerdote (2001) is available in a previous version of this article.

Table 3. Mean covariate differences in matched groups

	Average		$M = 1$		$M = 4$		$M = 16$		$M = 64$		$M = 2490$	
	PSID	Treated	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Age	0.06	−0.80	−0.02	(0.65)	−0.06	(0.60)	−0.30	(0.41)	−0.57	(0.57)	−0.86	(0.68)
Education	0.04	−0.54	−0.10	(0.44)	−0.20	(0.48)	−0.25	(0.39)	−0.24	(0.42)	−0.58	(0.66)
Black	−0.09	1.21	−0.00	(0.00)	0.09	(0.32)	0.35	(0.47)	0.70	(0.66)	1.30	(0.80)
Hispanic	−0.01	0.14	−0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.01	(0.03)	0.15	(1.30)
Married	0.12	−1.64	0.00	(0.00)	−0.06	(0.30)	−0.33	(0.46)	−0.90	(0.85)	−1.76	(1.02)
Earnings 13–24	0.09	−1.18	−0.01	(0.10)	−0.01	(0.12)	−0.05	(0.17)	−0.15	(0.30)	−1.26	(0.36)
Unemployed 13–24	−0.13	1.72	0.00	(0.00)	0.02	(0.17)	0.24	(0.40)	0.41	(0.72)	1.85	(1.36)
Earnings '75	0.09	−1.18	−0.04	(0.17)	−0.07	(0.15)	−0.11	(0.19)	−0.19	(0.26)	−1.26	(0.23)
Unemployed '75	−0.10	1.36	0.00	(0.00)	0.00	(0.05)	0.03	(0.28)	0.10	(0.41)	1.46	(1.44)
Log odds Prop score	−7.08	1.08	0.21	(0.99)	0.56	(1.13)	1.70	(1.14)	3.20	(1.49)	8.16	(2.13)

NOTE: In this table all covariates have been normalized to have mean zero and unit variance. The first two columns present the averages for the experimental treated and the PSID comparison units. The remaining pairs of columns present the average difference within the matched pairs and the standard deviation of this difference for matching based on 1, 4, 16, 64, and 2490 matches. For the last variable the logarithm of the odds ratio of the propensity score is used. This log odds ratio has mean −6.52 and standard deviation 3.30 in the sample.

Table 4. Simulation results Lalonde design (10,000 replications)

M	Estimator	Mean bias	Median bias	RMSE	MAE	SD	Mean SE	Coverage rate (95% CI)	Coverage rate (90% CI)
1	Covariate matching	−0.91	−0.81	1.36	0.86	1.02	1.44	0.97	0.94
	Bias-adj cov match	0.16	0.22	1.33	0.81	1.32	1.43	0.96	0.92
4	Covariate matching	−1.33	−1.25	1.57	1.25	0.84	1.22	0.93	0.83
	Bias-adj cov match	0.24	0.28	1.18	0.75	1.15	1.20	0.95	0.91
16	Covariate matching	−2.04	−2.00	2.17	2.00	0.74	1.11	0.61	0.41
	Bias-adj cov match	0.43	0.45	1.13	0.77	1.05	1.08	0.93	0.87
64	Covariate matching	−3.07	−3.08	3.13	3.08	0.61	1.01	0.09	0.02
	Bias-adj cov match	0.57	0.59	1.05	0.75	0.88	0.95	0.92	0.87
1	Pscore matching	−0.02	0.29	1.72	0.91	1.72	1.64	0.98	0.95
	Bias-adj pscore match	0.09	0.22	1.38	0.84	1.38	1.64	0.98	0.96
4	Pscore matching	−0.29	−0.07	1.36	0.77	1.33	1.43	0.98	0.96
	Bias-adj pscore match	0.09	0.22	1.23	0.74	1.22	1.45	0.97	0.95
16	Pscore matching	−0.89	−0.82	1.39	0.90	1.06	1.33	0.98	0.93
	Bias-adj pscore match	0.14	0.21	1.17	0.75	1.17	1.36	0.97	0.94
64	Pscore matching	−1.69	−1.66	1.89	1.66	0.84	1.13	0.78	0.62
	Bias-adj pscore match	0.27	0.31	1.06	0.72	1.03	1.16	0.96	0.92
	Mean difference	−22.41	−22.41	22.42	22.41	0.76	1.03	0.00	0.00
	Linear regression	−0.27	−0.25	1.33	0.88	1.30	1.48	0.97	0.94
	Quadratic regression	3.35	3.36	3.90	3.36	1.99	2.05	0.62	0.49
	Weighting on pscore	−0.14	−0.07	1.28	0.79	1.27	1.28	0.96	0.92
	Weighting and regression	0.15	0.19	1.09	0.68	1.08	1.13	0.97	0.94

The simulations are designed to mimic the Lalonde data. In each of the 10,000 replications we draw 185 treated and 2490 control observations and calculate 21 estimators for the average effect on the treated, τ_{treated}. In the simulation we have eight regressors, designed to match the following variables in the NSW dataset: `age`, `educ`, `black`, `married`, `re74`, `u74`, `re75`, and `u75`. In Appendix C we describe the precise data generating process for the simulations. For each estimator we report the mean and median bias, the root-mean-squared-error (RMSE), the median-absolute-error (MAE), the standard deviation, the average estimated standard error, and the coverage rates for nominal 95% and 90% confidence intervals based on the matching estimator for the variance. We implemented an extremely simple version of the bias adjustment, using only linear terms in the covariates. The results are reported in Table 4.

In terms of RMSE and MAE, the bias-adjusted matching estimator is best with 64 matches, but with this many matches the bias is substantial. With four matches the bias is considerably smaller, and the actual coverage rates of the 90% and 95% confidence intervals is close to the nominal coverage rate. The simple (not bias-adjusted) matching estimator does not perform as well, in terms of bias or RMSE. The pure regression adjustment estimators perform poorly. They have high RMSE and substantial bias. Coverage rates of confidence intervals centered on the bias-corrected matching estimator are closer to nominal levels than those centered on the simple matching estimator. Confidence intervals for the quadratic regression estimator have substantially lower than nominal coverage rates, although coverage rates for the linear regression estimator are close to the nominal rates. In this setting the weighting estimators do fairly well, both in terms of coverage rates and in terms of RMSE.

6. CONCLUSION

We propose a nonparametric bias-adjustment that renders matching estimators $N^{1/2}$-consistent. In simulations based on a realistic setting for nonexperimental program evaluations, a simple implementation of this estimator, where the bias-adjustment is based on linear regression, performs well compared to both matching estimators without bias-adjustment and regression-based estimators in terms of bias and mean-squared error. It also has good coverage rates for 90% and 95% confidence intervals, suggesting it may be a useful estimator in practice.

APPENDIX A: PROOFS

Before proving Theorem 2 we state two auxiliary lemmas. Let λ be a multi-index of dimension k, that is, a k-dimensional vector of nonnegative integers, with $|\lambda| = \sum_{i=1}^{k} \lambda_i$, and let Λ_l be the set of λ such that $|\lambda| = l$. Furthermore, let $x^\lambda = x_1^{\lambda_1}, \ldots, x_k^{\lambda_k}$, and let $\partial^\lambda g(x) = \partial^{|\lambda|} g(x) / \partial x_1^{\lambda_1}, \ldots, \partial x_k^{\lambda_k}$. For $d \geq 0$, define $|g|_d = \max_{|\lambda| \leq d} \sup_x |\partial^\lambda g(x)|$.

Lemma A.1 (Uniform convergence of series estimators of regression functions, Newey 1995). Suppose the conditions in Theorem 2 hold. Then for any $\xi > 0$ and nonnegative integer d,

$$|\hat{\mu}_w - \mu_w|_d = O_p\big(K^{1+2d}\big((K/N)^{1/2} + K^{-\xi}\big)\big)$$

for $w = 0, 1$.

Proof. Assumptions 3.1, 4.1, 4.2, and 4.3 in Newey (1995) are satisfied for $\mu_w(x)$ and $N_w \to \infty$, implying that Newey's theorems 4.2 and 4.4 apply. The result of the lemma holds because $N/N_w = O_p(1)$ for $w = 0, 1$.

Lemma A.2 (Unit-level bias correction). Suppose the conditions in Theorem 2 hold. Then

$$\max_{i=1,\ldots,N} \left|\hat{\mu}_w(X_i) - \hat{\mu}_w\big(X_{j_m(i)}\big) - \big(\mu_w(X_i) - \mu_w\big(X_{j_m(i)}\big)\big)\right| = o_p\big(N^{-1/2}\big)$$

for $w = 0, 1$.

Proof. Let $U_{m,i} = X_{j_m(i)} - X_i$. Use a Taylor series expansion around X_i to write

$$\left|\mu_w\big(X_{j_m(i)}\big) - \mu_w(X_i) - \sum_{1\le l\le k-1} \frac{1}{l!} \sum_{\lambda\in\Lambda_l} \partial^\lambda \mu_w(X_i) U_{m,i}^\lambda\right| \le \frac{C^k}{k!} \sum_{\lambda\in\Lambda_k} |U_{m,i}^\lambda| \le \frac{C^k}{k!} \sum_{\lambda\in\Lambda_k} \|U_{m,i}\|^k.$$

Because all moments of $N_{1-W_i}^{1/k}\|U_{m,i}\|$ and N/N_{1-W_i} are uniformly bounded, applying Bonferroni's and Markov's inequalities, we obtain that for any $\varepsilon > 0$:

$$\max_{i=1,\ldots,N} \left|\mu_w\big(X_{j_m(i)}\big) - \mu_w(X_i) - \sum_{1\le l\le k-1} \frac{1}{l!} \sum_{\lambda\in\Lambda_l} \partial^\lambda \mu_w(X_i) U_{m,i}^\lambda\right| = o_p(N^{-1+\varepsilon}).$$

Because we can choose $\varepsilon \le 1/2$, it follows that the left-hand side of the last equation is $o_p(N^{-1/2})$. Similarly, for any $\varepsilon > 0$:

$$\left|\hat{\mu}_w\big(X_{j_m(i)}\big) - \hat{\mu}_w(X_i) - \sum_{1\le l\le k-1} \frac{1}{l!} \sum_{\lambda\in\Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) U_{m,i}^\lambda\right| \le \frac{1}{k!} \sum_{\lambda\in\Lambda_k} |\hat{\mu}_w - \mu_w|_k \|U_{m,i}\|^k + \frac{C^k}{k!} \sum_{\lambda\in\Lambda_k} \|U_{m,i}\|^k.$$

Therefore, for arbitrary $\xi > 0$ and $\varepsilon > 0$:

$$\max_{i=1,\ldots,N} \left|\hat{\mu}_w\big(X_{j_m(i)}\big) - \hat{\mu}_w(X_i) - \sum_{1\le l\le k-1} \frac{1}{l!} \sum_{\lambda\in\Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) U_{m,i}^\lambda\right| = O_p\big(K^{1+2k}\big((K/N)^{1/2} + K^{-\xi}\big)\big) o_p(N^{-1+\varepsilon}) + o_p(N^{-1+\varepsilon}).$$

Because $\nu < 2/(4k+3)$, we can choose ξ and ε so that the left-hand side of the last equation becomes $o_p(N^{-1/2})$. Therefore,

$$\begin{aligned}
&\max_{i=1,\ldots,N} \left|\hat{\mu}_w\big(X_{j_m(i)}\big) - \hat{\mu}_w(X_i) - \big(\mu_w\big(X_{j_m(i)}\big) - \mu_w(X_i)\big)\right| \\
&\quad\le \max_{i=1,\ldots,N} \sum_{1\le l\le k-1} \frac{1}{l!} \sum_{\lambda\in\Lambda_l} |\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)| \cdot |U_{m,i}^\lambda| + o_p\big(N^{-1/2}\big) \\
&\quad\le |\hat{\mu}_w - \mu_w|_{k-1} \sum_{1\le l\le k-1} \frac{1}{l!} \sum_{\lambda\in\Lambda_l} \max_{i=1,\ldots,N} \|U_{m,i}\|^{|\lambda|} + o_p\big(N^{-1/2}\big) \\
&\quad= O_p\big(K^{2k-1}\big((K/N)^{1/2} + K^{-\xi}\big)\big) o_p\big(N^{-1/k+\varepsilon}\big) + o_p\big(N^{-1/2}\big),
\end{aligned}$$

for arbitrary $\xi > 0$ and $\varepsilon > 0$. Consider for a particular $\lambda \in \Lambda_l$ the term $(\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)) \cdot U_{m,i}^\lambda$. The second factor is, using the same argument as before, of order $O_p(N^{-l/k})$, Since $l \ge 1$, the second factor is at most $O_p(N^{-1/k})$, and because all the relevant moments exist $\max_i U_{m,i}^\lambda = o_p(N^{-1/k+\varepsilon})$ for any $\varepsilon > 0$. Now consider the first factor. By Lemma A.1, $|\sup(\partial^\lambda \hat{\mu}_w(x) - \partial^\lambda \mu_w(x))|$ is of order $O_p(K^{1+2k}((K/N)^{1/2} + K^{-\alpha}))$. Now, it can be easily seen that $\nu < 2/(4k^2 - k)$ guarantees that the result of Lemma A.2 holds.

Proof of Theorem 2

We focus on the result for the average treatment effect. The second part of the theorem for the average effect for the treated follows the same pattern. The difference $|\hat{B}_M^{\mathrm{m}} - B_M^{\mathrm{m}}|$ can be written as

$$\begin{aligned}
&|\hat{B}_M^{\mathrm{m}} - B_M^{\mathrm{m}}| \\
&\quad\le \frac{1}{N} \sum_{i=1}^N \frac{1}{M} \sum_{i=1}^M \left|\hat{\mu}_{1-W_i}(X_i) - \hat{\mu}_{1-W_i}\big(X_{j_m(i)}\big) - \big(\mu_{1-W_i}(X_i) - \mu_{1-W_i}\big(X_{j_m(i)}\big)\big)\right| \\
&\quad\le \max_{i=1,\ldots,N} \sum_{w=0,1} \left|\hat{\mu}_w(X_i) - \hat{\mu}_w\big(X_{j_m(i)}\big) - \big(\mu_w(X_i) - \mu_w\big(X_{j_m(i)}\big)\big)\right| \\
&\quad= o_p\big(N^{-1/2}\big),
\end{aligned}$$

by Lemma A.2.

APPENDIX B: THE AVERAGE EFFECT ON THE TREATED

Here we present the results for the average effect on the treated without proof. The formal proofs are similar to those for the case of the overall average effect. Estimation of τ_{treated} requires weaker assumptions than estimation of τ. In particular, Assumptions A.2 and A.3 can be weakened as follows.

Assumption A.2′. For almost every $x \in \mathbb{X}$,

(i) W is independent of $Y(0)$ conditional on $X = x$;
(ii) $\Pr(W = 1|X = x) < 1 - \eta$, for some $\eta > 0$.

Assumption A.3′. Conditional on $W_i = w$ the sample consists of independent draws from $Y, X|W = w$, for $w = 0, 1$. For some $r \ge 1$, $N_1^r/N_0 \to \theta$, with $0 < \theta < \infty$.

Using the same definition for $\hat{Y}_i(0)$ as before, we now estimate τ_{treated} and $\tau_{\text{cond,treated}}$ as

$$\hat{\tau}_{M,\text{treated}}^{\mathrm{m}} = \frac{1}{N_1} \sum_{W_i=1} (Y_i - \hat{Y}_i(0)) = \frac{1}{N_1} \sum_{W_i=1} \left(W_i - (1 - W_i)\frac{K_M(i)}{M}\right) Y_i.$$

Define the average bias

$$B_{M,\text{treated}}^{\mathrm{m}} = \frac{1}{N_1} \sum_{i=1}^N \frac{W_i}{M} \sum_{m=1}^M \big(\mu_0(X_i) - \mu_0\big(X_{j_m(i)}\big)\big),$$

the estimator for the average bias

$$\hat{B}_{M,\text{treated}}^{\mathrm{m}} = \frac{1}{N_1} \sum_{i=1}^N \frac{W_i}{M} \sum_{m=1}^M \big(\hat{\mu}_0(X_i) - \hat{\mu}_0\big(X_{j_m(i)}\big)\big),$$

the bias-corrected estimator for the average effect on the treated

$$\hat{\tau}^{\text{bcm}}_{M,\text{treated}} = \hat{\tau}^{\text{m}}_{M,\text{treated}} - \hat{B}^{\text{m}}_{M,\text{treated}},$$

the conditional variance

$$\mathbb{V}(\hat{\tau}^{\text{m}}_{M,\text{treated}}|\mathbf{X},\mathbf{W}) = \frac{1}{N_1^2}\sum_{i=1}^{N}\left(W_i - (1-W_i)\frac{K_M(i)}{M}\right)^2\sigma^2(X_i,W_i),$$

and its normalized version,

$$V^E_{\text{treated}} = N_1 \cdot \mathbb{V}(\hat{\tau}^{\text{m}}_{M,\text{treated}}|\mathbf{X},\mathbf{W}).$$

Also define $V^{\tau(X)}_{\text{treated}} = \mathbb{E}[(\tau(X) - \tau_{\text{treated}})^2|W = 1]$. Then the equivalent of Theorems 1 and 2 is:

Theorem 1′ (Asymptotic normality for the simple matching estimator for the average effect on the treated). Suppose Assumptions A.1, A.2′, A.3′, and A.4 hold. Then

$$\left(V^E_{\text{treated}} + V^{\tau(X)}_{\text{treated}}\right)^{-1/2}\sqrt{N_1} \times (\hat{\tau}^{\text{m}}_{M,\text{treated}} - B^{\text{m}}_{M,\text{treated}} - \tau_{\text{treated}}) \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem 2′ (Bias-corrected matching estimator for the average effect on treated). Suppose that Assumptions A.1, A.2′, A.3′, and A.4 hold. Assume also that (i) the support of X, $\mathbb{X} \subset \mathbb{R}^k$, is a Cartesian product of compact intervals; (ii) $K(N) = O(N^\nu)$, with $0 < \nu < \min(2/(4k+3), 2/(4k^2-k))$; and (iii) there is a constant C such that for each multi-index λ the λth partial derivative of $\mu_w(x)$ exists for $w = 0, 1$ and its norm is bounded by $C^{|\lambda|}$. Then,

$$\sqrt{N_1}(B^{\text{m}}_{M,\text{treated}} - \hat{B}^{\text{m}}_{M,\text{treated}}) \xrightarrow{p} 0$$

and

$$\left(V^E_{\text{treated}} + V^{\tau(X)}_{\text{treated}}\right)^{1/2}\sqrt{N_1}(\hat{\tau}^{\text{bcm}}_{M,\text{treated}} - \tau_{\text{treated}}) \xrightarrow{d} \mathcal{N}(0,1).$$

APPENDIX C: DATA GENERATING PROCESSES FOR THE SIMULATIONS

(i) *Covariates*. We draw separately from the covariate distribution given $W_i = 0$ and $W_i = 1$. The covariates are grouped into a set of five subsets, and between the subsets the covariates are drawn independently. The groups of covariates are {age}, {educ}, {black}, {married}, and {u74, u75, re74, re75}. For each of the joint distributions within a subset the distribution follows fairly closely to the corresponding distribution in the Lalonde data.

Among the control observations, the log of age has a normal distribution with mean 3.50 and standard deviation 0.30. Among the treated it has a normal distribution with mean 3.22 and standard deviation 0.25.

educ has a discrete distribution with points of support of all integers between 4 and 16. The probabilities for the 13 points of support for the controls are 0.02, 0.01, 0.02, 0.02, 0.06, 0.04, 0.07, 0.07, 0.34, 0.05, 0.08, 0.02, and 0.20. For the treated the probabilities are 0.02, 0.01, 0.01, 0.01, 0.10, 0.15, 0.17, 0.24, 0.21, 0.04, 0.02, 0.01, and 0.01.

black has a discrete distribution with, among the controls, the population fraction of blacks is 0.25. Among the treated the fraction is 0.84.

Among the controls, married has a binary distribution with mean 0.87, and among the treated it has a binary distribution with mean 0.19.

Among the controls the pr((u74, u75) = (0, 0)) = 0.88, pr((u74, u75) = (0, 1)) = 0.03 pr((u74, u75) = (1, 0)) = 0.02, and pr((u74, u75) = (1, 1)) = 0.07. Among the treated pr((u74, u75) = (0, 0)) = 0.28, pr((u74, u75) = (0, 1)) = 0.01, pr((u74, u75) = (1, 0)) = 0.12, and pr((u74, u75) = (1, 1)) = 0.59.

Among the controls, conditional on (u74, u75) = (0, 1), the log of re75 has a normal distribution with mean 2.32 and standard deviation 1.45, conditional on (u74, u75) = (1, 0), the log of re74 has a normal distribution with mean 2.16 and standard deviation 1.20, and conditional on (u74, u75) = (0, 0), the log of (re74, re75) has a joint normal distribution with mean (2.88, 2.86) and covariance matrix $\begin{pmatrix} 0.47 & 0.37 \\ 0.3700 & 0.5000 \end{pmatrix}$.

Among the treated, conditional on (u74, u75) = (0, 1), the log of re75 has a normal distribution with mean 0.60 and standard deviation 0.88, conditional on (u74, u75) = (1, 0), the log of re74 has a normal distribution with mean 1.26 and variance 0.78, and conditional on (u74, u75) = (0, 0), the log of (re74, re75) has a joint normal distribution with mean (1.53, 0.93) and covariance matrix $\begin{pmatrix} 1.08 & 0.57 \\ 0.57 & 1.42 \end{pmatrix}$.

(ii) *Conditional Outcome Distribution Given Covariates*. Conditional on the covariates, the outcome Y_i has a mixed discrete/continuous distribution, with separate coefficients for the controls and treated. The probability that the outcome is positive conditional on the covariate taking on the value x is $\exp(\gamma_w' h(x))/(1 + \exp(\gamma_w' h(x))$, for $w = 0, 1$. The vector of covariates contains 16 elements: an intercept, age, educ, black, married, u74, u75, re74, re75, black × u74, black × u75, black × re74, black × re75, u75 × u74, educ × re75, and re74 × re75. The coefficients γ are listed in Table C.1. Conditional on the outcome being positive, its logarithm has a normal distribution with mean $\beta_w' h(x)$,

Table C.1. Data generating process for Lalonde simulations

	γ_0	β_0	γ_1	β_1
Const.	2.7437	1.3486	7.1253	1.7159
age	−0.0331	−0.0054	0.0108	0.0023
educ	0.0022	0.0600	0.1003	0.0472
black	−0.7322	−0.1570	−7.3174	−0.6217
married	0.1582	0.1559	0.6020	−0.0065
u74	−1.5517	0.4833	−16.3253	−0.7556
u75	−0.7797	−0.0372	11.9438	1.5019
re74	−0.0110	0.0271	−1.6848	−0.2240
re75	0.0931	0.0494	8.1122	0.0178
black × u74	1.4475	−0.9685	18.2770	1.1383
black. × u75	−0.1650	−0.2274	−12.7834	−0.7507
black × re74	0.0501	0.0047	1.5954	0.2021
black × re75	0.0392	0.0011	−7.1228	−0.1245
u74 × u75	−1.0170	0.0217	−1.4175	−1.1367
educ × re75	−0.0007	−0.0017	−0.1118	0.0063
re74 × re75	−0.0005	−0.0003	0.0361	0.0051
σ		0.5313		0.9876

for $w = 0, 1$, and variance σ^2, for the same vector of functions of the covariates $h(x)$. Again the values for β_w are given in Table C.1.

ACKNOWLEDGMENTS

The authors thank the financial support for this research, generously provided through NSF grants SES-0350645 (Abadie), SBR-0452590, and SBR-0820361 (Imbens). A previous version of this article circulated under the title "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects" (Abadie and Imbens 2002).

[Received December 2007. Revised August 2009.]

REFERENCES

Abadie, A. (2005), "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19. [1]

Abadie, A., and Imbens, G. (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," Technical Working Paper T0283, NBER. [11]

——— (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74 (1), 235–267. [1-4,7]

Abadie, A., Drukker, D., Herr, H., and Imbens, G. (2003), "Implementing Matching Estimators for Average Treatment Effects in STATA," *The Stata Journal*, 4 (3), 290–311.

Busso, M., DiNardo, J., and McCrary, J. (2009), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," unpublished manuscript, Dept. of Economics, UC Berkeley. [7]

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors," *The Annals of Statistics*, 36 (2), 808–843. [1,3]

Dehejia, R., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [1,2,5]

Frölich, M. (2004), "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86 (1), 77–90. [7]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66 (2), 315–331. [1-3]

Heckman, J., and Hotz, J. (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training" (with discussion), *Journal of the American Statistical Association*, 84, 862–874. [1,5]

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098. [1,3]

Hirano, K., Imbens, G., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [1]

Horvitz, D., and Thompson, D. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685. [1]

Imbens, G. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review Papers and Proceedings*, 93 (2), 126–132. [1,5]

——— (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–30. [2]

Imbens, G. W., and Wooldridge, J. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [1]

Imbens, G., Newey, W., and Ridder, G. (2005), "Mean-Squared-Error Calculations for Average Treatment Effects," unpublished manuscript, Harvard University, Dept. of Economics. [1,3]

Imbens, G., Rubin, D., and Sacerdote, B. (2001), "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence From a Survey of Lottery Players," *American Economic Review*, 91, 778–794. [7]

Lalonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, 76, 604–620. [1,5]

Lechner, M. (2002), "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society, Ser. A*, 165, 59–82. [2]

Newey, W. (1995), "Convergence Rates for Series Estimators," in *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C. R. Rao* (eds. G. S. Maddala, P. C. B. Phillips, and T. N. Srinavasan), Cambridge: Blackwell. [5,8]

Quade, D. (1982), "Nonparametric Analysis of Covariance by Matching," *Biometrics*, 38, 597–611. [1,3]

Robins, J., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [1]

Rosenbaum, P. (1995), *Observational Studies*, New York: Springer-Verlag. [1]

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [2]

Rubin, D. (1973), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203. [1]

——— (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [1]

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328. [3]

Smith, J., and Todd, P. (2005), "Does Matching Address LaLonde's Critique of Nonexperimental Estimators," *Journal of Econometrics*, 125, 305–353. [1,5]

Zhao, Z. (2002), "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics and an Application," unpublished manuscript, Dept. of Economics, Johns Hopkins University. [7]