Simple and Bias–Corrected Matching Estimators for Average Treatment Effects[1]

Alberto Abadie — Harvard University[2]

Guido W. Imbens — UCLA[3], Berkeley, and NBER.

September 2001

[1]We wish to thank Jim Powell, Whitney Newey, Paul Rosenbaum and Ed Vytlacil for comments on an earlier version of this paper, and Don Rubin for many discussions on these topics.

[2]Kennedy School of Government, Harvard University, Cambridge, MA 02138.

[3]Department of Economics, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90095, imbens@econ.ucla.edu.

Simple and Bias–Corrected Matching Estimators for Average Treatment Effects

Alberto Abadie and Guido W. Imbens

Abstract

In this paper we analyze large sample properties of matching estimators, which have found wide applicability in evaluation research despite that fact that their large sample properties have not been established in many cases. We show that standard matching estimators have biases in large samples that do not vanish in the standard asymptotic distribution if the dimension of the covariates is at least four, and in fact dominate the variance if the dimension of the covariates is at least five. In addition, we show that standard matching estimators do not reach the semiparametric efficiency bound, although the efficiency loss is typically small. We then propose a bias-corrected matching estimator that has no asymptotic bias. In simulations the bias-corrected matching estimator performs well compared to simple matching estimators and to regression estimators in terms of bias and root-mean-squared-error.

1. Introduction

Estimation of average treatment effects is an important goal of much evaluation research. Often a reasonable starting point is the assume that assignment to the treatment is unconfounded, that is, based on observable pretreatment variables only, and that there is sufficient overlap in the distributions of the pretreatment variables (Rubin, 1978). Under those assumptions one can estimate the average effect within each subpopulation defined by the pretreatment variables by differencing average treatment and control outcomes. The population average treatment effect can then be estimated by averaging these conditional average treatment effects over the appropriate distribution of the covariates. Methods implementing this in parametric forms have a long history. See for example Cochran and Rubin (1973), Barnow, Cain, and Goldberger (), Rosenbaum and Rubin (1985), Rosenbaum (1995). Recently a number of nonparametric implementations of this idea have been proposed. Hahn (1998) calculates the efficiency bound and proposes an estimator based on nonparametric series estimation of the two conditional regression functions. Heckman, Ichimura and Todd (1997, 1998) focus on the average effect on the treated and consider estimators based on local linear kernel estimation of the two regression functions. Robins and Rotnitzky (1995), and Robins, Rotnitzky and Zhao (1995) propose efficient estimators that combine weighting and regression adjustment. Hirano, Imbens and Ridder (2000) propose an estimator that weights the units by the inverse of their assignment probabilities, and show that nonparametric series estimation of this conditional probability, labeled the propensity score by Rosenbaum and Rubin (1983), leads to an efficient estimator. Ichimura and Linton (2001) consider higher order expansions of such estimators to analyze optimal bandwidth choices.

Here we propose an estimator that matches each treated unit to a fixed number of controls and each control unit to a fixed number of treated units. Various other versions of matching estimators have been proposed in the literature. For example, Rosenbaum (1988,1995), Gu and Rosenbaum (1993) Rubin (1973a,b), and Dehejia and Wahba (1999) focus on the case

where only the treated units are matched, and with typically many controls relative to the number of trainees so that each control is used as a match only once. In contrast, we match all units, treated as well as controls, and explicitly allow a unit to be used more as a match more than once. This modification will typically somewhat increase the variance of the estimator, but will also generally lower the bias. Matching estimators have great intuitive appeal as they do not require the researcher to set any smoothing parameters other than the number of matches which has a clear and interpretable bias/variance tradeoff. However, as we show, somewhat surprisingly given the widespread popularity of matching estimators, the large sample properties of the simple matching estimators with a fixed number of matches are not necessarily very attractive. First, although these estimators are consistent, if the dimension of the continuous pre-treatment variables is larger than three, simple matching estimators have biases that do not vanish in the asymptotic distribution, and that can dominate the large sample variance. This crucial role for the dimension also arises in nonparametric differencing methods (e.g., Yatchew, 1999). In addition, even if the dimension of the covariates is low enough for the bias to vanish asymptotically, the simple matching estimator with a fixed number of matches is not efficient.

We also consider combining matching with additional bias reductions based on a nonparametric extension of the regression adjustment proposed in Rubin (1973b) and Quade (1982). This regression adjustment removes the bias of the estimators, without affecting the variance. Compared to estimators based on regression adjustment without matching (e.g., Hahn, 1998; Heckman, Ichimura and Todd, 1997, 1998) or estimators based on weighting by the inverse of the propensity score, (Hirano, Imbens, and Ridder, 2000) the proposed estimators are more robust as the matching ensures that they do not rely on the (asymptotically) correct specification of the regression function or the propensity score for consistency.

As the bias correction does not affect the variance, the bias corrected matching estimators still do not reach the semiparametric efficiency bound with a fixed number of matches. Only if the number of matches increases with the sample size do the matching estimators

reach the efficiency bound. However, as we shall show, the efficiency loss from using only a small number of matches is very modest, and it can often be bounded. An advantage of the matching methods is that the variance can be estimated conditional on the smoothing parameter (i.e., the number of matches), whereas in the regression estimators often only estimators for the limiting variance are available. We find that using the variance conditional on the covariates and the number of matches leads to more accurate confidence intervals. In simulations we find that the proposed estimators perform well with relatively few matches, and that its large sample distribution is in those cases a good approximation to the finite sample distribution.

In the next section we introduce the notation and define the estimators. In Section 3 we discuss the large sample properties of matching estimators. In section 4 we analyze bias corrections. In Section 5 we apply the estimators to data from an employment training program previously analyzed by Lalonde (1986), Heckman and Hotz (1989), and Dehejia and Wahba (1999). In Section 6 we carry out a small simulation study to investigate the properties of the various estimators.

2. Notation and Basic Ideas

2.1 Notation

We are interested in estimating the average effect of a binary treatment on some outcome. Let $(Y(0), Y(1))$ denote for a typical unit the two potential outcomes given the control treatment and given the active treatment respectively. The variable W, for $W \in \{0, 1\}$ indicates the treatment received. We observe W and the outcome for this treatment,

$$Y = \begin{cases} Y(0) & \text{if } W = 0, \\ Y(1) & \text{if } W = 1, \end{cases}$$

as well as a vector of pretreatment variables X. The estimand of interest is the population average treatment effect

$$\tau = E[Y(1) - Y(0)],$$

or the average effect for the treated

$$\tau_t = E[Y(1) - Y(0)|W = 1].$$

See for some discussion of the various estimands Rubin (1973) and Heckman and Robb (1984).

We assume assignment to treatment is unconfounded (Rosenbaum and Rubin, 1983), or

$$W \perp \Big(Y(0), Y(1)\Big) \Big| X, \tag{1}$$

as well as that the probability of assignment is bounded away from zero and one: for some $c > 0$

$$c < Pr(W = 1|X = x) < 1 - c, \tag{2}$$

for all x in $\mathcal{X}$, the support of X, which is a compact subset of $\Re^k$. The dimension of X will be seen to play an important role in the properties of the matching estimators. We assume that all covariates are have continuous distributions. Discrete covariates can be easily dealt with and their number does not affect the analysis. The combination of the two assumptions is referred to as strong ignorability (Rosenbaum and Rubin, 1983). These assumptions are strong, and in many cases may not be satisfied. In many studies, however, researchers have found it useful to consider estimators based on these or similar assumptions. See, for example, Cochran (1968), Cochran and Rubin (1973), Rubin (1973a,b), Barnow, Cain and Goldberger (197?) Rosenbaum and Rubin (1984), Heckman and Robb (1984), Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), Rosenbaum (1995), Heckman, Ichimura and Todd (1997, 1998), Ashenfelter and Card (1985), Card and Sullivan (1988), Hahn (1998), Dehejia and Wahba (1999), Lechner (1998), Hirano, Imbens and Ridder (2000) and Hotz, Imbens and Mortimer (2000). If the first assumption, unconfoundedness is deemed implausible in a given application, methods allowing for selection on unobservables such as instrumental variables (e.g., Heckman and Robb, 1984; Angrist, Imbens and Rubin, 1996), sensitivity analyses (Rosenbaum and Rubin, 1984), or bounds calculations (Manski, 1990)

may be considered. See for general discussion of such issues Heckman and Robb (1984), Heckman, Lalonde and Smith (2000), and Angrist and Krueger (2000). The importance of the restriction on the probability of assignment has been discussed in Rubin (1977), Heckman, Ichimura and Todd (1997, 1998), and Dehejia and Wahba (1999).

Under strong ignorability the average treatment effect for the subpopulation with pre-treatment variables equal to $X = x$,

$$\tau(x) \equiv E[Y(1) - Y(0)|X = x]$$

can be estimated from the data on (Y, W, X) because

$$E[Y(1) - Y(0)|X = x] = E[Y|W = 1, X = x] - E[Y|W = 0, X = x].$$

To get the average effect of interest we average this conditional treatment effect over the marginal distribution of X:

$$\tau = E[\tau(X)],$$

or over the conditional distribution to get the average effect for the treated:

$$\tau_t = E[\tau(X)|W = 1].$$

Next we introduce some additional notation. Let $\mu(x, w) = E[Y|X = x, W = w]$ be the conditional regression function, and $\sigma^2(x, w) = E[(Y - \mu(x, w)))^2|X = x, W = w]$ the conditional variance function. By unconfoundedness, $E[Y(0)|X = x] = E[Y(0)|X = x, W = 0] = E[Y|X = x, W = 0] = \mu(x, 0)$ and $E[Y(1)|X = x] = E[Y(1)|X = x, W = 1] = E[Y|X = x, W = 1] = \mu(x, 1)$. Let $f_w(x)$ be the conditional density of X given $W = w$, and let $e(x) = Pr(W = 1|X = x)$ be the propensity score (Rosenbaum and Rubin, 1983). The numbers of control and treated units are N_0 and N_1 respectively, with $N = N_0 + N_1$ the total number of units. Also, let $\mathcal{I}_0$ and $\mathcal{I}_1$ be the set of indices for the control and treated units respectively:

$$\mathcal{I}_w = \{i = 1, \ldots, N|W_i = w\},$$

for $w = c, t$. Let $||x|| = (x'x)^{1/2}$, for $x \in \mathcal{X}$ be the standard vector norm. Occasionally we will be using alternative norms of the form $||x||_V = (x'Vx)^{1/2}$ for some positive definite symmetric matrix V. In particular we will use the norm with V equal to the inverse of the diagonal matrix with the variances of X on the diagonal, or with V equal to the inverse of the covariance matrix of X, the Mahalanobis metric. In practice, as we discuss below, one may wish to change the weights for the covariates by adjusting the values on the diagonal of V.

Given a random sample of size N, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $j_m(i)$ be the index j that solves

$$\sum_{l \in \mathcal{I}_{1-W_i}} 1\Big\{||X_l - X_i|| \leq ||X_j - X_i||\Big\} = m,$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $j_m(i)$ is the index of the unit that is the m^{th} closest to unit i in terms of the distance measure based on the norm $||\cdot||$, among the units with the treatment opposite to that of unit i. In particular, $j_1(i)$, sometimes for notational convenience denoted by $j(i)$, is the nearest match for unit i. For simplicity we ignore the possibility of ties. As the large sample properties of matching estimators are straightforward with discrete covariates, focussing on the case with only continuous regressors where the probability of such ties is zero is not restrictive. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches:

$$\mathcal{J}_M(i) = \{j_1(i), \ldots, j_M(i)\}.$$

Finally, let $K_M(i)$ denote the number of times unit i is used as a match given M matches per unit:

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

In many procedures matching is carried out without replacement, so that every unit is used as a match at most once, and $K_M(i) \leq 1$. In our set up, with both treated and control units matched it is imperative that units can be used as matches more than once and the distribution of $K_M(i)$ is important in terms of the variance of the estimators.

2.2 Estimators

All estimators we consider are of the form

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{Y}_i(1) - \hat{Y}_i(0)\Big),$$

where one of the two potential outcomes $Y_i(0)$ and $Y_i(1)$ is observed and the other one is estimated. The estimators differ in the manner in which the unobserved potential outcome is estimated.

The first estimator, the simple matching estimator uses the following estimates for the missing potential outcomes:

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

and

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

The basic matching estimator we shall study is

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

Consider the case with a single match ($M = 1$). Each unit gets matched in this estimator, so that on average each units gets used as a match once. However, some units get used as a match more than once. This procedure differs from standard pairwise matching procedures where one constructs a number of distinct pairs. In our approach it could be that the nearest match for treated unit i is control unit $i' = j(i)$, but that the nearest match for control unit i' is $i'' = j(i')$, not necessarily the original treated unit i. As a result, differences $\hat{Y}_i(1) - \hat{Y}_i(0)$ and $\hat{Y}_{i'}(1) - \hat{Y}_{i'}(0)$ are not necessarily independent, and in fact maybe perfectly correlated if i is matched to i' ($j(i) = i'$) and i' is matched to i ($j(i') = i$). Matching with replacement will lead to a lower variance, although in general will not achieve the efficiency bound, but this comes at the price of a lower match quality, and thus typically a higher bias.

We shall compare the matching estimators to covariance-adjustment or regression estimators where

$$\bar{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \hat{\mu}(X_i, 0) & \text{if } W_i = 1, \end{cases} \tag{4}$$

and

$$\bar{Y}_i(1) = \begin{cases} \hat{\mu}(X_i, 1) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases} \tag{5}$$

with corresponding estimator

$$\hat{\tau}^{reg} = \frac{1}{N} \sum_{i=1}^{N} \left(\bar{Y}_i(1) - \bar{Y}_i(0) \right). \tag{6}$$

The $\hat{\mu}(x, 0)$ and $\hat{\mu}(x, 1)$ are consistent estimators for the population regression functions $\mu(x, 0)$ and $\mu(x, 1)$ respectively. This distinction between regression estimators and matching estimators is somewhat vague. With $\hat{\mu}(x, w)$ a nearest neighbour estimator with a fixed number of neighbours the regression estimator is identical to the matching estimator with the same number of matches. These two estimators will differ in the way they change with the number of observations. We classify as matching estimators those estimators where there is a finite and fixed number of matches, and as regression estimators those where $\hat{\mu}(x, w)$ is a consistent estimator for $\mu(x, w)$. The estimators considered by Hahn (1998) and Heckman, Ichimura and Todd (1997, 1998) are of the regression type. Hahn shows that using series estimators for $\hat{\mu}(x, 1)$ and $\hat{\mu}(x, 0)$ this leads to efficient estimators for the average treatment effect τ.

In addition we develop a bias–corrected matching estimator where the difference within the matches is adjusted:

$$\tilde{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \left(Y_j + \hat{\mu}(X_i, 0) - \hat{\mu}(X_j, 0) \right) & \text{if } W_i = 1, \end{cases} \tag{7}$$

and

$$\tilde{Y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \left(Y_j + \hat{\mu}(X_i, 1) - \hat{\mu}(X_j, 1) \right) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases} \tag{8}$$

with corresponding estimator

$$\hat{\tau}_M^{bcm} = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i(1) - \tilde{Y}_i(0)\right). \tag{9}$$

Rubin (1979) discusses such estimators in the context of matching without replacement and linear covariance adjustment.

To set the stage for some of the discussion below, note that the bias for the matching estimator with a single match consists of terms of the form $\mu(X_i,0) - \mu(X_{j(i)},0)$. The bias correction subtracts corresponding terms of the form $\hat{\mu}(X_i,0) - \hat{\mu}(X_{j(i)},0)$. This term has the obvious effect of removing some of the bias if $\hat{\mu}(x,0)$ is close to $\mu(x,0)$. The bias of the regression estimator on the other hand consists of terms of the form $\hat{\mu}(X_i,0) - E[\mu(X_i,0)]$. The bias corrected matching estimator differs from the regression estimator by terms of the form $Y_{j(i)} - \hat{\mu}(X_{j(i)},0)$, with expectation $\mu_w(X_{j(i)}) - E[\hat{\mu}(X_{j(i)},0)]$. If $E[\hat{\mu}(x,0)]$ is equal to $\mu(x,0)$, adding such terms does not affect the bias and merely introduces noise. However, if there is a substantial difference between $E[\hat{\mu}(x,0)]$ and $\mu(x,0)$, and if in addition X_i is close to $X_{j(i)}$, this term may remove a substantial amount of bias.

In the end, the bias-adjusted matching estimator combines some of the bias reductions from the matching, by comparing units with similar values of the covariates, and the bias-reduction from the regression. Compared to only regression adjustment it relies less on the accuracy of the estimator of the regression function since it only needs to adjust for relatively small differences in the covariates. The remaining bias consists of terms of the form

$$\mu(X_i,0) - \mu(X_{j_m(i)},0) - \left(\hat{\mu}(X_i,0) - \hat{\mu}(X_{j_m(i)},0)\right).$$

As long as either X_i is close to $X_{j_m(i)}$, or $\hat{\mu}(\cdot)$ is close to $\mu(\cdot)$, the remaining bias is small.

We are interested in the properties of the simple and bias corrected matching estimators in large samples, that is, as N increases, for fixed M. The properties of interest include bias and variance. Of particular interest is the dependence of these results on the dimension of the covariates.

3. Simple Matching Estimators

In this section we investigate the properties of the basic matching estimator $\hat{\tau}_M^{match}$ defined in (3). Define $\varepsilon_i = Y_i - \mu(X_i, W_i)$, so that

$$E[\varepsilon|X = x, W = w] = 0,$$

and

$$V(\varepsilon)|X = x, W = w) = \sigma^2(x, w).$$

Now define the two $N \times N$ matrices $\mathbf{A}_0(\mathbf{X}, \mathbf{W})$ and $\mathbf{A}_1(\mathbf{X}, \mathbf{W})$, with typical element

$$\mathbf{A}_{1,ij} = \begin{cases} 1 & \text{if } i = j,\ W_j = 1 \\ 1/M_i & \text{if } j \in \mathcal{J}_M(i),\ W_j = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

and

$$\mathbf{A}_{0,ij} = \begin{cases} 1 & \text{if } i = j,\ W_j = 0 \\ 1/M_i & \text{if } j \in \mathcal{J}_M(i),\ W_j = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

and define $\mathbf{A} = \mathbf{A}_1 - \mathbf{A}_0$. Let $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, ε, $\mathbf{W}$, $\mathbf{Y}$, and $\mathbf{X}$, be the matrices with ith row equal to $Y_i(0)$, $Y_i(1)$, ε_i, W_i, Y_i, X_i respectively, let the vector $\hat{\mathbf{Y}}(1) - \hat{\mathbf{Y}}(0)$ be the N dimenisional vector with ith element $\hat{Y}_i(1) - \hat{Y}_i(0)$, and let ι_N be the N-dimensional vector with all elements equal to one. Then

$$\hat{\mathbf{Y}}(1) = \mathbf{A}_1\mathbf{Y}(1) = \mathbf{A}_1\mathbf{Y},$$

and

$$\hat{\mathbf{Y}}(0) = \mathbf{A}_0\mathbf{Y}(0) = \mathbf{A}_0\mathbf{Y}.$$

We can now write the estimator $\hat{\tau}_M^{match}$ as

$$\hat{\tau}_M^{match} = \iota_N'\Big(\hat{\mathbf{Y}}(1) - \hat{\mathbf{Y}}(0)\Big)Bigl/N = \Big(\iota_N'\mathbf{A}_1\mathbf{Y} - \iota_N'\mathbf{A}_0\mathbf{Y}\Big)\Big/N = \iota_N'\mathbf{A}\mathbf{Y}/N$$

$$= \iota_N'\mathbf{A}\mu(\mathbf{X}, \mathbf{W})/N + \iota_N'\mathbf{A}\varepsilon/N.$$

Using the fact that $\mathbf{A}_1\mu(\mathbf{X},\mathbf{W}) = \mathbf{A}_1\mu(\mathbf{X},1)$ and $\mathbf{A}_0\mu(\mathbf{X},\mathbf{W}) = \mathbf{A}_0\mu(\mathbf{X},0)$ we can write this as

$$\hat{\tau}_M^{match} = \iota_N'\Big(\mu(\mathbf{X},1) - \mu(\mathbf{X},0)\Big)/N + \iota_N'\mathbf{A}\varepsilon/N \tag{12}$$

$$+\iota_N'(\mathbf{A}_1 - I_N)\mu(\mathbf{X},1)/N - \iota_N'(\mathbf{A}_0 - I_N)\mu(\mathbf{X},0)/N$$

If the matching is exact, and $X_i = X_{j_m(i)}$, then the last two terms are equal to zero. The expectation of the first two terms is equal to τ. Conditional on $\mathbf{X}$ and $\mathbf{W}$ only the second term is stochastic, and this term is of relevance for the variance of the estimator. We will analyze this term in Section 3.2 The last two terms constitute the bias, and will be analyzed in Section 3.1.

3.1 Bias

The bias of the simple matching estimator is equal to the expectation of the last two terms in (12), $E[\iota_N'(\mathbf{A}_1 - I_N)\mu(\mathbf{X},1)/N - \iota_N'(\mathbf{A}_0 - I_N)\mu(\mathbf{X},0)/N]$. To investigate this term further, consider the i^{th} element of $(\mathbf{A}_1 - I_N)\mu(\mathbf{X},1) - (\mathbf{A}_0 - I_N)\mu(\mathbf{X},0)$. Suppose $W_i = 1$. Then the i^{th} element is equal to

$$\mu(X_i,0) - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\mu(X_j,0) = \frac{1}{M}\sum_{j=1}^{M}\Big(\mu(X_i,0) - \mu(X_{j_m(i)},0)\Big).$$

Thus the components of the bias consist of the difference between expected value of $Y_i(0)$ given X_i, and the average of the expected values of the matches. To investigate the nature of this bias we expand the difference $\mu(X_i,0) - \mu(X_{j_m(i)},0)$ around X_i:

$$\mu(X_i,0) - \mu(X_{j_m(i)},0) = \frac{\partial\mu}{\partial x}(X_i,0)'(X_i - X_{j_m(i)})$$

$$+\frac{1}{2}(X_i - X_{j_m(i)})'\frac{\partial^2\mu}{\partial x\partial x'}(X_i,0)(X_i - X_{j_m(i)}) + O(||(X_i - X_{j_m(i)})||^3).$$

In order to study the components of the bias it is useful to consider the distribution of the matching discrepancy, the difference between the value of the covariate X_i and the value of the covariate for its mth nearest match, $X_{j_m(i)}$. We denote this difference by $D_{m,i} =$

$D_{m,i}(\mathbf{X},\mathbf{W}) = X_i - X_{j_m(i)}$. We focus on this discrepancy for the treated observations with $W_i = 1$. The argument for the control observations is analogous.

Lemma 1 (Distribution of Matching Discrepancy)
Conditional on $X_i = x$ and $W_i = 1$, and N_0,
(i), $D_{m,i} = O_p(N_0^{-1/k})$,
(ii)

$$E[D_{m,i}] = \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f_0(x)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{f_0(x)}\frac{\partial f_0}{\partial x}(x)\frac{1}{N_0^{2/k}} + o\left(\frac{1}{N_0^{2/k}}\right),$$

(iii),

$$E[D_{m,i}D_{m,i}'] = \Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\left(f_0(x)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{N_0^{2/k}}\cdot I_N + o\left(\frac{1}{N_0^{2/k}}\right),$$

(iv), $E\|D_{m,i}\|^3 = o\left(N_0^{-2/k}\right)$.

Note that although the stochastic order of the discrepancy is $O_p(N_0^{-1/k})$, the expectation is of lower order, namely $O(N_0^{-2/k})$. The reason is that the leading stochastic term has a symmetric distribution with mean zero. The leading term of the bias therefore depends on the expected value of the next, $O(N_0^{-2/k})$, term in the stochastic expansion and the variance of the leading, $O(N^{-1/k})$, term.

Now let us consider the bias for unit i conditional on X_i and $\mathbf{W}$:

$$\begin{aligned} B_i &= \left(\hat{Y}_i(1) - \hat{Y}_i(0)\right) - \left(Y_i(1) - Y_i(0)\right) \\ &= W_i\cdot(Y_i(0) - \hat{Y}_i(0)) - (1-W_i)\cdot(Y_i(1) - \hat{Y}_i(1)). \end{aligned}$$

Lemma 2 (unit-level bias)
If $W_i = 1$, then

$$E[B_i|\mathbf{W}, X_i = x] = \Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{f(z)}\frac{1}{N^{2/k}}\frac{\partial f}{\partial x}(x)'\frac{\partial \mu(0)}{\partial x}(x)$$

$$+\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{N^{2/k}}\cdot\text{trace}\left(\frac{\partial^2\mu(0)}{\partial x\partial x'}(x)\right)+o_p\left(N^{-2/r}\right),$$

and if $W_i=0$*, then*

$$E[B_i|\mathbf{W},X_i=x]=\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(1)(x)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{f(1)(x)}\frac{1}{N^{2/k}}\frac{\partial f(1)}{\partial x}(x)'\frac{\partial\mu(1)}{\partial x}(x)$$

$$+\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{N^{2/k}}\cdot\text{trace}\left(\frac{\partial^2\mu(0)}{\partial x\partial x'}(x)\right)+o_p\left(N^{-2/r}\right).$$

To get the overall bias we take the expectation over X_i conditional on $\mathbf{W}$, and then integrate over the distribution of $\mathbf{W}$.

Lemma 3 (BIAS)

The bias of the simple matching estimator is

$$\text{Bias}=p\cdot\int_x\left(\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{f(z)}\frac{1}{N^{2/k}}\frac{\partial f}{\partial x}(x)'\frac{\partial\mu(0)}{\partial x}(x)\right.$$

$$\left.+\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{N^{2/k}}\cdot\text{trace}\left(\frac{\partial^2\mu(0)}{\partial x\partial x'}(x)\right)\right)f(1)(x)dx$$

$$+(1-p)\cdot\int_x\left(\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(1)(x)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{f(1)(x)}\frac{1}{N^{2/k}}\frac{\partial f(1)}{\partial x}(x)'\frac{\partial\mu(1)}{\partial x}(x)\right.$$

$$\left.+\Gamma\left(\frac{k+2}{k}\right)\frac{1}{k}\left(f(z)\frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)}\right)^{-2/k}\frac{1}{N^{2/k}}\cdot\text{trace}\left(\frac{\partial^2\mu(0)}{\partial x\partial x'}(x)\right)\right)f(0)(x)dx+o_p\left(N^{-2/r}\right).$$

3.2 VARIANCE

In this section we investigate the variance of the simple matching estimator $\hat{\tau}_M^{match}$. Because the first term in expression (12) is deterministic conditional on $\mathbf{X}$ and $\mathbf{W}$, the conditional variance of the estimator is equal to the conditional variance of the second term, $\iota_N'\mathbf{A}\varepsilon/N$. Conditional on $\mathbf{X}$ and $\mathbf{W}$, the variance of $\hat{\tau}$ is

$$\text{Var}(\hat{\tau}_M^{match}|\mathbf{X},\mathbf{W})=\text{Var}(\iota_N'\mathbf{A}\varepsilon/N|\mathbf{X},\mathbf{W})=\frac{1}{N^2}\iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N,\tag{13}$$

where

$$\Omega = E\Big[\varepsilon\varepsilon'|\mathbf{X},\mathbf{W}\Big].$$

The covariance matrix Ω is a diagonal matrix with ith diagonal element equal to the conditional variance of Y_i given X_i and W_i:

$$\omega_{ii} = \sigma^2(X_i, W_i).$$

Note that (13) gives the exact variance, not relying on large sample approximations.

Estimating the variance, or its limiting normalized version $V_\tau = \text{plim}\iota_N'\mathbf{A}\Omega\mathbf{A}'\iota_N/N$, is difficult because it involves the conditional variances $\sigma^2(x,w)$. In principle one can estimate these consistently, first using nonparametric regression to obtain $\mu(x,w)$, and then using nonparametric regression again to obtain $\sigma^2(x,w)$. Although this leads to a consistent estimator for V_τ, it would require exactly the type of nonparametric regression that the simple matching estimator allows one to avoid. Often, it may therefore be attractive to estimate the variance under additional assumptions that would allow one to avoid such nonparametric regression.

If one is willing to assume homoskedasticity, the variance simplifies considerably. Under homoskedasticity with $\sigma^2(x,1) = \sigma^2(x,0) = \sigma^2$, the variance reduces to

$$N \cdot \text{Var}(\hat{\tau}_M^{match}|\mathbf{X},\mathbf{W}) = \sigma^2 \cdot \iota_N'\mathbf{A}\mathbf{A}'\iota_N/N.$$

We can further express the variance in this special case in terms of the number of times each observation is matched, using the following lemma.

Lemma 4 *Let* $\mathbf{A}$ *be the matrix defined in (10). Then*

$$\iota_N'\mathbf{A}\mathbf{A}'\iota_N/N = 3 + \frac{1}{M^2}\sum_{i=1}^{N} K_M(i)^2/N.$$

The other component of the variance, σ^2, can also be estimated in a straightforward manner if we assume a constant treatment effect. Define the N-dimensional vector of residuals

$$\nu = \mathbf{A}\varepsilon.$$

The expected value of ν_i is zero, and the variance is

$$V(\nu_i) = \sigma^2 \cdot (1 + 1/M).$$

Hence, we can estimate σ^2 using an estimated residual as

$$\hat{\sigma}^2 = \frac{M}{N(M+1)}(\mathbf{AY} - \hat{\tau} \cdot \iota_N)'(\mathbf{AY} - \hat{\tau} \cdot \iota_N).$$

3.3 Efficiency

In practice there is no reason to go beyond the expression for the variance given in (13) which conditions on the covariates and the treatment indicators, and thus on the number of times each observation is used as a match. In fact, one would expect the conditional variance, with the conditioning on ancillary statistics, to lead to more accurate confidence intervals than the unconditional variance. For comparison purposes with other estimators for whom only an unconditional variance formula is available, however, it is of interest to calculate the approximate large sample variance, based on the expected value of the conditional variance. This also allows us to investigate the efficiency of the matching estimators. In general the key to the efficiency properties of the matching estimators is the distribution of $K_M(i)$, the number of times each unit is used as a match. It is difficult to work out the approximate distribution of this number for the general case. Here we investigate the form of the variance for the special case with homoskedastic residuals and a scalar covariate.

Lemma 5 *Suppose $\sigma^2(x,w) = \sigma^2$, and $k = dim(X) = 1$. Then, for fixed M the normalized variance V_τ converges to*

$$V_\tau = \text{plim } \sigma^2 \iota' \mathbf{A}\mathbf{A}' \iota / N = V_{\text{eff}} \cdot \left(1 + \frac{1}{2M}\right) - \frac{\sigma^2}{2M},$$

where V_{eff} is the semiparametric efficiency bound:

$$V_{\text{eff}} = \sigma^2 \cdot E\left[\frac{1}{e(X)} + \frac{1}{1-e(X)}\right].$$

The estimator is not efficient in general. However, the efficiency loss disappears if one increases the number of matches. In practice the efficiency loss from using two or three matches is very small. For example, the variance with a single match is at most 50% higher than the variance of the efficient esetimator, and with three matches the variance is at most 16% higher. The key to the efficiency loss is the variance of $K_M(i)/M$. As the number of matches increases, this variance converges to zero and the estimator becomes efficient.

4. Bias Corrected Matching

In this section we analyze the properties of the bias corrected matching estimator $\hat{\tau}_M^{bcm}$. First we introduce an infeasible version of the bias corrected estimator:

$$\underline{Y}_{c,i} = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} \left(Y_j + \mu(X_i,0) - \mu(X_j,0)\right) & \text{if } W_i = 1, \end{cases} \tag{14}$$

and

$$\underline{Y}_{t,i} = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} \left(Y_j + \mu(X_i,1) - \mu(X_j,1)\right) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases} \tag{15}$$

with corresponding estimator

$$\hat{\tau}_M^{i-bcm} = \frac{1}{N}\sum_{i=1}^{N} \left(\underline{Y}_{t,i} - \underline{Y}_{c,i}\right). \tag{16}$$

The bias correction in the infeasible bias-corrected matching estimator uses the actual, rather than the estimated regression functions. Using these it is obvious that the correction removes all the bias. The correction term solely depends on the covariates and the treatment indicators, so that conditionally it has the same variance as the simple matching estimator. It is therefore a clear improvement over the simple matching estimator in terms of mean-squared-error:

Theorem 1 (Infeasible Bias Corrected Matching Estimator versus Simple Matching Estimator)
(i) $E[\hat{\tau}_M^{i-bcm}] = \tau$,
(ii) $V(\hat{\tau}_M^{i-bcm}|\mathbf{X},\mathbf{W}) = V(\hat{\tau}_M^{match}|\mathbf{X},\mathbf{W})$.

Proof: This follows immediately from the discussion. The conditional variance is not affected by the addition of the bias correction terms that depend only on $\mathbf{X}$ and $\mathbf{W}$, and the infeasible estimator has zero bias so the bias of the simple matching estimator cannot be lower. □

The second key result concerns the difference between the infeasible and feasible bias-corrected matching estimator. Given sufficient smoothness the difference between the infeasible and feasible bias-corrected matching estimators is of sufficiently low order that it can be ignored in large sample approximations.

Theorem 2 (Infeasible versus Feasible Bias Corrected Matching Estimator) *Suppose Assumptions ?-? hold. Then*

$$\sqrt{N} \cdot \left(\hat{\tau}_M^{i-bcm} - \hat{\tau}_M^{bcm}\right) \xrightarrow{d} 0.$$

Proof: See Appendix.

5. Estimates using Lalonde Data

In this section we apply the estimators to a subset of the data analyzed by Lalonde (1986) and Dehejia and Wahba (1999). We use data from a randomized evaluation of a job training program and a subsample from the Panel Study of Income Dynamics. Using the experimental data we obtain an unbiased estimate of the average effect of the training. We then see how well the non-experimental matching estimates compare using the experimental trainees and the controls from the PSID.

Table 1 presents summary statistics for the three groups. The first two columns present the summarys statistics for the experimental trainees. The second pair of columns presents

the results for the experimental controls. The third pair of columns presents summary statistics for the non-experimental control group constructed from the PSID. The last two columns present t-statistics for the hypothesis that the population averages for the trainees and the experimental controls, and for the trainees and the PSID controls, respectively, are zero. Note the large differences in background characteristics between the trainees and the PSID sample. This is what makes drawing causal inferences from comparisons between the PSID sample and the trainee group a tenuous task. From the last two rows we can obtain an unbiased estimate of the effect of the training on earnings in 1978 by comparing the averages for the trainees and the experimental controls, $6.35 - 4.55 = 1.80$ (s.e. 0.63).

Table 2 presents estimates of the causal effect of training on earnings using various matching and regression adjustment estimators. The top part of the table reports estimates for the experimental data (experimental trainees and experimental controls), and the bottom part reports estimates based on the experimental trainees and the PSID controls. The first set of rows in each case reports matching estimates, based on a number of matches including 1, 4, 16, 64 and 2490. The matching estimates include simple matching, and regression adjusted matching estimates where the regression can be based on all observations or only on the matched observations. The second part reports estimates based on linear regression with no controls, all covariates linearly and all covariates with quadratic terms and a full set of interactions. The experimental estimates range from 1.17 (regression using the matched observations, with a single match) to 2.27 (quadratic regression). The non-experimental estimates have a much wider range, from -15.20 (simple difference) to 3.26 (quadratic regression). Using a single match, however, there is little variation in the estimates, ranging only from 2.09 to 2.56. The regression adjusted matching estimator, with the regression based only on the matched observations, does not vary much with the number of matches, with estimates of 2.45 ($M = 1$), 2.51 ($M = 4$), 2.48 ($M = 16$), and 2.26 ($M = 16$), and only with $M = 2490$ does the estimate deteriorate to 0.84. The matching estimates all use the identity matrix as the weight matrix, after normalizing the covariates to have zero mean and unit

variance.

To see how well the matching peforms 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 for the PSID controls and the experimental trainees. One can see that before matching, the averages for some of the variables are more than a 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 indicicator variables the matching is exact: every trainee is matched to someone with the same ethnicity, marital status and employment history for the years 1974 and 1975. 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 trainees and controls 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 goes down, with even the indicator variables no longer matched exactly, but in most cases the average difference is still far smaller than the standard deviation till we get to 16 or more matches. The last row reports matching differences for logistic estimates of the propensity score. Although the matching is not directly on the propensity score, with single matches the average difference in the propensity score is only 0.21, whereas without matching the difference between trainees and controls is 8.16, 40 times higher.

6. Simulations

In this section we discuss some simulations designed to assess the performance of the various matching estimators in this context. To make the evaluation more credible we simulate data sets that in a number of ways resemble the Lalonde data set analyzed in the previous section fairly closely.

In the simulation we have nine regressors, designed to match the following variables in the

Lalonde data set: age, education, black, hispanic, married, earnings1974, unemployed1974, earnings1975, unemployed1975. For each simulated data set we sample with replacement 185 observations from the empirical covariate distribution of the trainees, and 2490 observations from the empirical covariate distribution of the PSID controls. This gives us the joint distribution of covariates and treatment indicators. For the conditional distribution of the outcome given covariates, we estimated a two-part model on the PSID controls, where the probability of zero earnings is a logistic function of the covariates with a full set of quadratic terms and interactions. Conditional on being positive, the log of earnings is a function of the covariates with again a full set of quadratic terms and interactions. We then assume a constant treatment effect of 2.0.

For each data set simulated in this way we report results for the same set of estimators. For each estimator we report the mean and median bias, the root-mean-squared-error, the median-absolute-error, the standard deviation, the average estimated standard error, and the coverage rates for nominal 95% and 90% confidence intervals. The results are reported in Table 4.

In terms of rmse and mae the matching estimator using regression adjustment with the regression only using the matched observations is best with 4 or 16 matches, although the matching estimator with regression adjustment based on all observations is close. However, the matched regression estimator is far better in terms of bias. The simple matching estimator does not perform as well neither in terms of bias or rmse. The pure regression adjustment estimators do not perform very well. They have high rmse and substantial bias.

In terms of coverage rates the matching estimators have higher than nominal coverage rates, which is somewhat surprising. The regression estimators have lower than nominal coverage rates.

7. Conclusions

In this paper we derive large sample properties of simple matching estimators. We point

out that with more than three continuous covariates the bias will dominate the variance in large samples. We suggest a bias-adjustment that removes the asymptotic bias. The resulting estimator is simple to implement and appears to perform well in simulations.

References

Ashenfelter, O., and D. Card, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648-660.

Barnow, J. Cain, and A. Goldberger

Card, D., and Sullivan, (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica*, vol. 56, no. 3 497–530.

Cochran, W., (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies", *Biometrics* 24, 295-314.

Cochran, W., and D. Rubin (1973) "Controlling Bias in Observational Studies: A Review" *Sankhya*, 35, 417-46.

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

Friedlander, D., and P. Robins, (1995), "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods", *American Economic Review*, Vol. 85, p 923–937.

Gradshteyn, I.S. and I.M. Ryzhik, (2000), *Table of Integrals, Series, and Products*. 6th ed. New York: Academic Press.

Gu, X., and P. Rosenbaum, (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms", *Journal of Computational and Graphical Statistics*, 2, 405-20.

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

Heckman, J., and J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association*.

Heckman, J., and R. Robb, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*,

Cambridge, Cambridge University Press.

Heckman, J., H. Ichimura, and P. Todd, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies* 64, 605-654.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching As An Econometric Evaluations Estimator," *Review of Economic Studies* 65, 261-294.

Hirano, K., G. Imbens, and G. Ridder, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score" NBER Working Paper.

Ichimura, H., and O. Linton, (2001), "Trick or Treat: Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." unpublished manuscript, London School of Economics.

Lechner, M, (1998), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics*.

Ming, K., and P. Rosenbaum, (2000), "Substantial Gain in Bias Reduction from Matching with a Variable Number of Controls", *Biometrics*.

Moller, J., (1994), *Lectures on Random Voronoi Tessellations*, Springer Verlag, New York.

Okabe, A., B. Boots, K. Sugihara, and S. Nok Chiu, (2000), *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd Edition, Wiley, New York.

Olver, F.W.J., (1974), *Asymptotics and Special Functions*. Academic Press, New York.

Quade, D., (1982), "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38:, 597-611.

Robins, J., and A. Rotnitzky, (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, Vol. 90, No. 429, 122-129.

Rosenbaum, P., (1988), "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 79, 565-74.

Rosenbaum, P., (1988), "Optimal Matching in Observational Studies", *Journal of the Amer-*

ican Statistical Association, 84, 1024-32.

ROSENBAUM, P., (1995), *Observational Studies*, Springer Verlag, New York.

ROSENBAUM, P., (2000), "Covariance Adjustment in Randomized Experiments and Observational Studies," forthcoming, *Statistical Science*.

ROSENBAUM, P., AND D. RUBIN, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

ROSENBAUM, P., AND D. RUBIN, (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", *American Statistician*, 39, 33-38.

ROTNITZKY, A., AND J. ROBINS, (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring," *Biometrika*, Vol. 82, No. 4, 805-820.

RUBIN, D., (1973a), "Matching to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.

RUBIN, D., (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.

RUBIN, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1-26.

RUBIN, D. B., (1978a), "Bayesian inference for causal effects: The Role of Randomization", *Annals of Statistics*, 6:34–58.

RUBIN, D., (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

RUBIN, D., AND N. THOMAS, (1992), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20 (2) 1079-1093.

RUBIN, D., AND N. THOMAS, (1992), "Characterizing the Effect of Matching using Linear Propensity Score Methods with Normal Distributions," *Biometrika* 79, 797-809.

RUBIN, D., AND N. THOMAS, (1996), "Matching Using Estimated Propensity Scores: Relat-

ing Theory to Practice," *Biometrics* 52, 249-264.

Shadish, W., T. Cook and D. Campbell, (2001), *Experimental and Quasi-Experimental Designs*, Houghton Mifflin, Boston, MA.

Stoyan, D., W. Kendall, and J. Mecke, (1995), *Stochastic Geometry and its Applications*, 2nd Edition, Wiley, New York.

Stroock, D.W., (1994), *A Concise Introduction to the Theory of Integration.* Birkhäuser, Boston.

Yatchew, A., (1999), "Differencing Methods in Nonparametric Regression: Simple Techniques for the Applied Econometrician", Working Paper, Department of Economics, University of Toronto.

Appendix

Assumption 1 (Random Sampling)

$\{Y_i, X_i, W_i)\}_{i=1}^N$ are independent and identically distributed.

Assumption 2 (Covariate Support)

The support of X is $\mathcal{X} \subset \Re^r$, the Cartesian product of compact intervals, with absolutely continous density $f_X(x)$ that is bounded and bounded away from zero on $\mathcal{X}$.

Assumption 3 (Finite Moments)

The first two moments of Y_w and $Y(0)$ given $X = x$ are finite for all $x \in \mathcal{X}$.

Assumption 4 (Propensity Score)

The propensity score $e(x) = Pr(T = 1|X = x)$ is bounded away from zero and one.

Assumption 5 (Unconfoundedness)

$$W \perp \Big(Y(1), Y(0)\Big) \Big| X.$$

Assumption 6 (Rates of Power Series)

$$K^4/N \to 0.$$

Assumption 7 (Derivatives of Expectations) *There is a C such that for each multi-index λ the λth partial derivative of $\mu_w(x)$ exist for $w = c, t$ and are bounded by $C^{|\lambda|}$.*

First we state some additional lemmas.

Theorem 3 (Uniform Convergence of Series Estimators of Derivatives of Regression Functions, Newey 1993)

Suppose Assumptions ?-? hold. Then for any alpha > 0 and non-negative integer d,

$$|\hat{\mu}_w(\cdot) - \mu_w(\cdot)|_d = O_p\left(K^{1+2r}\left((K/N)^{1/2} + K^{-\alpha}\right)\right),$$

for $w = c, t$.

Proof: Assumptions 3.1, 4.1, 4.2 and 4.3 in Newey (1993) are satisfied, implying that Newey's Theorem 4.4 applies.

Lemma 6 (Distribution of Matching Discrepancy)
Suppose Assumptions ??-?? hold. Then

$$D_m(X_i) = O_p(N^{-1/r})$$

Lemma 7 (Infeasible versus Feasible Bias Correction)
Suppose Assumptions ??-?? hold. Then

$$\hat{\mu}_w(X_i) - \hat{\mu}_w(X_{j_m(i)}) - \Big(\mu_w(X_i) - \mu_w(X_{j_m(i)})\Big) = o_p(N^{-1/2}),$$

for $w = c, t$.

Proof

Fix the non-negative integer $L > (r-2)/2$. Let λ be a multi-index of dimension r, that is, an r-dimensional vector of non-negative integers, with $|\lambda| = \sum_{i=1}^r \lambda_i$, and let Λ_l be the set of λ such that $|\lambda| = l$. Furthermore, let $x^\lambda = x_1^{\lambda_1} \dots x_r^{\lambda_r}$, and let $\partial^\lambda g(x) = \partial^{|\lambda|} g(x)/\partial x_1^{\lambda_1} \dots \partial x_r^{\lambda_r}$. Finally, let $D_m(X_i) = X_{j_m(i)} - X_i$. Use a Taylor series expansion around X_i to write

$$\hat{\mu}_w(X_{j_m(i)}) = \hat{\mu}_w(X_i) + \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) D_m(X_i)^\lambda + \sum_{\lambda \in \Lambda_{L+1}} \partial^\lambda \hat{\mu}_w(\tilde{x}) D_m(X_i)^\lambda.$$

First consider the last sum, $\sum_{\lambda \in \Lambda_{L+1}} \partial^\lambda \hat{\mu}_w(\tilde{x}) D(X_i)^\lambda$. By Assumption 7, the first factor in each term is bounded by $C^{|\lambda|} = C^{L+1}$. The second factor in each term is of the form $\prod_{j=1}^r D_m(X_i)_j^{\lambda_j}$. The factor $D_m(X_i)_j^{\lambda_j}$ is of order $O_p(N^{-\lambda_j/r})$, so that the product is of the order $O_p(N^{-\sum_{j=1}^r \lambda_j/r}) = O_p(N^{-(L+1)/r}) = o_p(N^{-1/2})$ because $L > (r-2)/2$. Hence, we can write

$$\hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) = \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \partial^\lambda \hat{\mu}_w(X_i) D_m(X_i)^\lambda + o_p(N^{-1/2}).$$

Using the same argument as we used for the estimated regression function $\hat{\mu}_w(x)$ we have for the true regression function $\mu_w(x)$,

$$\mu_w(X_{j_m(i)}) - \mu_w(X_i) = \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \partial^\lambda \mu_w(X_i) D_m(X_i)^\lambda + o_p(N^{-1/2}).$$

Now consider the difference between these two expressions:

$$\hat{\mu}_w(X_{j_m(i)}) - \hat{\mu}_w(X_i) - \Big(\mu_w(X_{j_m(i)}) - \mu_w(X_i)\Big)$$

$$= \sum_{l=1}^{L} \sum_{\lambda \in \Lambda_l} \Big(\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)\Big) \cdot D_m(X_i)^\lambda + o_p(N^{-1/2}).$$

Consider for a particular $\lambda \in \Lambda_l$ the term $\Big(\partial^\lambda \hat{\mu}_w(X_i) - \partial^\lambda \mu_w(X_i)\Big) \cdot D_m(X_i)^\lambda$. The second factor is, using the same argument as before, of order $O_p(N^{-l/r})$. Since $l \geq 1$, the second factor is at most $O_p(N^{-1/r})$. Now consider the first factor. By Theorem 3, this is of order $O_p\left(K^{1+2r}\left((K/N)^{1/2} + K^{-\alpha}\right)\right)$. With $K = N^\delta$, this is $O_p\left(N^{\delta(3/2+2r)-1/2} + N^{-\alpha\delta(1+2r)}\right)$. We can choose α large enough so that for any given δ the first term dominates. Hence the order of the product is $O_p\left(N^{\delta(3/2+2r)-1/2}\right) \cdot O_p(N^{-1/r})$ Given that by Assumption ?? $\delta < 2/(3r+4r^2)$ we have $\delta(3/2+2r) - 1/2 < 1/r - 1/2$, and therefore the order is $o_p(N^{-1/2})$. □

Proof of Theorem 2

The difference $\hat{\tau}^{-bcm} - \hat{\tau}^{bcm}$ can be written as

$$\hat{\tau}^{-bcm} - \hat{\tau}^{bcm} = \frac{1}{N}\left(\sum_{i|w_i=c} \hat{\mu}(X_i,1) - \hat{\mu}(1)(X_{j_m(i)}) - \Big(\mu(X_i,1) - \mu(1)(X_{j_m(i)})\Big)\right.$$

$$\left. + \sum_{i|w_i=t} \hat{\mu}(X_i,0) - \hat{\mu}(0)(X_{j_m(i)}) - \Big(\mu(X_i,0) - \mu(0)(X_{j_m(i)})\Big)\right).$$

Each of these terms are of order $o_p(N^{-1/2})$. There are N of them, with a factor $1/N$. Hence the sum is of order $o_p(N^{-1/2})$.

Proof of Lemma 1

$$\Pr(k_{m,N} = 1 | X_1 = x)$$

$$= \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) (\Pr(\|X-z\| > \|x-z\|))^{N-m} (\Pr(\|X-z\| \leq \|x-z\|))^{m-1}$$

$$= \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) (1 - \Pr(\|X-z\| < \|x-z\|))^{N-m} (\Pr(\|X-z\| \leq \|x-z\|))^{m-1}.$$

Hence,

$$f_{D_{m,N}}(v) = N \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) f(z+v) (1 - \Pr(\|X-z\| \leq \|x-z\|))^{N-m}$$

$$\times (\Pr(\|X-z\| \leq \|x-z\|))^{m-1}.$$

Therefore,

$$ED_{m,N} = N \left(\begin{array}{c} N-1 \\ m-1 \end{array} \right) I_{m,N},$$

where

$$I_{m,N} = \int_{\Re^k} v\, f(z+v) (1 - \Pr(\|X-z\| \leq \|x-z\|))^{N-m} (\Pr(\|X-z\| \leq \|x-z\|))^{m-1}\, dv.$$

Change variables to polar coordinates to get:

$$I_{m,N} = \int_0^\infty r^{k-1} \left(\int_{S_k} r\omega\, f(z+r\omega) \lambda_{S_k}(d\omega) \right) (1 - \Pr(\|X-z\| \leq r))^{N-m} (\Pr(\|X-z\| \leq r))^{m-1}\, dr$$

$$= \int_0^\infty r^{k-1} \left(\int_{S_k} r\omega\, f(z+r\omega) \lambda_{S_k}(d\omega) \right) \left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{N-m}$$

$$\times \left(\int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{m-1} dr$$

$$= \int_0^\infty e^{-Np(r)}\, a(r)\, dr,$$

where

$$p(r) = -\log \left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z+s\omega) \lambda_{S_k}(d\omega) \right) ds \right),$$

and

$$a(r) = r^k \cdot \left(\int_{S_k} \omega \, f(z + r\omega) \lambda_{S_k}(d\omega) \right) \frac{\left(\int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{m-1}}{\left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^m}.$$

That is, $a(r) = q(r)b(r)$, $q(r) = r^k c(r)$, and $b(r) = (g(r))^{m-1}$, where

$$c(r) = \frac{\int_{S_k} \omega \, f(z + r\omega) \lambda_{S_k}(d\omega)}{1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds},$$

$$g(r) = \frac{\int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds}{1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds}.$$

Note that

$$\frac{dg}{dr}(r) = r^{k-1} \frac{\int_{S_k} f(z + r\omega) \lambda_{S_k}(d\omega) ds}{\left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^2}.$$

Therefore, the first non-zero derivative of g at zero is

$$\frac{d^k g}{dr^k}(0) = (k-1)! f(z) \int_{S_k} \lambda_{S_k}(d\omega).$$

Using standard results on higher order derivatives of composite functions (see, e.g., Gradshteyn and Ryzhik 2000), it can be seen that the first non-zero derivative of b at zero is:

$$\frac{d^{(m-1)k} b}{dr^{(m-1)k}}(0) = \frac{((m-1)k)!}{(k!)^{m-1}} \cdot \left(\frac{d^k g}{dr^k}(z) \right)^{m-1}$$

$$= \frac{((m-1)k)!}{k^{m-1}} \cdot \left(f(z) \int_{S_k} \lambda_{S_k}(d\omega) \right)^{m-1}.$$

It has been shown above that the first non-zero derivative of $q(r)$ at zero is

$$\frac{d^{k+1} q}{dr^{k+1}}(0) = (k+1)! \left(\int_{S_k} \omega \omega' \, \lambda_{S_k}(d\omega) \right) \frac{df}{dx}(z)$$

$$= (k+1)! \frac{\int_{S_k} \lambda_{S_k}(d\omega)}{k} \frac{df}{dx}(z).$$

Therefore, the first non-zero derivative of $a(r)$ at zero is

$$\frac{d^{mk+1}a}{dr^{mk+1}}(0) = \begin{pmatrix} mk+1 \\ k+1 \end{pmatrix} \frac{d^{(m-1)k}b}{dr^{(m-1)k}}(0) \frac{d^{k+1}q}{dr^{k+1}}(0) = \frac{(mk+1)!}{k^m} \left(f(z) \int_{S_k} \lambda_{S_k}(d\omega) \right)^m \frac{1}{f(z)} \frac{df}{dx}(z).$$

So, in a neigborhood of zero we have that

$$a(r) = \sum_{t=0}^{\infty} a(1) r^{mk+1+t}, \quad \text{with} \quad a(1) = \frac{1}{(mk+1+t)!} \frac{d^{mk+1+t}a}{dr^{mk+1+t}}(0).$$

It has been shown above that $p(0) = 0$ and the first non-zero derivative of $p(r)$ at zero is

$$\frac{d^k p}{dr^k}(0) = (k-1)! f(z) \int_{S_k} \lambda_{S_k}(d\omega).$$

In a neigborhood of zero we have that:

$$p(r) = \sum_{t=0}^{\infty} p(1) r^{k+t}, \quad \text{with} \quad p(1) = \frac{1}{(k+t)!} \frac{d^{k+t}p}{dr^{k+t}}(0).$$

Applying Theorem 8.1 in Olver (1974), we get

$$I_{m,N} = \Gamma\left(\frac{mk+2}{k}\right) \frac{a_0}{k p_0^{(mk+2)/k}} \frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right)$$

$$= \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{k} \left(\frac{f(z)}{k} \int_{S_k} \lambda_{S_k}(d\omega) \right)^{-2/k} \frac{1}{f(z)} \frac{df}{dx}(z) \frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right).$$

Now, since

$$\frac{N^m/(m-1)!}{N \begin{pmatrix} N-1 \\ m-1 \end{pmatrix}} - 1 = o(1),$$

we have that

$$ED_{m,N} = \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{(m-1)!k} \left(f(z) \frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)} \right)^{-2/k} \frac{1}{f(z)} \frac{df}{dx}(z) \frac{1}{N^{2/k}} + o\left(\frac{1}{N^{2/k}}\right).$$

To get the result for $ED_{m,N}D'_{m,N}$, notice that

$$ED_{m,N}D'_{m,N} = N \begin{pmatrix} N-1 \\ m-1 \end{pmatrix} I_{m,N},$$

where

$$\begin{aligned} I_{m,N} &= \int_{\Re^k} vv' \, f(z+v) \left(1 - \Pr(\|X - z\| \leq \|x - z\|)\right)^{N-m} \left(\Pr(\|X - z\| \leq \|x - z\|)\right)^{m-1} dv \\ &= \int_0^\infty r^{k-1} \left(\int_{S_k} r^2 \omega\omega' \, f(z + r\omega) \lambda_{S_k}(d\omega) \right) \left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{N-m} \\ &\qquad \times \left(\int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{m-1} dr \\ &= \int_0^\infty e^{-Np(r)} \, a(r) \, dr, \end{aligned}$$

$$p(r) = -\log \left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right),$$

and

$$a(r) = r^{k+1} \cdot \left(\int_{S_k} \omega\omega' \, f(z + r\omega) \lambda_{S_k}(d\omega) \right) \frac{\left(\int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{m-1}}{\left(1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds \right)^{m}}.$$

That is, $a(r) = q(r)b(r)$, $q(r) = r^{k+1}c(r)$, and $b(r) = (g(r))^{m-1}$, where

$$c(r) = \frac{\int_{S_k} \omega\omega' \, f(z + r\omega) \lambda_{S_k}(d\omega)}{1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds},$$

$$g(r) = \frac{\int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds}{1 - \int_0^r s^{k-1} \left(\int_{S_k} f(z + s\omega) \lambda_{S_k}(d\omega) \right) ds}.$$

As shown above, $q(0) = 0$ and the first non-zero derivative of $q(r)$ at zero is

$$\frac{d^{k+1}q}{dr^{k+1}}(0) = (k+1)! f(z) \int_{S_k} \omega\omega' \, \lambda_{S_k}(d\omega) = (k+1)! f(z) \frac{\int_{S_k} \lambda_{S_k}(d\omega)}{k} \cdot I,$$

and

$$\frac{d^{(m-1)k}b}{dr^{(m-1)k}}(0) = \frac{((m-1)k)!}{k^{m-1}} \cdot \left(f(z) \int_{S_k} \lambda_{S_k}(d\omega) \right)^{m-1}.$$

Therefore, the first non-zero derivative of $a(r)$ at zero is

$$\begin{aligned}\frac{d^{mk+1}a}{dr^{mk+1}}(0) &= \begin{pmatrix} mk+1 \\ k+1 \end{pmatrix} \frac{d^{(m-1)k}b}{dr^{(m-1)k}}(0) \frac{d^{k+1}q}{dr^{k+1}}(0) \\ &= (mk+1)! \left(f(z) \frac{\int_{S_k} \lambda_{S_k}(d\omega)}{k} \right)^m \cdot I.\end{aligned}$$

So, in a neigborhood of zero we have that

$$a(r) = \sum_{t=0}^{\infty} a(1) r^{mk+1+t}, \quad \text{with} \quad a(1) = \frac{1}{(mk+1+t)!} \frac{d^{mk+1+t}a}{dr^{mk+1+t}}(0).$$

It has been shown above that $p(0) = 0$ and the first non-zero derivative of $p(r)$ at zero is

$$\frac{d^k p}{dr^k}(0) = (k-1)! f(z) \int_{S_k} \lambda_{S_k}(d\omega).$$

In a neigborhood of zero we have that:

$$p(r) = \sum_{t=0}^{\infty} p(1) r^{k+t}, \quad \text{with} \quad p(1) = \frac{1}{(k+t)!} \frac{d^{k+t}p}{dr^{k+t}}(0).$$

Applying Theorem 8.1 in Olver (1974), we get

$$\begin{aligned}I_{m,N} &= \Gamma\left(\frac{mk+2}{k}\right) \frac{a_0}{k p_0^{(mk+2)/k}} \frac{1}{N^{(mk+2)/k}} + o\left(\frac{1}{N^{(mk+2)/k}}\right) \\ &= \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{k} \left(\frac{f(z)}{k} \int_{S_k} \lambda_{S_k}(d\omega) \right)^{-2/k} \frac{1}{N^{(mk+2)/k}} \cdot I + o\left(\frac{1}{N^{(mk+2)/k}}\right).\end{aligned}$$

Now, since

$$\frac{N^m/(m-1)!}{N \begin{pmatrix} N-1 \\ m-1 \end{pmatrix}} - 1 = o(1),$$

we have that

$$ED_{m,N}D'_{m,N} = \Gamma\left(\frac{mk+2}{k}\right) \frac{1}{(m-1)!k} \left(f(z) \frac{\pi^{k/2}}{\Gamma\left(1+\frac{k}{2}\right)} \right)^{-2/k} \frac{1}{N^{2/k}} \cdot I + o\left(\frac{1}{N^{2/k}}\right).$$

Using the same techniques as for the first two moments, $E\|D_{m,N}\|^3 = N \cdot I_{m,N}$, where

$$I_{m,N} = \int_0^\infty e^{-Np(r)}\, a(r)\, dr,$$

$$p(r) = -\log\left(1 - \int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right) ds\right),$$

and

$$a(r) = r^{k+2} \cdot \left(\int_{S_k} f(z+r\omega)\lambda_{S_k}(d\omega)\right) \frac{\left(\int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right) ds\right)^{m-1}}{\left(1 - \int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right) ds\right)^m}.$$

That is, $a(r) = q(r)b(r)$, $q(r) = r^{k+2}c(r)$, and $b(r) = (g(r))^{m-1}$, where

$$c(r) = \frac{\int_{S_k} f(z+r\omega)\lambda_{S_k}(d\omega)}{1 - \int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right) ds},$$

$$g(r) = \frac{\int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right) ds}{1 - \int_0^r s^{k-1}\left(\int_{S_k} f(z+s\omega)\lambda_{S_k}(d\omega)\right) ds}.$$

All the derivatives of q at zero of order smaller than $k+2$ are equal to zero. In addition, all the derivatives of b at zero of order smaller than $(m-1)k$ are equal to zero. Therefore, all the derivatives of a at zero of order smaller than $mk+2$ are equal to zero. Applying the results in Olver (1974), we obtain

$$I_{m,N} = O\left(\frac{1}{N^{(mk+3)/k}}\right).$$

Therefore

$$E\|D_{m,N}\|^3 = O\left(\frac{1}{N^{3/k}}\right) = o\left(\frac{1}{N^{2/k}}\right).$$

□

Proof of Lemma 4 Order the observations so that the first N_1 observations have $t_i = 1$ and the last N_0 observations have $t_i = 0$. Then partition the matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{pmatrix} \mathcal{I}_{N_1} & -\mathbf{A}_{01}/M \\ \mathbf{A}_{10}/M & -\mathcal{I}_{N_0} \end{pmatrix},$$

with $\mathcal{I}_{N_1}$ and $\mathcal{I}_{N_0}$ identity matrices of rank N_1 and N_0 respectively. Note that $\mathbf{A}_{01}$ and $\mathbf{A}_{10}$ are $N_0 \times N_1$ and $N_1 \times N_0$ dimensional matrices with all elements equal to zero or one. In addition, $\mathbf{A}_{01}\iota_{N_1}$ is a N_0 dimensional vector with all elements equal to M, and $\mathbf{A}_{10}\iota_{N_0}$ is a N_1 dimensional vector with all elements equal to M. The vector $\mathbf{A}'_{01}\iota_{N_1}$ has ith element equal to $K_M(i)$, and $\mathbf{A}'_{10}\iota_{N_0}$ has ith element equal to $K_M(i + N_1)$. Therefore,

$$\iota'_{N_1}\mathbf{A}_{01}\mathbf{A}'_{01}\iota_{N_1} = (\mathbf{A}'_{01}\iota_{N_1})'(\mathbf{A}'_{01}\iota_{N_1}) = \sum_{i \in I_1} K_M(i)^2,$$

because the typical elemenent of $\mathbf{A}'_{01}\iota_{N_1}$ is equal to $K_M(i)$. Similarly

$$\iota'_{N_1}\mathbf{A}_{01}\mathbf{A}'_{01}\iota_{N_1} = \sum_{i \in I_0} K_M(i)^2.$$

After these preliminaries consider the matrix $\mathbf{A}\mathbf{A}'$:

$$\mathbf{A}\mathbf{A}' = \begin{pmatrix} \mathcal{I}_{N_1} + \mathbf{A}_{01}\mathbf{A}'_{01}/M^2 & (\mathbf{A}'_{10} + \mathbf{A}_{01})/M \\ (\mathbf{A}_{10} + \mathbf{A}'_{01})/M & \mathcal{I}_{N_0} + \mathbf{A}_{10}\mathbf{A}'_{10}/M^2 \end{pmatrix}.$$

Hence

$$\begin{aligned} \iota'_N\mathbf{A}\mathbf{A}'\iota_N &= \iota'_{N_1}(\mathcal{I}_{N_1} + \mathbf{A}_{01}\mathbf{A}'_{01}/M^2)\iota_{N_1} + 2\iota'_{N_1}(\mathbf{A}'_{10}/M + \mathbf{A}_{01}/M)\iota_{N_0} \\ &\quad + \iota'_{N_0}(\mathcal{I}_{N_0} + \mathbf{A}_{10}\mathbf{A}'_{10}/M^2)\iota_{N_0} \\ &= N_1 + \iota'_{N_1}\mathbf{A}_{01}\mathbf{A}'_{01}\iota_{N_1}/M^2 + 2(N_1 M + N_0 M) + N_0 + \iota'_{N_0}\mathbf{A}_{10}\mathbf{A}'_{10}\iota_{N_0}/M^2 \\ &= 3N + \frac{1}{M^2}\sum_{i=1}^{N} K_M(i)^2. \end{aligned}$$

□

Lemma 8 *The exact conditional distribution of $K_m(i)$ is,*

$$K_M(i) \;\Big|\; T_1,\ldots,T_N,\{X_j\}_{j\in I_1},T_i=1 \;\sim\; \text{Binomial}\left(N_0,\int_{A_M(i)} f_0(z)dz\right),$$

and

$$K_M(i) \;\Big|\; T_1,\ldots,T_N,\{X_j\}_{j\in I_0},T_i=0 \;\sim\; \text{Binomial}\left(N_1,\int_{A_M(i)} f_1(z)dz\right).$$

Let us consider in more detail, for the special case considered in this section with a scalar covariate, the set $A_M(i)$. First, let $\bar{r}_t(x)$ be the number of units with $T_i = t$ and $X_i \geq x$. Then, define $X_{(i,k)} = X_j$ if $\bar{r}_{T_i}(X_i) - \bar{r}_{T_i}(X_j) = k$, and $\bar{r}_{T_i}(X_i) - \lim_{x\uparrow X_j} \bar{r}_{T_i}(x) = k-1$.

Lemma 9 *The set $A_M(i)$ is equal to the interval*

$$A_M(i) = (X_i/2 + X_{(i,-M)}/2, X_i/2 + X_{(i,M)})/2),$$

with width $(X_{(i,M)} - X_{(i,-M)})/2$.

Lemma 10 *Given $X_i = x$, and $T_i = 1$*

$$2N_1 \cdot \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz \xrightarrow{d} \text{Gamma }(2M,1),$$

and given $X_i = x$ and $T_i = 0$,

$$2N_0 \cdot \frac{f_0(x)}{f_1(x)} \cdot \int_{A_M(i)} f_1(z)dz \xrightarrow{d} \text{Gamma }(2M,1),$$

Proof: We only prove the first part of the lemma. The second part follows exactly the same proof. First we establish that

$$2N_1 f_1(x) \cdot \int_{A_M(i)} dz = N_1 f_1(x)\left(X_{(i,M)} - X_{(i,-M)}\right) \xrightarrow{d} \text{Gamma }(2M,1).$$

Let $F_1(x)$ be the distribution function of X given $T = 1$. Then $Y = F_1(X_{(i,+M)}) - F_1(X_{(i,-M)})$ is the difference in order statistics of the uniform distribution, $2M$ orders apart. Hence the

exact distribution of Y is Beta with parameters $2M$ and N_1. For large N_1, the distribution of $N_1 \cdot Y$ is then Gamma with parameters $2M$ and 1. Now approximate $N_1 Y$ as

$$N_1 Y = N_1 \cdot \left(F_1(X_{(i,M)}) - F_1(X_{(i,-M)}) \right) \approx N_1 f_1(x) \cdot \left(X_{(i,M)} - X_{(i,-M)} \right),$$

which demonstrates the first claim.

Second, we show that

$$2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz - 2N_1 f_1(x) \cdot \int_{A_M(i)} dz = o_p(1).$$

This difference can be written as

$$2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz - 2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(x)dz = 2N_1 \frac{f_1(x)}{f_0(x)} \cdot \left(\int_{A_M(i)} (f_0(z) - f_0(x))dz \right).$$

Because

$$f_0(z) - f_0(x) = f_0'(\tilde{z})'(z - x),$$

we have

$$||f_0(z) - f_0(x)|| \leq \sup_z ||f_0'(z)|| \cdot ||z - x||,$$

and hence we can bound the difference by

$$= 2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} dz \cdot \sup_z ||f_0'(z)|| \cdot \max_i \sup_{z,x \in A_M(i)} ||z - x||.$$

By the first part of the proof

$$2N_1 \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} dz = O_p(1).$$

By Assumption ?? $\sup_z ||f_0'(z)||$ is bounded, and by Lemma **??** $\max_i \sup_{z,x \in A_M(i)} ||z - x||$ is $o_p(1)$. Hence the product is $o_p(1)$. □

Lemma 11 *The number of matches $K_M(i)$ is $Z + o_p(1)$, where*

$$E[Z^2] = M^2 \cdot \frac{(1-p)^2}{p} \int \frac{f_0(x)^2}{f_1(x)} dx + M^2 \cdot \frac{p^2}{(1-p)} \int \frac{f_1(x)^2}{f_0(x)} dx$$
$$+M + M \cdot \frac{(1-p)^2}{2p} \int \frac{f_0(x)^2}{f_1(x)} dx + M \cdot \frac{p^2}{2(1-p)} \int \frac{f_1(x)^2}{f_0(x)} dx + o(1).$$

Proof: Condition on $T_i = 1$ and $X_i = x$. By Lemma 10, we can write

$$2N_1 \cdot \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz = W + o_p(1),$$

where

$$W \sim \text{Gamma } (2M, 1).$$

Let Z be a random variable with conditional distribution of Z given W equal to a binomial distribution with parameters N_0 and $Wf_0(x)/(F_1(x)N_1)$. Then, because the distribution of $K_m(i)$ given $2N_1 \cdot \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz$ is the same as the distribution of Z given W, and because the distribution of $2N_1 \cdot \frac{f_1(x)}{f_0(x)} \cdot \int_{A_M(i)} f_0(z)dz$ converges to that of W, the distribution of $K_M(i)$ converges to that of Z. This demonstrates the first part of the claim.

For the second part we calculate the second moment of Z, initially conditional on W, $\{T_j\}_{j=1,\ldots,N}$, and $X_i = x$. Suppose $T_i = 1$. Then

$$E[Z^2|W, T_i = 1, X_i = x] = \frac{N_0^2 f_0^2(x)}{4N_1^2 f_1^2(x)}W^2 + \frac{N_0 f_0(x)}{2N_1 f_1(x)}W - \frac{N_0 f_0(x)^2}{4N_1 f_1(x)^2}W^2.$$

The latter term is of lower order and will be ignored. Then taking the expectation over W, (with $E[W] = 2M$ and $E[W^2] = 4M^2 + 2M$), still conditioning on $T_i = 1$ and $X_i = x$ gives

$$E[Z^2|T_i = 1, X_i = x] = M^2 \cdot \frac{(1-p)^2 f_0(x)^2}{p^2 f_1(x)^2} + M \cdot \frac{(1-p)f_0(x)}{pf_1(x)} + M \cdot \frac{(1-p)^2 f_0(x)^2}{2p^2 f_1(x)^2}.$$

Similarly,

$$E[Z^2|T_i = 0, X_i = x] = M^2 \cdot \frac{p^2 f_1(x)^2}{(1-p)^2 f_0(x)^2} + M \cdot \frac{pf_1(x)}{(1-p)f_0(x)} + M \cdot \frac{p^2 f_1(x)^2}{2(1-p)^2 f_0(x)^2}.$$

Then taking the expectation over T and X gives the desired result. □

Proof of Lemma 5:

For a calculation of the efficiency bound in the general case see Hahn (1998). To show the result in the lemma, we will demonstrate that

$$3 + \frac{1}{M^2}E[K_M^2] = E\left[\frac{1}{e(X)} + \frac{1}{1-e(X)}\right].$$

First, write

$$e(x) = \frac{pf_1(x)}{pf_1(x) + (1-p)f_0(x)},$$

so that

$$\begin{aligned} E\left[\frac{1}{e(x)}\right] &= \int_x \frac{pf_1(x) + (1-p)f_0(x)}{pf_1(x)} \left(pf_1(x) + (1-p)f_0(x)\right) dx \\ &= 1 + \int_x \frac{(1-p)f_0(x)}{pf_1(x)} \left(pf_1(x) + (1-p)f_0(x)\right) dx \\ &= 1 + (1-p) + \int_x \frac{(1-p)f_0(x)}{pf_1(x)} (1-p)f_0(x) dx \\ &= 1 + (1-p) + \frac{(1-p)^2}{p} \int_x \frac{f_0^2(x)}{f_1(x)} dx. \end{aligned}$$

Using the same argument for $E[1/(1-e(X))]$, and combining the results, we get

$$V_{\text{eff}}/\sigma^2 = E\left[\frac{1}{e(X)} + \frac{1}{1-e(X)}\right] = 3 + \frac{(1-p)^2}{p} \int_x \frac{f_0^2(x)}{f_1(x)} dx + \frac{p^2}{1-p} \int_x \frac{f_0^2(x)}{f_1(x)} .dx$$

Hence we can write

$$\frac{(1-p)^2}{p} \int_x \frac{f_0^2(x)}{f_1(x)} dx + \frac{p^2}{1-p} \int_x \frac{f_0^2(x)}{f_1(x)} .dx = (V_{\text{eff}} - 3)/\sigma^2.$$

Plugging this into the variance for the matching estimator, using the expression for the second moment of K_M, we get

$$\begin{aligned} \sigma^2 \left(3 + \frac{1}{M^2} E[K_M^2]\right) &= V_{\text{eff}} + \frac{\sigma^2}{M} + \frac{1}{M}(V_{\text{eff}} - 3\sigma^2)/2 \\ &= V_{\text{eff}} + \frac{1}{2M}(V_{\text{eff}} - \sigma^2)/2 = V_{\text{eff}} \cdot \left(1 + \frac{1}{2M}\right) - \frac{\sigma^2}{2M} \end{aligned}$$

□

Table 1: Summary Statistics Lalonde Data

	Experimental Data				PSID		T-statistic	
	Trainees (N=185)		Controls (N=260)		Controls (N=2490)		Train/	Train/
	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)	Contr	PSID
Age	25.8	(7.16)	25.05	(7.06)	34.85	(10.44)	[1.1]	[-16.0]
Education	10.4	(2.01)	10.09	(1.61)	12.12	(3.08)	[1.4]	[-11.1]
Black	0.84	(0.36)	0.83	(0.38)	0.25	(0.43)	[0.5]	[21.0]
Hispanic	0.06	(0.24)	0.11	(0.31)	0.03	(0.18)	[-1.9]	[1.5]
Married	0.19	(0.39)	0.15	(0.36)	0.87	(0.34)	[1.0]	[-22.8]
Earnings '74	2.10	(4.89)	2.11	(5.69)	19.43	(13.41)	[-0.0]	[-38.6]
Unempl. '74	0.71	(0.46)	0.75	(0.43)	0.09	(0.28)	[-1.0]	[18.3]
Earnings '75	1.53	(3.22)	1.27	(3.10)	19.06	(13.60)	[0.9]	[-48.6]
Unempl '75	0.60	(0.49)	0.68	(0.47)	0.10	(0.30)	[-1.8]	[13.76]
Earnings '78	6.35	(7.87)	4.55	(5.48)	21.55	(15.56)	[2.7]	[-23.1]
Unempl. '78	0.24	(0.43)	0.35	(0.48)	0.11	(0.32)	[-2.65]	[4.0]

The first two columns give the average and standard deviation of the 185 trainees from the experimental data set. The second pair of columns give the average and standard deviation of the 260 controls from the experimental data set. The third pair of columns give the averages and standard deviations of the 2490 controls from the nonexperimental PSID sample. The seventh column gives t-statistics for the difference between the averages for the experimental trainees and controls. The last column gives the t-statistics for the differences between the averages for the experimental trainees and the PSID controls. The last two variables, earnings '78 and unemployed '78 are post-training. All the others are pre-training variables. Earnings data are in thousands of dollars.

Table 2: Experimental and Non-experimental Estimates of Average Treatment Effects for Lalonde Data

	$M=1$		$M=4$		$M=16$		$M=64$		All Controls	
	mean	(s.d.)	est	(s.e.)	est	(s.e.)	est	(s.e.)	est	(s.e.)
Panel A: Experimental Estimates										
dif	1.23	(0.84)	1.99	(0.79)	1.76	(0.79)	2.20	(0.80)	1.79	(0.75)
reg-all	1.22	(0.84)	1.94	(0.79)	1.57	(0.80)	1.80	(0.82)	1.72	(0.77)
reg-matched	1.17	(0.84)	1.84	(0.80)	1.55	(0.80)	1.74	(0.82)	1.72	(0.77)
Regression Estimates										
dif	1.79	(0.67)								
linear	1.72	(0.65)								
quadratic	2.27	(0.73)								
Panel B: Non-experimental Estimates										
dif	2.09	(1.67)	1.62	(1.71)	0.47	(1.21)	-0.11	(0.90)	-15.20	(0.60)
reg-all	2.56	(1.69)	2.54	(1.68)	1.80	(1.14)	2.29	(0.82)	0.84	(0.62)
reg-matched	2.45	(1.68)	2.51	(1.68)	2.48	(1.15)	2.26	(0.83)	0.84	(0.62)
Regression Estimates										
dif	-15.20	(0.66)								
linear	0.84	(0.86)								
quadratic	3.26	(0.98)								

Panel A reports the results for the experimental data (experimental controls and trainees), and Panel B the results for the nonexperimental data (PSID controls with experimental trainees). In each panel the top part reports results for the matching estimators, with the number of matches equal to 1, 4, 16, 64 and 2490 (all controls). The second part reports results for three regression adjustment estimates, based on no covariates, all covariates entering linearly and all covariates entering with a fully set of quadratic terms and interactions. The outcome is earnings in 1978 in thousands of dollars.

Table 3: Mean Covariate Differences in Matched Groups

	Average		$M = 1$		$M = 4$		$M = 16$		$M = 64$		$M = 2490$	
	PSID	Trainees	dif	(s.d.)	dif	(s.d.)	dif	(s.d.)	dif	(s.d.)	dif	(s.d.)
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 '74	0.09	-1.18	-0.01	(0.10)	-0.01	(0.12)	-0.05	(0.17)	-0.15	(0.30)	-1.26	(0.36)
Unempl '74	-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)
Unempl '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)

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 trainees and the PSID controls. 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 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

M	Estimator	mean bias	median bias	rmse	mae	s.d.	mean s.e.	coverage (nom. 95%)	coverage (nom. 90%)
1	dif	-0.48	-0.43	0.86	0.56	0.71	1.02	0.98	0.96
	reg all	0.25	0.27	0.75	0.53	0.71	1.01	0.99	0.97
	reg matched	0.05	0.08	0.77	0.50	0.77	1.01	0.99	0.97
4	dif	-0.85	-0.84	1.04	0.84	0.59	0.93	0.94	0.88
	reg all	0.15	0.16	0.62	0.40	0.60	0.89	0.99	0.98
	reg matched	0.05	0.06	0.61	0.41	0.61	0.89	0.99	0.99
16	dif	-1.81	-1.80	1.90	1.80	0.59	0.66	0.20	0.11
	reg all	-0.51	-0.47	0.88	0.58	0.71	0.60	0.82	0.74
	reg matched	0.17	0.17	0.62	0.40	0.60	0.59	0.94	0.89
64	dif	-3.29	-3.27	3.34	3.27	0.60	0.50	0.00	0.00
	reg all	-0.84	-0.84	1.17	0.88	0.82	0.45	0.51	0.43
	reg matched	0.16	0.17	0.66	0.42	0.65	0.43	0.79	0.71
All (2490)	dif	-19.05	-19.04	19.06	19.04	0.62	0.43	0.00	0.00
	reg all	-2.04	-2.05	2.26	2.05	0.99	0.37	0.09	0.07
	reg matched	-2.04	-2.05	2.26	2.05	0.99	0.37	0.09	0.07
difference		-19.05	-19.04	19.06	19.04	0.62	0.66	0.00	0.00
linear regression		-2.04	-2.05	2.26	2.05	0.99	0.97	0.44	0.32
quadratic regression		2.73	2.68	3.05	2.68	1.35	1.16	0.38	0.27

For each estimator summary statistics are provided for 10,000 replications of the data set. Results are reported for five values of the number of matches ($M = 1, 4, 16, 64, 2490$), and for three estimators: the difference in matched outcomes, the difference adjusted by the regression of all treated and controls on the covariates, and the difference adjusted by the regression on only the matched treated and controls. The last three rows report results for the simple average treatment-control difference, the ordinary least squares estimator, and the ordinary least square estimator using a full set of quadratic terms and interactions. For each estimator we report the mean and median bias, the root-mean-squared-error, the median-absolute-error, the standard deviation of the estimators, the average estimate of the standard error, and the coverage rate of the nominal 95% and 90% confidence intervals.