A MATCHING ESTIMATOR BASED ON A BILEVEL OPTIMIZATION PROBLEM

Juan Díaz, Tomás Rau, and Jorge Rivera*

Abstract—This paper proposes a novel matching estimator where neighbors used and weights are endogenously determined by optimizing a covariate balancing criterion. The estimator is based on finding, for each unit that needs to be matched, sets of observations such that a convex combination of them has the same covariate values as the unit needing matching or with minimized distance between them. We implement the proposed estimator with data from the National Supported Work Demonstration, finding outstanding performance in terms of covariate balance. Monte Carlo evidence shows that our estimator performs well in designs previously used in the literature.

I. Introduction

MATCHING estimators have been widely used in the impact evaluation literature during the past decades. These methods essentially rest on the imputation of a potential outcome to an individual, built as a weighted average of the observed outcomes of his or her closest neighbors from the corresponding counterfactual set. Given that, the individual treatment effect is usually defined as the difference between the imputed and actual outcome. One popular estimator is the simple matching estimator (also known as nearest neighbor matching estimator) studied by Abadie and Imbens (2006), where the outcome to be imputed is defined as the average of the outcomes from a certain number of closest neighbors, with closeness defined by distance induced by a norm. The choice of the number of neighbors is up to the researcher, and the weights are simply the reciprocal of this number. As Imbens and Wooldridge (2009) state, "Little is known about the optimal number of matches, or about data-dependent ways of choosing it."

In this paper, we propose a matching estimator where the number of units used in each match and the weighting scheme are endogenously determined from the solution of an optimization problem. The first task deals with finding sets of observations such that a convex combination of them has exactly the same covariate values as the unit to be matched, when possible, or otherwise where their distance is minimized. Since this problem may have more than one solution, the second task consists of implementing a refinement criterion that looks for the set with the closest covariate values to the unit to be matched. These problems (involving the two tasks described) can be written as one optimization problem whose admissible points belong to the solution set of another optimization problem. In the optimization literature this is called a bilevel optimization problem (BLOP).[1]

To fix ideas, assume we are interested in finding a match for individual i with a unique real-valued characteristic $X_i \in \mathbb{R}$ and outcome variable $Y_i \in \mathbb{R}$. Also assume that individuals in the counterfactual group are indexed by $j = 1, \ldots, n,\ n > 2$, with characteristics (covariates) and outcomes given by $X_j \in \mathbb{R}$ and $Y_j \in \mathbb{R}$, respectively. Without loss of generality, let us assume that $X_1 < ... < X_k < X_i < X_{k+1} < \ldots < X_n$. The first task of this optimization problem is to look for individuals in the counterfactual group such that the convex combination of their characteristics is as close as possible to X_i. This problem has many solutions. For instance, combining X_k and X_{k+1} will match X_i, but the combination of X_1, X_2, and X_n will also get an exact match. Hence, the second task is another optimization problem that identifies the solution that minimizes the sum of distances between X_i and the characteristics from individuals used from the counterfactual group. In this example, the solution of the BLOP is the weighting scheme given by

$$\lambda_k = \frac{X_{k+1} - X_i}{X_{k+1} - X_k}, \quad \lambda_{k+1} = \frac{X_i - X_k}{X_{k+1} - X_k},$$
$$\lambda_j = 0,\ j \neq k,\ k+1,$$

which exactly matches X_i.[2] Hence, our approach uses only units k and $k+1$ to perform the match, and according to this approach, the potential outcome to be imputed to individual i is

$$\widehat{Y}_i = \sum_{j=1}^{n} \lambda_j\, Y_j = \lambda_k\, Y_k + \lambda_{k+1}\, Y_{k+1}.$$

When $X \in \mathbb{R}^K$ is a K-dimensional vector of characteristics ($K > 1$), solving the BLOP could be difficult due to the great number of alternatives that any optimization algorithm has to evaluate in order to achieve a global solution. To circumvent this complexity, we present an equivalent formulation of the BLOP as a linear programming problem, for which there is a vast literature in optimization that allows us to solve the BLOP efficiently.

Our estimator is related to recent methods that use covariate balance as a metric for selecting either the propensity

Received for publication January 31, 2012. Revision accepted for publication August 28, 2014. Editor: David S. Lee.

* Díaz: Universidad de Chile; Rau: Pontificia Universidad Católica de Chile; Rivera: Universidad de Chile.

We thank Roberto Cominetti, Guido Imbens, Esteban Puentes, and Sebastian Otero for helpful discussions and seminar participants at the 2011 annual meeting of SECHI and the 2012 NASM (Evanston). We also thank George Vega and Daniel Espinoza for helping us with the library and Justin McCrary for providing us copies of the code used to generate the results in Busso, DiNardo, and McCrary (2014). An R library that implements the proposed estimator can be downloaded from http://www.economia.puc.cl/tomas-rau. The usual disclaimer applies. We are grateful for the funding provided by Fondecyt, project 1095181, and Instituto Sistemas Complejos de Ingeniería. T.R. thanks the funding provided by the Iniciativa Científica Milenio, project NS100041.

A supplemental appendix is available online at http://www.mitpress journals.org/doi/suppl/10.1162/REST_a_00504.

[1] For details see Colson, Marcotte, and Savard (2007).

[2] It is straightforward to check that $X_i = \lambda_k X_k + \lambda_{k+1} X_{k+1}$. Additionally, as will be discussed later, the number of matches will be determined by a refinement criterion that minimizes the sum of distances between X_i and the characteristics from the units used from the counterfactual group.

The Review of Economics and Statistics, October 2015, 97(4): 803–812

doi:10.1162/REST_a_00504

score model or tuning parameters. Diamond and Sekhon (2013) propose a search algorithm to iteratively check and improve covariate balance. This is done by using a generalized version of the Mahalanobis distance that places different weights on the covariates used in the matching algorithm. The algorithm then iteratively chooses the weighting matrix that provides the best covariate balance according to a loss function depending on measures of imbalance such as the Kolmogorov-Smirnov test statistic. These weights are common for all the units to be matched, generating a distance metric to perform the matching optimizing postmatching covariate balance. Our approach differs from theirs since it looks for different weights that minimize the distance between the match (built with these weights) and the unit. This optimization is done for each observation that needs to be matched so the weights are different observation by observation. Thus, the method we develop is an optimization problem (not an iterative algorithm) that jointly finds the weights and determines the number of neighbors used for each observation. The choice of neighbors is determined by the fact that units with zero weight are then not used in the match. Meanwhile, Graham, Pinto, and Egel (2012) propose a modified inverse probability weighting estimator that also maximizes covariate balance and has the property of being doubly robust (see Robins, Rotnitzky, & Zhao, 1994, 1995). These weights are optimally estimated to balance the covariates and inversely depend on estimates of the propensity score. Finally, Imai, and Ratkovic (2014) propose an inverse propensity score weighting estimator that simultaneously optimizes the covariate balance and the prediction of treatment assignment. The last two approaches are inverse probability weighting methods that require the estimation of the propensity score. Our method, however, is a nonparametric matching method in characteristics with an optimal choice of weights that by construction improves the covariate balance.

Our paper is also related to the literature that tries to determine the number of neighbors to use when doing matching or tuning parameters. Although there is not a theoretical foundation, many researchers use cross-validation to select the number of neighbors. For instance, Frölich (2004) uses that in his investigations of finite sample properties of matching estimators. Also, Galdo, Smith, and Black (2008) discuss how to select the bandwidth when doing matching following a weighted cross-validation strategy. The BLOP strategy solves the issue of the number of neighbors selection since the optimal weights that are different from 0 define the units from the counterfactual group that participate in the optimal convex combination. Indeed, from Caratheodory's theorem (see Rockafellar, 1972), the number of units should be at most the pretreatment characteristic vector dimension plus 1. This is relevant since when the estimation is implemented employing the entire sample (as we propose), there is no need to fix the number of neighbors to be used for estimation.

Regarding large sample properties, the BLOP matching estimator is consistent and asymptotically normal, under regularity conditions assumed in the literature (see Abadie & Imbens, 2006).[3] We also provide a consistent estimator of its marginal variance.

To assess the performance of the BLOP estimator in finite samples, we implement different empirical designs and Monte Carlo exercises. We use data from the National Supported Work Demonstration (NSW) to see its performance compared to its natural competitor: the nearest-neighbor matching estimator. Then, using the data-generating processes from Busso, DiNardo, and McCrary (2014), we implement Monte Carlo experiments in order to assess its performance in terms of absolute bias, variance, and covariate balance. We find significant improvements in comparison to other matching estimators employed in the literature, especially when the propensity score is underspecified.

This paper is organized as follows. The BLOP matching estimator is introduced in section II, while in section III, we study its performance using data from the NSW to assess its performance compared to a simple matching estimator and perform the empirical Monte Carlo exercise. In section IV, we study the finite sample performance of our estimator under misspecification. Section V concludes.

II. The BLOP Matching Estimator

In this section we illustrate and formalize our proposed matching estimator beginning with introducing some basic notations concerning binary program evaluation and certain mathematical concepts needed for properly setting up the method.

A. *Basic Concepts and Notation*

Following Imbens and Wooldridge (2009), we denote by $K \in \mathbb{N}$ the number of pretreatment characteristics or covariates. For a unit i, $W_i \in \{0, 1\}$ indicates whether the treatment was received ($W_i = 1$) or not ($W_i = 0$), $X_i \in \mathbb{R}^K$ is the vector of covariates, and $Y_i = W_i\, Y_i(1) + (1 - W_i)\, Y_i(0) \in \mathbb{R}$ is the observed outcome, with $Y_i(1)$ and $Y_i(0)$ the outcome that this unit would have obtained because of treatment or its absence, respectively.

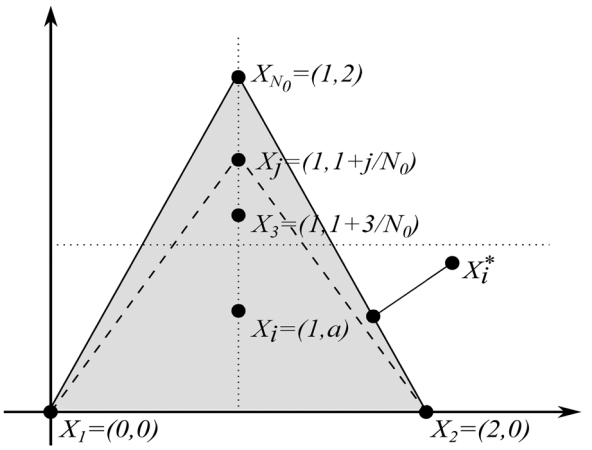
The matching metric we use is given by the Euclidean norm in $\mathbb{R}^K$, denoted by $\|\cdot\|$, which means that the nearest neighbors to each unit are determined according to the Euclidean distance between corresponding covariates.[4]

As far as mathematical concepts are concerned, a vector sum $\lambda_1 X_1 + \ldots + \lambda_N X_N$ is called a convex combination of vectors $X_1, \ldots, X_N \in \mathbb{R}^K$ if the coefficients λ_j are all nonnegative and $\lambda_1 + \ldots + \lambda_N = 1$. The set of these weights is the simplex of dimension N, hereafter denoted by $\Delta_N = \{(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N_+,\ \sum_{j=1}^N \lambda_j = 1\}$. The convex hull of $\{X_1, \ldots, X_N\}$ is the set of all convex combinations

[3] Large sample properties are developed in detail in the online appendix.

[4] Our approach can be extended to consider any other norm or even a balancing score (see Rosenbaum & Rubin, 1983) by properly configuring the setting we will present.

FIGURE 1.—SPATIAL CONFIGURATION OF COVARIATES $\{X_1, \ldots, X_{N_0}\}$ AND THEIR CONVEX HULL



of these vectors, which throughout this paper is denoted as $\text{co}\{X_1, \ldots, X_N\} = \left\{ \sum_{j=1}^{N} \lambda_j X_j,\ (\lambda_1, \ldots, \lambda_N) \in \Delta_N \right\}$.

We recall that the projection of $X \in \mathbb{R}^K$ onto $\text{co}\{X_1, \ldots, X_N\}$ is, by definition, the nearest vector to X belonging to that set. Denoting it by $\text{Proj}(X)$, it follows that[5]

$$\|X - \text{Proj}(X)\| = \min_{Z \in \text{co}\{X_1, \ldots, X_N\}} \|X - Z\| = \min_{(\lambda_1, \ldots, \lambda_N) \in \Delta_N} \left\| X - \sum_{j=1}^{N} \lambda_j X_j \right\|. \quad (1)$$

B. *Motivating Example*

Consider we are interested in evaluating the effect of a binary treatment due to a social program on some outcome variable. Let the dimension of the covariate vector be $K = 2$, and take a treated unit i, with observed outcome $Y_i(1) = Y_i \in \mathbb{R}$, and covariates $X_i = (1, a) \in \mathbb{R}^2$, with $0 \leq a \leq 1$. In this program, there are $N_0 \geq 3$ control units indexed by $j = 1, \ldots, N_0$, each one having observed outcome $Y_j(0) = Y_j \in \mathbb{R}$, and covariates $X_1 = (0, 0)$, $X_2 = (2, 0)$ and $X_j = (1, 1 + j/N_0)$ for $j = 3, \ldots, N_0$. Figure 1 illustrates the configuration of these covariates. There, the convex hull of $\{X_1, \ldots, X_{N_0}\}$ is the shaded triangle with vertices X_1, X_2, and X_{N_0}.

For a given $j \in \{3, \ldots, N_0\}$, highlighted by dashed lines in figure 1, we have that X_i belongs to the interior of the triangle with vertices X_1, X_2, and X_j, which is equivalent to stating that there is a vector of weights $(\lambda_1, \ldots, \lambda_{N_0}) \in \Delta_{N_0}$ such that $X_i = \sum_{s=1}^{N_0} \lambda_s X_s$, with $\lambda_s > 0$ for $s = 1, 2, j$, and 0 otherwise. In fact, for a given j as before, it is easy to see the nonnull components of this vector of weights are given by

$$\lambda_1 = \frac{j + (1 - a) N_0}{2 (j + N_0)}, \quad \lambda_2 = \frac{j + (1 - a) N_0}{2 (j + N_0)}, \quad \lambda_j = \frac{a N_0}{j + N_0}. \quad (2)$$

Consequently, varying j from 3 to N_0, we have at least $N_0 - 2$ convex combinations of $X_1, \ldots, X_{N_0}$ that match X_i in an exact manner. Leading from this is the question of which of these convex combinations should be chosen to match X_i. To answer this question, we first note that any of these convex combinations exactly match X_i, which makes us consider a refinement criterion. Thus, based on a continuity assumption of the relationship between covariates and outcome, the refined solution proposed is the convex combination that while performing a perfect match has the closest covariates to X_i. From figure 1, this refined solution is defined by the triangle whose vertices are X_1, X_2, and X_3. Given that, and using these covariates with weights from equation (2) with $j = 3$, we have a perfect match, that is,

$$X_i = \left(\frac{3 + (1 - a) N_0}{2 (3 + N_0)} \right) X_1 + \left(\frac{3 + (1 - a) N_0}{2 (3 + N_0)} \right) X_2 + \left(\frac{a N_0}{3 + N_0} \right) X_3.$$

At this stage, it is important to note that when the matching is performed using the nearest-neighbor approach, a perfect match is not guaranteed. For instance, when the parameter a is equal to 1 and the number of neighbors chosen is equal to 3, we have that the nearest neighbors to unit i are units 3, 4, and 5. It is easy to check that convex combinations of these units' covariate will not achieve a perfect match for X_i. This occurs since the nearest-neighbor approach does not incorporate an explicit covariate balancing criterion when choosing the neighbors.

Our refined solution solves an optimization problem that optimizes individual covariate balance. In this regard, the objective function proposed is the weighted sum of these distances with weights used to perform the convex combination. As seen in the next section, by using this objective function, the proposed optimization problem becomes a linear program implying a straightforward numerical implementation. Formally, when evaluated in the convex combination with covariates X_1, X_2, X_j and weights from relation (2), the objective function value is

$$\lambda_1 \|X_i - X_1\| + \lambda_2 \|X_i - X_2\| + \lambda_j \|X_i - X_j\| = (\sqrt{1 + a^2} + a) \left(1 - \frac{a N_0}{j + N_0} \right), \quad (3)$$

which clearly attains a minimum value when $j = 3$. This corresponds to the triangle with the closest covariates to X_i that was referred to previously as the refined solution to our problem. Consequently, among the convex combinations performing the best possible balance of X_i, our refined solution minimizes the measure of performance, equation (3). Thus,

[5] The uniqueness of this point comes from the fact that $\text{co}\{X_1, \ldots, X_N\}$ is a convex and compact set. See Rockafellar (1972) for details on properties of convex sets.

the weighting scheme is given by relation (2) evaluated at $j = 3$, implying that the missing potential outcome to impute to unit i according to this approach is

$$\widehat{Y}_i(0) = \left(\frac{3 + (1-a)\,N_0}{2\,(3+N_0)}\right) Y_1 + \left(\frac{3 + (1-a)\,N_0}{2\,(3+N_0)}\right) Y_2 + \left(\frac{a\,N_0}{3+N_0}\right) Y_3,$$

where the estimated individual treatment effect for this unit is given by $Y_i - \widehat{Y}_i(0)$.

When the unit needing matching does not belong to the counterfactual group's convex hull of covariates, we propose to find the convex combination that exactly balances the projection of its covariates onto that convex hull. Illustrated in figure 1, this case occurs when X_i^*, instead of X_i, is the vector of covariates of unit i. Denoting by $\text{Proj}(X_i^*)$ the projection of X_i^* onto $\text{co}\{X_1, \ldots, X_{N_0}\}$, our solution uses covariates X_2 and X_{N_0} to perform that projection, with optimal weights proportional to the distance from $\text{Proj}(X_i^*)$ to X_2 and $\text{Proj}(X_i^*)$ to X_{N_0}.

Remark 1. The weighting scheme described depends on the covariates of the unit that needs to be matched and the covariates of the units participating in the matching (i.e., having strictly positive weights). That we optimize over the entire sample of opposites for determining such optimal weights does not imply we are necessarily using all of them when matching X_i. Indeed, as a direct consequence of Carathedory's theorem (see Rockafellar, 1972), the number of units used is usually fewer than or equal to the number of covariates plus 1: $K+1$, usually far from the number of opposites in the sample. Consequently, by using this approach, we would avoid an exogenous number of neighbors, a crucial tuning parameter for most matching methods currently available (see Imbens & Wooldridge, 2009).

C. *Formal Aspects of the Bilevel Matching Estimator*

Following the notation introduced in section IIA, the sample of covariates, outcomes, and treatment assignment is denoted by $\{(X_i, Y_i, W_i)\}_{i=1}^N$, with $N \in \mathbb{N}$ the sample size, and N_1 and N_0 the number of treated and control units, respectively. We note that for each unit i, the number of units in its opposite treatment group is $N_{1-W_i} \in \{N_0, N_1\}$. By reordering, in what follows we assume that control units are indexed $1, \ldots, N_0$; thus, the treated ones are labeled by $N_0 + 1, \ldots, N_0 + N_1\ (= N)$.

For each unit to be matched, we show that the proposed weighting scheme solves a linear optimization problem; thus, the estimation can be efficiently implemented using optimization routines available on the our web pages. To do so, without loss of generality, we describe the approach for a treated unit $i \in \{N_0 + 1, \ldots, N\}$. Now, since $\text{Proj}(X_i) = X_i$ if and only if $X_i \in \text{co}\{X_1, \ldots, X_{N_0}\}$, there is no reason to present the method separately for the cases mentioned in section IIB, that is, whether X_i belongs to the convex hull of opposite units' covariates. Hence, we present the method for matching the projection that allows us to write the problem, as a linear program, as shown next. Given that, we now note any weighting scheme that serves to match $\text{Proj}(X_i) \in \text{co}\{X_1, \ldots, X_{N_0}\}$ solves the next optimization problem (see relation [1] in section IIA):

$$\mathcal{P}_i \quad : \quad \min_{(\lambda_1,\ldots,\lambda_{N_0}) \in \Delta_{N_0}} \left\| X_i - \sum_{j=1}^{N_0} \lambda_j X_j \right\|. \tag{4}$$

From the example in section IIB, we have that the solution set of problem (4), namely, $\text{argmin}(\mathcal{P}_i) \subseteq \Delta_{N_0}$, may have more than one element. For any of them, say, $(\lambda_1, \ldots, \lambda_{N_0}) \in \text{argmin}(\mathcal{P}_i)$, the measure of performance, evaluated in the convex combination that uses this vector of weights, is given by

$$\sum_{j=1}^{N_0} \lambda_j \, \|X_i - X_j\|.$$

Consequently, adapting the refinement criterion introduced in section IIB to this setting, the weighting scheme we are looking for solves the next optimization problem:

$$\begin{aligned} &\min_{(\lambda_1,\ldots,\lambda_{N_0})} \sum_{j=1}^{N_0} \lambda_j \, \|X_i - X_j\| \\ &\text{s.t.} \\ &\sum_{j=1}^{N_0} \lambda_j X_j = \text{Proj}(X_i), \\ &\sum_{j=1}^{N_0} \lambda_j = 1, \quad \lambda_j \geq 0,\ j = 1, \ldots, N_0. \end{aligned} \tag{5}$$

Due to the set of admissible points of problem (5) are the solutions of another optimization problem, namely, equation (4), it is called a bilevel optimization problem, BLOP (see Colson et al., 2007).

The main difficulty in solving linear optimization problem (5) is determining the projection of X_i onto the convex hull of opposite units covariates, $\text{Proj}(X_i)$, which can be solved efficiently using methods currently available in the optimization literature (see Botkin & Stoer, 2005).

When i is a control unit, the BLOP must be solved using covariates $X_{N_0+1}, \ldots, X_N$; thus, the optimal weighting scheme belongs to Δ_{N_1}. By properly configuring the BLOP in terms of covariates, the solution of problem (5) for any unit $i \in \{1, \ldots, N\}$ is denoted by

$$\lambda(i) = \left(\lambda_1(i), \ldots, \lambda_{N_{1-W_i}}(i)\right) \in \Delta_{N_{1-W_i}}. \tag{6}$$

Given that, by setting

$$\widehat{Y}_i^b = W_i \sum_{j=1}^{N_0} \lambda_j(i)\, Y_j + (1 - W_i) \sum_{j=1}^{N_1} \lambda_j(i)\, Y_{N_0+j},$$

the missing potential outcome we impute to unit i is

$$\widehat{Y}_i(0) = \begin{cases} Y_i & \text{if} \quad W_i = 0 \\ \widehat{Y}_i^b & \text{if} \quad W_i = 1 \end{cases},$$

$$\widehat{Y}_i(1) = \begin{cases} \widehat{Y}_i^b & \text{if} \quad W_i = 0 \\ Y_i & \text{if} \quad W_i = 1. \end{cases}$$

Thus, our proposed matching estimator is defined as follows:

Definition 1. *The BLOP matching estimator for the average treatment effect and the average treatment effect on the treated are, respectively,*

$$\widehat{\tau}^b = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right),$$

$$\widehat{\tau}^b_{tre} = \frac{1}{N_1} \sum_{i=1}^{N} W_i \left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right). \tag{7}$$

D. Variance and Large Sample Properties

We present marginal variance estimators for the average treatment effect and the average treatment effect on the treated. Following Abadie and Imbens (2006), a variance estimator of the conditional mean of $\widehat{\tau}^b$ is given by

$$\widehat{\mathbb{V}}\left(\mathbb{E}(\widehat{\tau}^b \mid \{X_i, W_i\}_{i=1}^N)\right) = \frac{1}{N^2} \sum_{i=1}^{N} \left(\left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\tau}^b\right)^2 - \left(1 + c_i^{[2]}\right) \widehat{\sigma}_i^2 \right), \tag{8}$$

while the estimator for the expected value of the conditional variance is

$$\widehat{\mathbb{E}}\left(\mathbb{V}\left(\widehat{\tau}^b \mid \{X_i, W_i\}_{i=1}^N\right)\right) = \frac{1}{N^2} \sum_{i=1}^{N} \left(1 + c_i^{[1]}\right)^2 \widehat{\sigma}_i^2, \tag{9}$$

where $c_i^{[\alpha]}$, $\alpha = 1, 2$, is the sum of weights, to the power of α, associated to unit i when used as a counterfactual individual,[6] and $\widehat{\sigma}_i^2$ is a BLOP matching estimator of the conditional variance $\sigma^2(X_i, W_i) = \mathbb{V}(Y \mid X = X_i, W = W_i)$.[7] Using equations (8) and (9), we have a consistent estimator of the variance of $\widehat{\tau}^b$,

$$\widehat{V}(\widehat{\tau}^b) = \frac{1}{N^2} \sum_{i=1}^{N} \left(\left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\tau}^b\right)^2 + \left[\left(1 + c_i^{[1]}\right)^2 - \left(1 + c_i^{[2]}\right) \right] \widehat{\sigma}_i^2 \right),$$

and restricting this estimator to the subsample of treated units, after some simple manipulation, we have the estimator of the variance of $\widehat{\tau}^b_{tre}$:

$$\widehat{V}(\widehat{\tau}^b_{tre}) = \frac{1}{N_1^2} \sum_{i=1}^{N} \left(W_i \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\tau}^b_{tre}\right)^2 + (1 - W_i) \left[\left(c_i^{[1]}\right)^2 - c_i^{[2]} \right] \widehat{\sigma}_i^2 \right).$$

Asymptotic properties are studied in the online appendix, where we show, under suitable conditions, the consistency of the variance estimator for $\widehat{\tau}^b$, the asymptotic normality of the BLOP matching estimator for the average treatment effect, and that the conditional bias of $\widehat{\tau}^b$ is $O_p\left(N^{-1/K}\right)$. Some of the proofs are straightforward extensions of those from Abadie and Imbens (2006) for the simple matching estimator.

III. Empirical Application and Monte Carlo Evidence

In this section we implement the proposed estimator to data from the National Supported Work (NSW) Demonstration and evaluate its performance in an empirical Monte Carlo design based on these data.

A. NSW Demonstration

In order to assess the performance of the BLOP matching estimator, we provide estimates of the average treatment effect on the treated (ATT) using two control groups: the control group from the experimental sample from Lalonde (1986) and a control group from the Panel Study of Income Dynamics (PSID) used by Dehejia and Wahba (1999), Smith and Todd (2005), and Abadie and Imbens (2011), among others.

Table 1 presents some summary statistics of the data. As can be seen from the experimental data, treated and control units are well balanced in terms of sample means. Indeed, we fail to reject the null hypothesis of the differences being equal to 0 for the nine covariates considered. However, when comparing the treated units from the experimental data with those from the control group in the nonexperimental data (PSID), we can see that the samples differ significantly in terms of first moment for eight of the nine covariates. The only case in which we fail to reject the null hypothesis of means equality is for the Hispanic dummy variable. Thus, using the nonexperimental data is an interesting scenario to check the balancing properties of the proposed estimator.

In table 2 we present the ATT results for both control groups. In addition, we compare our estimator with those

[6] We recall that control units are indexed by $1, \ldots, N_0$, while the treated ones are labeled by $N_0+1, \ldots, N_0+N_1$. This implies that for a given integer α, when i is a treated unit, $c_i^{[\alpha]} = \sum_{j=1}^{N_0} \left(\lambda_{i-N_0}(j)\right)^\alpha$, and when i is a control unit, $c_i^{[\alpha]} = \sum_{j=1}^{N_1} \left(\lambda_i(N_0 + j)\right)^\alpha$. In each case, this comes directly from the the solution of the optimization problem (5).

[7] The BLOP matching estimator of the conditional variances requires solving problem (5) for each unit that needs to be matched, using covariates from units in the same treatment group and leaving the ith unit out, instead of using covariates from units in the the opposite treatment group. See the online appendix for more details.

TABLE 1.—SUMMARY STATISTICS

	Experimental Data				Nonexperimental PSID		*P*-Value	
	Treated (185)		Control (260)		Control (2490)			
Variable	Mean	(SD)	Mean	(SD)	Mean	(SD)	Treat/Control Experiment	Treat/ Control PSID
Age	25.8	(7.2)	25.1	(7.1)	34.9	(10.4)	0.27	0.00
Education	10.3	(2.0)	10.1	(1.6)	12.1	(3.1)	0.15	0.00
Black	0.84	(0.36)	0.83	(0.38)	0.25	(0.43)	0.65	0.00
Hispanic	0.06	(0.24)	0.11	(0.31)	0.03	(0.18)	0.06	0.13
Married	0.19	(0.39)	0.15	(0.36)	0.87	(0.34)	0.33	0.00
Earnings '74	2.10	(4.89)	2.11	(5.69)	19.4	(13.41)	0.98	0.00
Earnings '75	1.53	(3.22)	1.27	(3.10)	19.1	(13.60)	0.39	0.00
Unemployed '74	0.71	(0.46)	0.75	(0.43)	0.09	(0.30)	0.33	0.00
Unemployed '75	0.60	(0.49)	0.68	(0.47)	0.10	(0.28)	0.07	0.00

Earnings are expressed in thousands of 1978 dollars. The last two columns show the *p*-values for the differences in means test between treated and controls from the two samples.

TABLE 2.—ESTIMATES FOR THE NSW DATA

		Neighbors				
	Optimal	$k = 1$	$k = 4$	$k = 16$	$k = 64$	All
A: Experimental control group						
BLOP	1,728.9	–	–	–	–	–
	(764.1)					
NN covariates	–	1,223.2	1,994.6	1,753.3	2,204.9	1,794.3
		(867.3)	(763.9)	(759.1)	(754.2)	(712.6)
NN *p*-score	–	1,608.2	1,970.9	1,863.8	1,847.1	1,794.3
		(824.1)	(751.9)	(731.8)	(729.9)	(712.6)
B: Nonexperimental control group (PSID)						
BLOP	2,338.9	–	–	–	–	–
	(868.3)					
NN covariate	–	2,073.5	1,618.7	469.2	−111.6	−15,204.8
		(1,678.6)	(1,544.1)	(1,137.3)	(865.0)	(627.2)
NN *p*-score	–	2,141.7	2,061.8	1,014.4	−182.1	−15,204.8
		(1,555.6)	(1,483.1)	(1,378.2)	(991.0)	(627.2)

The estimator presented is for the ATT. BLOP refers to our proposed estimator, and NN refers to nearest-neighbor matching estimator. For the experimental sample, the optimal number of matches was 2.66 (ranging from 1 to 8), and for the nonexperimental, it was 2.44 (ranging from 1 to 5). Standard errors follow the approach of Abadie and Imbens (2006).

from the nearest-neighbor matching estimator with differing number of matches (from 1 to all). As can be observed, with the experimental data (NSW control group), the estimate is very close to the benchmark (US$1,794) and is comparable to those obtained with nearest neighbor using 16 neighbors. However, for the BLOP estimator, only an average of 2.66 neighbors per observation was needed (ranging from 1 to 8).[8] With the BLOP matching estimator, we do not need to worry about the choice of number of matches.

When analyzing the results for the PSID control group, we can see that the nearest-neighbor estimator performs relatively well with 1 and 4 neighbors but poorly when the number of neighbors increases. The BLOP estimator gives a slightly higher estimate than the nearest neighbor, and it does not explode (in terms of bias) as the nearest neighbor since it chooses optimally the number of neighbors. In this case, for each unit to match, only 2.44 neighbors were needed on average. In figure 2 we present the histograms for the number of units used for the experimental and the PSID sample. In the upper panel we show the histogram for the experimental data. As can be observed, it is highly left-skewed, with a median value equal to 2. The lower panel shows the histogram for the number of units used with the PSID control group. As with the experimental data, the histogram is left-skewed but less so. Additionally, the median value of the number of units used in this case is equal to 2 as well.

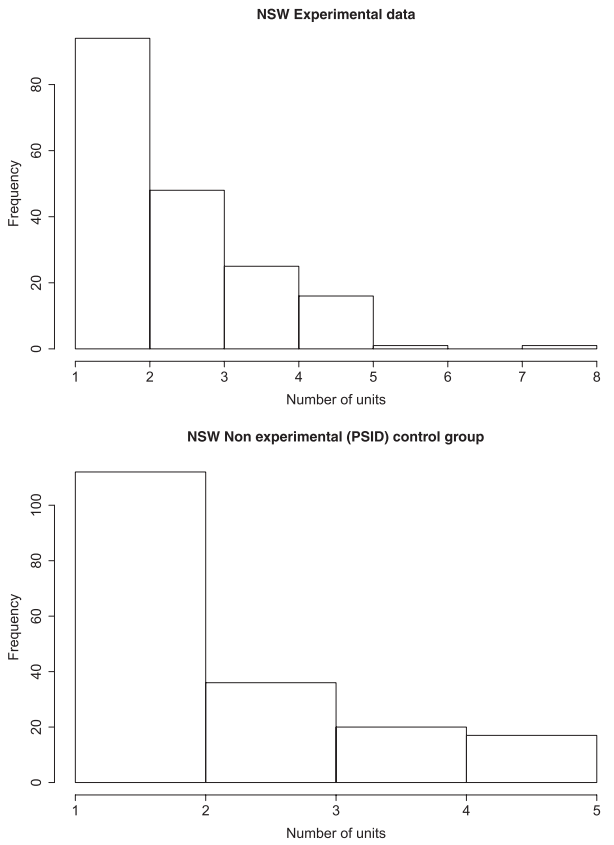
As we mentioned before, the nonexperimental data from the PSID are quite different from the experimental data control group. Thus, it is interesting to analyze the postmatching balance of our estimator. To do so, we compare the sample mean and Kolmogorov-Smirnov distances among the treated units and their counterfactuals, constructed with the units chosen for each match and the computed optimal weights. Table 3 presents the results of the postmatching balance. The BLOP estimator is able to balance the nine covariates, for sample means (failing to reject the null of means equality at 1%). For the continuous covariates we perform a Kolmogorov-Smirnov test and compute the *p*-values implementing bootstrapping with 500 repetitions. For three of the four continuous covariates, we cannot reject the null of equality of treated units distribution and their counterfactuals.

B. *Empirical Monte Carlo*

In this section we implement an empirical Monte Carlo experiment to evaluate the performance of the BLOP matching estimator for the ATT. We compare its performance in terms of bias, variance, and postmatching covariate balance

[8] In each match, a neighbor j is considered "used" if $\lambda_j > 0$, which translates to $\lambda_j > 1e-15$ in the numeric implementation.

FIGURE 2.—FREQUENCIES OF THE NUMBER OF NEIGHBORS USED FOR EACH UNIT MATCHED



to the nearest-neighbor matching estimator (on covariates and propensity score) and the normalized inverse probability weighting estimators (IPW).[9]

The empirical Monte Carlo design is taken from Busso et al. (2014). They focus on the African American subsample in the experimental group (156 individuals); the control group is taken from the PSID (624). The covariates considered are similar to those in section IIIA (age, education, marital status, earnings in 1974 and 1975, and unemployment in 1974 and 1975), plus a dummy for high school dropouts. Also, interactions between the 1974 and 1975 unemployment dummies and between 1974 and 1975 earnings are included. Finally, earnings in 1974 and 1975 squared complete the full set of covariates. Let X_i be the list of covariates (excluding squared terms and interactions) and Z_i the full set of covariates (including squared terms and interactions).

The data generation process (DGP) is explained in detail in Busso et al. (2014) but we touch on it briefly. The main framework is given by the following equations,

$$Y_i(0) = m(Z_i) + \sigma\epsilon, \tag{10}$$

$$T_i^* = \alpha + \beta Z_i - U_i, \tag{11}$$

where Z_i is a function of the covariates (described previously) and U_i follows a standard logistic distribution. Each sample

[9] See Imbens (2004) for a discussion of this estimator.

TABLE 3.—POSTMATCHING BALANCE: NSW DATA WITH PSID CONTROLS

Variable	Mean Treated (1)	Mean Control (2)	t-Statistic p-value (3)	KS p-value (4)
Age	25.82	27.06	0.092	0.004
Education	10.35	10.33	0.934	0.800
Black	0.84	0.84	1.000	–
Hispanic	0.06	0.06	1.000	–
Married	0.19	0.19	0.995	–
Earnings '74	2,095.57	2,211.63	0.818	0.590
Earnings '75	1,532.06	2,563.57	0.013	0.062
Unemployed '74	0.71	0.71	0.968	–
Unemployed '75	0.60	0.60	0.953	–

Postmatching balance is evaluated by comparing treated units and their counterfactuals built with units from the control group and their optimal weights from the BLOP. Column 1 shows the sample mean of covariates of treated units from the NSW experimental sample. Column 2 is the sample mean of covariates of counterfactual values from the PSID sample using optimal weights. Column 3 shows the p-values of the difference in means test between columns 1 and 2. Column 4 shows the p-values for the Kolmogorov-Smirnov test for equality of distribution between treated units and their counterfactual values for the nonbinary covariates.

is constructed such that covariates (X_i) are drawn from a population model (that uses first and second moments from the original sample). Then U_i are drawn from a standard logistic distribution and T^* is generated using equation (11), where we use, instead of α and β, the coefficients from a logistic regression estimated on the original NSW sample. Hence, a latent treatment variable T^* is generated using equation (11). Using a linear function for $m(\cdot)$ and sampling ϵ from a standard normal distribution (independent of X_i), we use equation (10) to generate $Y_i(0) = \delta_0' Z_i + \epsilon_{0i}$. Instead of δ_0, we use the coefficients from a regression of $Y_i(0)$ on Z_i using control observations in the NSW sample. The root mean squared error of the regression is assigned to σ_0^2. $Y_i(1)$ is constructed analogously, regressing $Y_i(1)$ on Z_i using the treated units from the NSW sample. Finally, we construct $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

We draw 5,000 random samples of size $N = 400$. The population treatment effect on the treated for this benchmark case is \$2,334. Since the benchmark case has poor or bad overlap, Busso et al. (2014) also consider a case with good overlap to evaluate the performance of different estimators under the two cases.[10]

In table 4, we present the estimates of different matching estimators to assess the BLOP's performance in comparison to commonly used estimators such as nearest neighbor in characteristics, nearest neighbor in propensity score, and the normalized inverse probability weighting (IPW) estimator (Hirano, Imbens, & Ridder 2003). In this exercise, the propensity score is correctly specified so nearest-neighbor matching estimators on the propensity score and IPW are well specified. As can be seen, for the design with poor overlap (panel A) the BLOP estimator (on characteristics) performs well in terms of bias when compared to the nearest-neighbor estimator in characteristics. As noted, when the number of neighbors increases, the bias increases by a factor of 11 and its variance decreases to a third, a remark on the importance

[10] In this case, the parameters of the selection to treatment equation are divided by 5, so the random component in the assignment to treatment (U_i) is relatively more important, implying a better overlap of the data.

TABLE 4.—EMPIRICAL MONTE CARLO RESULTS

	BLOP Covariates	BLOP p-Score	NN Covariates $k=1$	NN Covariates $k=4$	NN Covariates $k=16$	NN p-Score $k=1$	NN p-Score $k=4$	NN p-Score $k=16$	IPW
A: Bad overlap									
Absolute Bias × 1,000	97.89	28.84	116.11	353.11	1,369.62	25.15	213.79	1,194.95	147.70
Variance × n	3,379.62	5,139.79	3,442.77	1,908.3	1,014.64	5,341.96	3,090.08	1,686.45	3,610.69
B: Good overlap									
Absolute Bias × 1,000	10.69	34.56	71.19	24.10	233.15	27.43	2.16	51.62	13.82
Variance × n	795.36	990.36	859.22	678.79	633.60	1,039.88	787.57	720.98	829.45

The data-generating process follows Busso et al. (2014) with $n = 400$. BLOP Covariates correspond to our estimator performing the matching on characteristics and BLOP p-score to our estimator doing the matching on the estimated propensity score from a correctly specified model. NN Covariates corresponds to the k-nearest-neighbor matching estimator on characteristics and NN p-score to the k-nearest-neighbor matching estimator on the propensity score. IPW corresponds to the normalized inverse probability weighting estimator. Bad overlap corresponds to the NSW DGP that mimics the overlap of the original NSW sample. Good overlap, in panel B, corresponds to the NSW DGP in which the coefficients of the selection equation are divided by 5.

of the choice of the number of matches.[11] One plausible explanation for the dramatic increase in the bias is related to the overlap being poor. Thus, increasing the number of neighbors necessarily implies including units far away from the units that need to be matched. When compared to the IPW, the BLOP matching estimator in covariates performs slightly better than the IPW in terms of bias and variance. We also compute the BLOP estimator on the (correctly specified) propensity score. As shown in table 4, the BLOP matching estimator on the propensity score performs similar to the IPW in terms of bias and variance. Nearest-neighbor matching estimators on the propensity score perform well with few neighbors (between 1 and 4), but when they increase to 16, the performance is poor in terms of bias.

When analyzing the results with good overlap (see table 4, panel B) the results are qualitatively similar to those with poor overlap. The BLOP matching estimator in covariates and the IPW performs similarly well in terms of bias and variance. Nearest-neighbor matching estimator on covariates also performs well with four neighbors in terms of bias that increases by a factor of 10 with sixteen neighbors. Nearest-neighbor matching on propensity score achieves its best performance in term of bias with four neighbors, which increases by a factor of 24 with sixteen neighbors. Thus, the BLOP matching estimator has excellent performance that solves an important issue: it determines the number of neighbors used (by finding the weights that optimize covariate balance), an open question for nearest-neighbor estimators.

Next, we check the postmatching balance performance of the estimators analyzed in table 4. In table 5 we show the p-values of the difference in means test for all the covariates between the means of treated individuals and means of their counterfactual values in the control group, using the weighting scheme of each matching estimator. The BLOP matching estimator has the higher average p-value in both designs (panels A and B). In general, matching estimators on the (correctly specified) propensity score and IPW perform well. However, when overlap is poor, nearest-neighbor matching estimator on covariates performs poorly relative to the BLOP or IPW estimators when the number of neighbors increases (both designs).

[11] See Härdle (1990) for details about the trade-off bias-variance in nearest-neighbor estimators.

Hence, the empirical Monte Carlo evidence suggests that when the propensity score is correctly specified, matching on propensity score with few neighbors (between 1 and 4) and reweighting (IPW) performs well in terms of bias, variance, and balance. The BLOP matching estimator also performs well without needing to estimate the propensity score and without the need of arbitrarily fixing the number of neighbors. The next section analyzes the case of misspecification in the selection equation and, hence, the propensity score estimation.

IV. Finite Sample Properties and Misspecification

In this section we implement Monte Carlo simulations with different designs taken from Busso et al. (2014) and evaluate the BLOP's performance comparing it to other matching estimators in cases of correct and misspecification of the outcome and selection equations.

Design 1 considers a linear model for both the outcome and the selection equation. The number of observations is $N = 200$, and there are four covariates (X_1, X_2, X_3, X_4) normally distributed with zero mean and a block diagonal matrix given by Σ.[12] This structure permits correlation between only X_1 and X_2 and only X_3 and X_4. The main framework follows equations (10) and (11). As in the empirical Monte Carlo of section III, Z_i is a function of the covariates specified below and U_i follows a standard logistic distribution. Hence, a latent treatment variable T^* is generated using equation (11). Using a linear function for $m(\cdot)$ and sampling ϵ from a standard normal distribution (independent of X_i) and setting $\sigma = 1$, we use equation (10) to generate $Y_i(0)$. We assume a constant treatment effect equal to 1 to generate $Y_i(1) = T_i + Y_i(0)$ where $T = \mathbb{1}(T^* > 0)$ where $\mathbb{1}(\cdot)$ is an indicator function that is equal to 1 when the argument is true and 0 otherwise.

Design 2 considers a linear model with interactions for both the outcome and the selection equation. Hence, Z_i includes the four linear terms (X_1, X_2, X_3, X_4) plus the six interactions (X_1X_2, X_1X_3, and so on).

For both designs, we consider cases of correct and misspecification for the propensity score. Table 6 shows the results of this Monte Carlo exercise after 5,000 replications. Focusing

[12] The lower right and upper left blocks of Σ are given by $\frac{1}{3}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$.

TABLE 5.—POSTMATCHING BALANCE EMPIRICAL MONTE CARLO

Variable	BLOP Covariates	BLOP p-Score	NN Covariates $k=1$	NN Covariates $k=4$	NN Covariates $k=16$	NN p-Score $k=1$	NN p-Score $k=4$	NN p-Score $k=16$	IPW
A: Bad overlap									
Age	0.22	0.32	0.11	0.04	0.00	0.30	0.28	0.20	0.24
Education	0.43	0.24	0.24	0.28	0.20	0.25	0.31	0.43	0.27
Dropout	0.82	0.23	0.51	0.24	0.40	0.23	0.30	0.42	0.25
Married	0.74	0.45	0.32	0.13	0.00	0.39	0.39	0.14	0.34
Unemployed '74	0.47	0.49	0.07	0.01	0.00	0.42	0.45	0.18	0.31
Unemployed '75	0.99	0.30	0.41	0.28	0.02	0.44	0.46	0.14	0.26
Earnings '74	0.62	0.52	0.49	0.04	0.00	0.38	0.34	0.04	0.39
Earnings '75	0.33	0.54	0.99	0.86	0.18	0.26	0.31	0.12	0.39
Average p-value	0.59	0.39	0.39	0.23	0.10	0.33	0.36	0.21	0.31
Min p-value	0.22	0.23	0.07	0.01	0.00	0.23	0.28	0.04	0.24
B: Good overlap									
Age	0.78	0.54	0.44	0.49	0.43	0.48	0.56	0.57	0.57
Education	0.79	0.50	0.44	0.50	0.48	0.50	0.57	0.60	0.58
Dropout	0.95	0.49	0.64	0.48	0.33	0.48	0.57	0.58	0.57
Married	0.92	0.58	0.50	0.50	0.38	0.48	0.57	0.59	0.57
Unemployed '74	0.89	0.63	0.42	0.40	0.16	0.51	0.60	0.63	0.55
Unemployed '75	0.95	0.57	0.44	0.53	0.41	0.51	0.59	0.62	0.55
Earnings '74	0.53	0.66	0.90	0.34	0.04	0.45	0.53	0.54	0.59
Earnings '75	0.91	0.69	0.93	0.49	0.12	0.44	0.51	0.52	0.59
Average p-value	0.89	0.58	0.59	0.46	0.29	0.48	0.56	0.58	0.57
Min p-value	0.78	0.49	0.42	0.34	0.04	0.44	0.51	0.52	0.55

P-values of the difference in means test between means of treated individuals and means of their counterfactual values in the control group, using the weighting scheme of each matching estimators. BLOP Covariates perform the matching on characteristics and BLOP p-score on the estimated propensity score from a correctly specified model. NN Covariates corresponds to the k-nearest-neighbor matching estimator on characteristics and NN p-score to the one on the propensity score. IPW corresponds to the normalized inverse probability weighting estimator. Bad overlap corresponds to the NSW DGP that mimics the overlap of the original NSW sample. Good overlap in panel B corresponds to the NSW DGP in which the coefficients of the selection equation are divided by 5.

on absolute bias and empirical variance of the estimators, we see that the BLOP in characteristics performs well in terms of absolute bias compared to the nearest-neighbor matching estimator (NN) in design 1 (table 6, columns 1 to 3). The BLOP estimator based on the correctly specified propensity score (linear, column 1) is very competitive with nearest-neighbor estimates on the propensity score and IPW. When the propensity score is misspecified (interactions only, column 2), the performance of the BLOP on characteristics is outstanding relative to the other estimators. When overspecified (linear plus interactions, column 3), the comparison is similar to the case in which the propensity score is correctly specified.

In design 2 (table 6, columns 4 to 6), the true specification of the propensity score (and $m(Z_i)$) is a linear plus interaction equation (column 6). When the propensity score is underspecified (linear, column 4), the nearest-neighbor matching estimator (on covariates) performs well with 1 and 4 neighbors. With 16 neighbors, nearest-neighbor matching on covariates performs similar to the BLOP, but with 64 neighbors, its bias greatly increases. Estimates based on the propensity score perform poorly in terms of bias with 1 and 4 neighbors but improve significantly with 64 neighbors.

The second case of underspecification is when the selection equation considers only interactions but no linear terms (interactions, column 5 in table 6) and the results are similar to the previous case. When the selection equation is correctly specified (linear plus interactions, column 6 in table 6), estimates based on the propensity score perform very well. Specifically nearest neighbor on propensity score with 1 and 4 neighbors, IPW and the BLOP on the propensity score are very close in absolute bias and variance. Thus, under misspecification, the BLOP (on covariates) performs very well in comparison to other estimators based on the propensity score, especially when this is underspecified.

V. Conclusion

The main advantages of the BLOP matching estimator are that it directly determines the weights and the matches used while optimizing the postmatching covariate balance. As stated in previous research, there is little known about the optimal number of matches or about data-dependent ways of finding it (Imbens & Wooldridge, 2009). Additionally, there is no clear way to measure covariate balance (Diamond & Sekhon, 2013). In this paper we contribute to the matching literature by linking the choice of matches and weights to the improvement of postmatching covariate balance.

The method we develop is not an algorithm that iteratively checks covariate balance until convergence. Instead, it is an optimization problem that incorporates an individual covariate balancing criterion in the objective function that determines the weights used in each match. It can be written as a linear program that allows us to use standard optimization techniques to solve the problem quickly. We provide an R package called blopmatching that implements the proposed estimator.

The empirical analysis shows that the BLOP matching estimator provides an outstanding postmatching balance and performs well in terms of bias and variance when compared to nearest-neighbor matching estimators (for both covariates and propensity score) and the normalized inverse probability weighting estimator. Major improvements are observed when there is underspecification of the selection equation

TABLE 6.—MONTE CARLO EVIDENCE, MISSPECIFIED MODELS

Variable	Design 1			Design 2		
	Linear (1)	Interactions (2)	Linear + Interactions (3)	Linear (4)	Interactions (5)	Linear + Interactions (6)
A: Absolute Bias × 1,000						
BLOP covariates	110.6	110.6	110.6	86.9	86.9	86.9
BLOP *p*-score	6.1	595.6	11.9	101.8	175.5	1.7
NN covariates						
$k = 1$	183.5	183.5	183.5	26.7	26.7	26.7
$k = 4$	274.7	274.7	274.7	45.4	45.4	45.4
$k = 16$	428.5	428.5	428.5	97.9	97.9	97.9
$k = 64$	560.7	560.7	560.7	138.0	138.0	138.0
NN *p*-score						
$k = 1$	6.6	595.8	12.6	101.3	174.3	2.0
$k = 4$	26.6	592.6	31.5	101.3	169.5	6.6
$k = 16$	107.1	584.1	108.4	91.7	156.7	29.9
$k = 64$	389.4	571.6	395.6	34.2	122.9	83.1
IPW	3.4	595.3	10.0	108.7	177.9	3.3
B: Variance ×n						
BLOP covariates	8.3	8.3	8.3	7.1	7.1	7.1
BLOP *p*-score	9.6	11.2	10.5	8.2	8.3	9.2
NN covariates						
$k = 1$	7.9	7.9	7.9	7.1	7.1	7.1
$k = 4$	6.1	6.1	6.1	5.3	5.3	5.3
$k = 16$	6.0	6.0	6.0	5.1	5.1	5.1
$k = 64$	6.6	6.6	6.6	5.4	5.4	5.4
NN *p*-score						
$k = 1$	10.2	12.1	11.2	8.9	9.1	9.9
$k = 4$	7.1	8.1	7.6	6.1	6.2	6.6
$k = 16$	5.9	7.2	6.4	5.3	5.6	5.5
$k = 64$	6.0	7.1	6.5	5.4	5.6	5.3
IPW	7.1	7.2	7.8	5.2	5.6	5.8

The data-generating process follows Busso et al. (2014) with $n = 200$ and 5,000 repetitions. Design 1 corresponds to a linear equation for the outcome equation and linear equation for the selection equation. Design 2 corresponds to a linear plus interaction equation for the outcome and selection equations. Thus, in design 1, column 1 corresponds to a correctly specified model for the propensity score, column 2 to a misspecified model, and column 3 to an overspecified model for the propensity score. For design 2, columns 4 and 5 correspond to a missspecified model for the propensity score and column 6 to a correctly specified model for the propensity score.

for estimating the propensity score. Hence, our method gives researchers a new alternative matching estimator that prevents the selection of an arbitrary number of neighbors or the estimation of the propensity score.

REFERENCES

Abadie, A., and G. W. Imbens, "Large Sample Properties of Matching Estimator for Average Treatment Effect," *Econometrica* 74 (2006), 235–267.

——— "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics* 29 (2011), 1–11.

Botkin, N., and J. Stoer, "Minimization of Convex Functions on the Convex Hull of a Point Set," *Math. Meth. Oper. Res.* 62 (2005), 167–185.

Busso, M., J. DiNardo, and J. McCrary, "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," this REVIEW 96 (2014), 885–897.

Colson, B., P. Marcotte, and G. Savard, "An Overview of Bilevel Optimization," *Annals of Operations Research* 153 (2007), 235–256.

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

Diamond, A., and J. S. Sekhon, "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," this REVIEW 95 (2013), 932–945.

Frölich, M., "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," this REVIEW 86 (2004), 77–90.

Galdo, J. C., J. Smith, and D. Black, "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data," *Annals of Economics and Statistics/Annales d'économie et de statistique* 91/92 (2008), 189–216.

Graham, B. S., C. C. D. X. Pinto, and D. Egel, "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies* 79 (2012), 1053–1079.

Härdle, W., *Smoothing Techniques: With Implementation in S* (New York: Springer-Verlag, 1990).

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

Imai, K., and M. Ratkovic, "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 76 (2014), 243–263.

Imbens, G. W., "Nonparametric Estimation of Average Treatment Effects Under Exogenety: A Review," this REVIEW 86 (2004), 4–29.

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

Lalonde, R. J., "Evaluating Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76 (1986), 604–620.

Robins, J. M., A. Rotnitzky, and L. P. Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89 (1994), 846–866.

——— "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association* 90 (1995), 106–121.

Rockafellar, R., *Convex Analysis* (Princeton, NJ: Princeton University Press, 1972).

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

Smith, J. A., and P. E. Todd, "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (2005), 305–353.