Nonparametric Estimation of Causal Effects in Observational Studies

Ricardo Paes de Barros**

Abstract

The paper develops a statistical procedure to provide consistent estimators for the average impact of an intervention or treatment (e.g., earnings) among subjects of a target population (e.g., young high school dropouts). The procedures we study belong to a class of estimators which can be expressed as a DIFFERENCE between the average outcome in the treated sample and an adequately chosen weighted average of the outcomes in the control group. We refer to estimators in this class as $D$-estimators. We show that $D$-estimators should only be used in circumstances in which subjects are randomly assigned to treatment given the vector of observed covariates. This condition has been referred to as the “strongly ignorable assumption” or “selection on observable model”. The consistency of $D$-estimators is proved under very weak restrictions on the distribution of the covariates.

Keywords: Average Treatment Effect, Matching Estimators.

JEL Codes: C14, C21.
1. Introduction

The objective of this paper is to develop a statistical procedure to provide consistent estimators for the average impact of an intervention or treatment (e.g., earnings) among subjects belonging to a target population (e.g., young high school dropouts).

The statistical technique to be investigated can be best viewed as an application of recent developments in both nonparametric regression (e.g., Stone (1977) and Devroye and Wagner (1980) and matching (e.g., Rosenbaum and Rubin (1983). Consistency is proved for an estimator which appropriately generalizes and unifies the matching and the nonparametric regression procedures as used to control for covariates in observational studies. Similarly to a recent work on nonparametric regression, but contrary to a previous work on matching, only weak conditions are imposed on the distribution of the covariates throughout this study. In fact, these conditions are required primarily to assure the whole problem is well-defined.

The asymptotic properties of several commonly used matching procedures such as caliper, nearest neighbor, and stratified matching have been largely ignored in the literature. The emphasis has been on bias reduction properties. In particular, to the best of our knowledge, no previous proof of consistency is available for matching procedures involving non-normal and continuous covariates. The consistency of such procedures is established in this article. For a general description of matching and related procedures, see Anderson et al. (1980, Ch. 6), Cochran (1965, Ch. 5), and Cochran and Rubin (1973).

In order to determine whether a treatment is effective in improving an outcome, information about a treated and a control population is routinely available. We assume that a random sample is drawn from each population. For each sampled subject, two attributes are measured: an outcome or response (e.g., earnings) and a vector of covariates or endowments (e.g., age and education). Subjects belonging to the two populations are supposed to differ in at least one important respect: while all treated subjects have been submitted to a specific intervention or treatment (e.g., a training program), all control subjects have not. In addition, the populations may also differ with respect to the distribution of the covariates. Finally, we note parenthetically that for all treated subjects the outcome is measured after treatment.

Let $Y, Z,$ and $T$ denote OUTCOME, COVARIATES and TREATMENT EFFECTS, respectively. Moreover, let $c, r, a,$ and $o$ denote the control, treated, target, and overall populations, respectively. So, for example, $T_a$ denotes the TREATMENT IMPACT on outcome for subjects in the TARGET population. By the same token, $Y_r$ denotes OUTCOME for subjects in the TREATED population. Next, let $i$ and $j$ be the index subjects sampled from the treated and control populations, respectively. Consequently, $\{(Y_{ri}, Z_{ri}) : 1 \leq i \leq n\}$ and $\{(Y_{sj}, Z_{sj}) : 1 \leq j \leq m\}$ denote the available sample information when $n$ and $m$ are the corresponding sample sizes.
Our objective can then be more precisely stated as to investigate how the information available from the treated and control samples, \([Y_{ri}, Z_{ri}) : 1 \leq i \leq n]\) and \([Y_{cj}, Z_{cj}) : 1 \leq j \leq m]\), can be used to construct consistent nonparametric estimators for \(E[T_a]\), the average impact of a treatment on outcome among subjects belonging to a given target population.

The estimators considered in this paper can be expressed as follows:

\[
D_{nm} = \left( \frac{1}{n} \right) \sum_{i=1}^{n} Y_{ri} - \sum_{j=1}^{m} P_{nm}(i,j)Y_{cj}
\]  

(1)

for suitable choices of \(W_{nm} = (P_{nm}(i,j) : 1 \leq i \leq n, 1 \leq j \leq m]\). Throughout the paper \(W_{nm}\) is assumed to be a weighting system in the sense that for given sample sizes \(n, m \geq l\)

\[
\sum_{j=1}^{m} P_{nm}(i,j) = 1 \text{ for } 1 \leq i \leq n
\]

and

\[
P_{nm}(i,j) \geq 0 \text{ for } 1 \leq i \leq n \text{ and } 1 \leq j \leq m
\]

More importantly, however, \(W_{nm}\) is not intended to be a collection of sequences of constants chosen independently of the sample information. In fact, these weights are considered to be random variables. They are, however, required to satisfy two conditions:

(1) \(P_{nm}(i,j)\) has to be conditionally independent of \([Y_{cj} : l \leq j \leq m]\) given the remaining sample information, \([Y_{ri}, Z_{ri}) : 1 \leq i \leq n]\) and \([Z_{cj} : 1 \leq j \leq m]\);

(2) for each \(n, m \geq 1\) and \(1 \leq j \leq m\), \([P_{nm}(i,j) : l \leq i \leq n]\) has to be a sequence of \(n\) identically distributed random variables.

In this article, estimators which can be expressed in this form are referred to as \(D\)-estimators. In Sections 4 and 5, we show how estimators which are based on matching or regression and used in observational studies to control for covariates can be generally expressed as \(D\)-estimators. To obtain an alternative representation for \(D\)-estimators, let \(q_{n,m}(j)\) denote the total weight attached to the outcome of the control subject \(j\). Then \(D_{nm}\) can be expressed as

\[
D_{nm} = \left( \frac{1}{n} \right) \sum_{i=1}^{n} Y_{ri} - \sum_{j=1}^{m} q_{n,m}(j)Y_{cj}
\]

where \(q_{n,m}(j) = \left( \frac{1}{n} \right) \sum_{i=1}^{n} P_{nm}(i,j)\).

For simplicity, the indices \(n\) and \(m\) are omitted everywhere up to Section 6. In Sections 6 and 7, where the asymptotic behavior of \(D\)-estimators is investigated, these indices are then reincorporated.
This article provides general sufficient conditions for $D$-estimators to converge in probability to $E[T_a]$ when both $n$ and $m$ get large. Nevertheless, the entire argument is clearly constituted of two parts. In the first segment, on identification, we investigate sufficient conditions for

$$E[T_a] = E[Y_r] - E[F_c(Z_r)],$$

(2)

where $F_c$ denotes the regression function for the CONTROL population, i.e.,

$$F_c(z) = E[Y_c|Z_c = z].$$

Henceforth, whenever equation (2) is satisfied we say that $E[T_a]$ is identified.

In the second segment, on consistency, we describe conditions which suffice for $D$-estimators to converge in probability to $E[Y_r] - E[F_c(Z_r)]$. It is interesting to note the relative independence of the conditions investigated in the two segments of the study.

In Section 2, we investigate conditions for identification. These conditions provide a transparent rationale for the use of $D$-estimators in observational studies. Furthermore, they also delineate the range of problems or studies to which $D$-estimators are a valid estimation procedure.

In Section 3, we review the concept of balancing score introduced by Rosenbaum and Rubin (1983). Then we state and prove two important properties common to balancing scores. These properties are then used in the subsequent sections to show how a balancing score can be used to reduce the computational costs involved in matching and nonparametric regression analyses.

In Sections 4 and 5, we review the matching and the nonparametric regression procedures, respectively, as used to control for covariates in observational studies. We also show how both procedures can generally be expressed as $D$-estimators.

In Section 6, we describe the underlying probability model and formally define the concepts of weighting systems and $D$-estimators. Moreover, those assumptions required to assure that the whole problem is well-defined are precisely stated.

In Section 7, consistency is proved under very weak conditions on the weighting system. As far as conditions for consistency are concerned, there is a trade-off between imposing conditions on the weighting system and imposing conditions on the distribution of the covariates. The trade-off is similar to that found in the theory of nonparametric regression (e.g., Stone, 1977). Due to this similarity, the general structure of our proofs resembles that used by Stone (1977) and Devroye and Wagner (1980). Finally, in Section 8, some final remarks and suggestions for future research are presented.

2. Identification

In a simple controlled experiment, subjects from a target population are randomly assigned to control or treatment. As a consequence, the outcome distri-
bution in the control subpopulation will be identical to the outcome distribution that would prevail in the treated subpopulation had treated subjects not been submitted to treatment. In other words, the random assignment assures that $Y_c$ and $Y_r - T_r$ are identically distributed. So, in particular,

$$E[T_r] = E[Y_r] - E[Y_c]$$  \(4\)

Clearly, the randomization has additional implications. It is of particular interest to us that also $Z_c$ and $Z_r$ as well as $T_r$ and $T_a$ are identically distributed. So

$$E[T_a] = E[T_r]$$  \(5\)

and

$$E[F_c(Z_r)] = E[F_c(Z_c)]$$  \(6\)

Moreover, by definition, equation (3),

$$F_c(Z_c) = E[Y_c | Z_c]$$  \(7\)

So, it follows from (6) and (7) that

$$E[Y_c] = E[F_c(Z_r)]$$

which together with (4) and (5) imply (2). In summary, in a controlled experiment

$$E[T_a] = E[Y_r] - E[F_c(Z_r)]$$

and

$$E[T_a] = E[Y_r] - E[Y_c]$$  \(8\)

Consequently, $E[T_a]$ is identified and can be easily estimated. In fact, the sample difference in average outcomes between treated and control subjects provides an unbiased and consistent estimator. Let this estimator be denoted by $M$. It can be expressed algebraically as

$$M = \left( \frac{1}{n} \right) \sum_{i=1}^{n} Y_{ri} - \left( \frac{1}{m} \right) \sum_{j=1}^{m} Y_{cj}$$

Hence $M$ is a $D$-estimator with $q = t/m$.

We are, however, more interested in designing techniques which permit valid statistical inferences to be made in circumstances in which the choice of the control and treated populations are, at least to some extent, outside the analyst’s control. Such studies are called quasi-experiments or comparative observational studies, as proposed by Cochran (1965).
In observational studies, the development of valid techniques relies, as we would expect, upon some prior knowledge about the relationship between the outcome distributions in the control and treated groups. Since the outcome distribution in the control group is to be used as a standard of comparison, this distribution must be related, in some appropriate sense, to the outcome distribution that would prevail in the treated group had treated subjects not been submitted to treatment.

In this study, to model this relationship, we assume that, in the ABSENCE of treatment, all control and treated subjects with EQUAL value for the covariates would have the same expected outcome independent of the group to which they belong. Nevertheless, the distributions of the covariates in the treated and control populations are not themselves constrained to be similar in any respect. In summary, it is just assumed that discounting for the treatment effect, Tr, the regression function for the treated group should be identical to the regression function for the control group, i.e.,

\[ \text{Assumption 1: } E[Y_r - T_r | Z_r = z] = E[Y_c | Z_c = z] \text{ a.s.}(P_{zr}) \]

where \( P_{zr} \) denotes the probability distribution of \( Z_r \). To assure that A1 is well-defined, two assumptions are required:

1. the first moments for \( Y_c \) and \( Y_r - T_r \) must exist and be finite;
2. the probability measure induced by \( Z_r \) must dominate the probability measure induced by \( Z_c \). See Section 6 for a detailed discussion of these conditions.

The entire analysis in this paper relies on Assumption 1. Even more importantly, this assumption plays a paramount role in assuring the validity of any matching procedure. A matching procedure is said to be valid if it is able to produce unbiased estimators when perfect matches are available for all treated subjects (see Rosenbaum and Rubin, 1983, 1985).

The assumptions used in observational studies vary considerably. Usually, they are stronger than A1. Even so, A1 is rarely explicitly mentioned in such studies, especially those done by econometricians. It is also true, however, that in some of these studies A1 is not indeed required. Let us then consider how A1 is related to those assumptions commonly found in the econometric literature. Initially, it is important to realize that econometricians have a strong preference for structural models. In the present context, this means that instead of assumptions like A1 they prefer to model how the control and treated subpopulations end up being selected from the overall population. The necessary identification is then obtained by imposing suitable restrictions on the selection or assignment process (see Heckman and Macurdy (1986) for a thorough discussion of this literature and of its applications to labor economics).

As described by Barnow et al. (1980, Sec. 4), these selection models can be organized into three classes (see also Heckman and Robb, 1985a,b) for a more detailed description of these models as well as for several extensions). In the
first class, which is discussed in detail by Cain (1975), a subject’s probability of being assigned to treatment is assumed to depend only on the value of the covariates. Conditional on the covariates, the assignment is assumed to be random. In this case, Heckman and Robb (1985a, p. 190) would say that there is “selection only on observables”; Rosenbaum and Rubin (1983) would say that assignment to treatment is “strongly ignorable” given the observed covariates. Both the selection on observables and the strongly ignorable assignment hypotheses imply A1. Indeed, the techniques developed in this chapter are designed to be applied precisely to this class of selection models; they are inappropriate when applied to models belonging to other classes.

To analyze formally the direct link between this class of selection models and A1, recall that \( Y_o \) and \( Z_o \) denote, respectively, the outcome and the covariates of subjects in the overall population. Assume \( Y_o \) is measured before any treatment. Moreover, let \( d \) indicate the assignment of a subject to treatment. Therefore,

\[
E[Y_o|d = 1, Z_o = z] = E[Y_r - T_r|Z_r = z]
\] (9)

and

\[
E[Y_o|d = 0, Z_o = z] = E[Y_c|Z_o = z]
\] (10)

The defining characteristic of selection models in this class is, as already mentioned, the dependence of the assignment probability only on \( Z_o \). So, in particular,

\[
P[d = 1|Y_o, Z_o] = P[d = 1|Z_o]
\] (11)

It can be easily shown that equation (11) implies the following assumption, commonly used to describe this kind of selection models.

(A1*) **Assumption 1**: \( E[Y_o|d, Z_o] = E[Y_o|Z_o] \)

Finally, notice that because of (9) and (10), A1* is equivalent to A1.

Now consider the second class of selection models. The models in this class rely on the existence of an observed variable which affects the selection process, but is NOT directly related to the outcome or the treatment impact. An instrumental variable procedure can then be used to estimate the average treatment impact consistently. This class of models has, however, an important drawback, which is not always appreciated in the literature. To understand the nature of this problem, note that in our sample the proportion of treated subjects is \( n/(n + m) \). Nothing in the sample design guarantees that this proportion is an unbiased or consistent estimator of the proportion of treated subjects in the overall population. In fact, treated subjects generally tend to be oversampled, generating what is called a choice-based sample. In dealing with choice-based samples some knowledge about the proportion of treated subjects in the overall population is required for
instrumental variables methods to produce consistent estimators (see Heckman and Robb, 1985a, p. 198). So, for those evaluation studies for which this additional information is not available, this class of models does not represent a feasible alternative.

The third class of selection models relies on functional form or distributional assumptions in order to obtain consistent estimators. These models, however, are very flexible to the extent that they permit all observed and unobserved variables to affect simultaneously the selection process and the outcome. So, in particular, the existence of selection based on unobservables is permitted.

We emphasize that studies based on selection models belonging to the second and third classes do NOT require A1 to be satisfied. Accordingly, the statistical procedure to be developed should ONLY be applied to studies for which selection models of the first kind are considered adequate.

Consider now the main implications of Assumption 1. First we show that when this assumption is satisfied the average treatment effect among treated subjects, $E[T_r]$, equals the average difference between the regression functions for treated and control subjects. So, in particular, any observed statistically significant difference between these regression functions should be indicative of some treatment effect. To prove this important fact (see equation (16) below), let $F'_r$ denote the regression function for treated subjects, i.e.,

$$F'_r(z) = E[Y_r | Z_r = z]$$

Recall that by definition, equation (3),

$$F_c(z) = E[Y_c | Z_c = z]$$

Then, notice that by the iterated expectation formula

$$E[T_r] = E[E[Y_r | Z_r]] - E[E[Y_r - T_r | Z_r]]$$

and by A1 and (13)

$$E[Y_r - T_r | Z_r = z] = F_c(z)$$

Hence, from (12), (14), and (15)

$$E[T_r] = E[F_r(Z_r) - F_c(Z_r)]$$

or

$$E[T_r] = E[Y_r] - E[F_c(Z_r)]$$

These equations, (16) and (17), are very suggestive of how $E[T_r]$ can be estimated. They form the basis of our estimation procedure and, consequently, will be
extensively explored later. Now, however, we recall that our final objective is to estimate $E[T_a]$ rather than $E[T_r]$. In order to obtain the identification for $E[T_a]$, additional prior information is required on how the treated and target populations are related to each other. The analysis that follows will be based on the simplifying assumption that the treated group is representative of the target population, in the sense that

(A2) **Assumption 2:** $E[T_a] = E[T_r]$

Nevertheless, the statistical procedure we study could also be applied under less restrictive assumptions as far as some suitable modifications are made.

From (16), (17), and A2,

$$E[T_a] = E[T_r] = E[F_r(Z_r) - F_c(Z_r)]$$

and

$$E[T_a] = E[Y_r] - E[F_c(Z_r)]$$

So, when A1 and A2 are satisfied, $E[T_a]$ is identified. As we showed previously, $E[T_a]$ is also identified in a controlled experiment. In fact, both A1 and A2 are implied by random assignment. We then find it useful to finish this section by describing what additional flexibility has been gained going from a controlled experiment to a sample design which is only required to satisfy A1-A2.

Quade (1982) defines a covariate as “concomitant” when its distribution is the same in both the control and treated populations. As already mentioned, throughout this article, the covariates are not required to have the same distribution in both groups. In other words, we do not require the covariates to be concomitant. For example, when age and education are the observed covariates, controls are allowed to be on average older and more educated than treated subjects. Thus, the following assumption is NOT required to be satisfied.

(A3) **Assumption 3:** $Z_c$ and $Z_r$ are identically distributed.

In fact, when the covariates are concomitant, A3 holds and A1 is satisfied,

$$E[Y_c] = E[Y_r - T_r]$$

or

$$E[T_r] = E[Y_r] - E[Y_c]$$

Moreover, when A2 is also satisfied, then

$$E[T_a] = E[Y_r] - E[Y_c]$$
and we are essentially back to the controlled experiment case (see equation (8), for example). As in that case, the sample difference between the average outcomes for treated and control subjects still provides an unbiased and consistent estimator for $E[T_a]$. In summary, our sample design differs from that of a controlled experiment exactly to the extent that we do not require the covariates to be concomitant, i.e., for us $A3$ may fail. We emphasize that once $A3$ is added to $A1-A2$, the sample design becomes essentially equivalent to a controlled experiment.

Whenever $A3$ fails, a direct comparison of average outcomes between treated and control subjects will be misleading. That is the case because the two groups differ not only due to treatment, but also due to differences in the underlying distributions of the covariates. Any direct comparison will confound these two differences – that is precisely why covariates are called confounding variables in this literature. In order to draw valid inferences we need a technique such as regression or matching that can account for the imbalance of covariates between the two groups.

3. Balancing Scores

The computational burden of any procedure used to control for non-concomitant covariates is exponentially related to the number of covariates involved. Therefore, any device able to reduce the dimensionality of the problem would be extremely helpful. The balancing score is one such device.

Let $X: R^a \rightarrow R^b$ be a measurable function and define $X_r = X(Z_r)$ and $X_c = X(Z_c)$. Let $P_{xr}$ and $P_{xc}$ denote the probability distribution of $X_r$ and $X_c$, respectively. Moreover, let $P_{x_r}[.|X_r]$ denote the probability distribution of $Z_r$ conditional on $X_r$; $P_{x_c}[.|X_c]$ is defined similarly. Then $X$ is called a balancing score for the pair $(Z_r, Z_c)$ when for any measurable set $A$ contained in $R^a$

$$P_{x_r}[A|X_r = x] = P_{x_c}[X_c = x] \ a.s.(P_{xr})$$

Examples of balancing scores range from the identity $X(z) = z$ to the propensity score, $E(z)$, introduced by Rosenbaum and Rubin (1983). The propensity score is defined as the probability of assignment to treatment for subjects in the overall population conditional on the value of their covariates. Using the notation introduced previously, see equations (9-11),

$$E(z) = P[d = 1|Z_o = z]$$

Rosenbaum and Rubin prove that $E$ is the coarsest balancing score in the sense that any other balancing score can be expressed as a function of $E$. Moreover, by construction, the propensity score is always a real-value function ($b = 1$), independent of the original dimension of the vector of covariates. As an intermediate example, let $Z = [F, h, w, S]$, where $F, h, w,$ and $S$ denote the property
income, hours worked, hourly wage, and some socioeconomic background variables, respectively. Suppose the probability of assignment to treatment depends only on total income, \( F + h.w \), and on the socioeconomic background variables. Then \( X : Z \to [F + h.w, S] \) would be a balancing score which would reduce the dimensionality of the problem by two (\( b = a - 2 \)).

Balancing scores have two very important properties which indeed follow almost directly from their definition. Such properties assure that in order to control for a vector of non-concomitant covariates in observational studies, it suffices to control for any of the balancing scores which could be constructed from them. To be useful, however, a balancing score must meet two requirements:

(1) \( X \) has to be a KNOWN function so that \( X_r \) and \( X_c \) can be actually treated as observed covariates;

(2) the dimensionality of the balancing score must be smaller than the dimensionality of the original vector of covariates (\( b < a \)) so that reduction in computational costs can actually be achieved.

We now prove the following properties:

(P1) **Property 1:** If \( P_{xr} = P_{xc} \) then \( P_{zr} = P_{zc} \).

(P2) **Property 2:** If \( E[Y_r - T_r|Z_r = z] = E[Y_c|Z_c = z] \) a.s. \( (P_{zr}) \) then

\[
E[Y_r - T_r|X_r = x] = E[Y_c|X_c = x] \text{ a.s.}(P_{xr})
\]

Consider Property 1. It states that if the balancing score is concomitant, so must be the covariates. The proof follows almost directly from the definition. Let \( A \) be any measurable set contained in \( R^a \), then by the iterated expectation formula, equation (20), and the hypothesis \( (P_{xr} = P_{xc}) \)

\[
P_{zr}(A) = \int P_{zr}[A|X_r = x].P_{xr}(dx)
\]

\[
= \int P_{zc}[A|X_c = x].P_{xc}(dx) = P_{zc}(A)
\]

So, Property 1 holds. Consider now Property 2. It states that if A1 is satisfied with respect to \( (Z_r, Z_c) \), then it also has to be satisfied with respect to any balancing score constructed from \( (Z_r, Z_c) \). To prove P2, we use:

(1) the fact that \( X_r \) and \( X_c \) are functions of \( Z_r \) and \( Z_c \), respectively;

(2) the iterated expectation formula; (3) equation (20); and (4) the hypothesis of the proposition.
\[ E[Y_r - T_r | X_r = x] = \int E[Y_r - T_r | Z_r = z] \cdot P_{zr} [d_z | X_r = x] \quad \text{[by (1) and (2)]} \]

\[ = \int E[Y_c | Z_c = z] \cdot P_{zc} [d_z | X_c = x] \quad \text{[by (3) and (4)]} \]

\[ = E[Y_c | X_c = x] \quad \text{a.s.}(P_{xr}) \quad \text{[by (1) and (2)]} \]

So, Property 2 holds.

Of course, the advantage of a balancing score is only illusory when the function \( X \) is actually unknown to the analyst, as it is usually the case in observational studies. Nevertheless, notice that Rosenbaum and Rubin (1984) consider a two-step procedure in which instead of controlling for a vector of covariates the analyst (1) estimates the propensity score and then (2) controls for the covariates by matching them with the estimated propensity score.

The consistency of such a class of procedures is not investigated in this article. Throughout the subsequent discussion, we assume that \( X \) is a KNOWN balancing score for the pair \((Z_r, Z_c)\).

4. Matching

\( D \)-estimators have been designed precisely to generalize and unify the matching and the nonparametric regression procedures as used to control for non-concomitant covariates in observational studies. These two procedures, however, are based on considerably distinct principles. Therefore, in the present and following sections, we present a convenient review of the basic principles behind each of these traditional procedures. We then show how the matching, \( M \), and the regression, \( R \), estimators can be generally expressed as \( D \)-estimators.

The basic principle behind matching is to convert non-concomitant covariates to concomitant ones. In other words, matching aims to ensure that A3 is satisfied at least asymptotically. To eliminate the original imbalance of covariates, a subsample is selected from the control sample such that in this subsample the distribution of the covariates is as similar as possible to the corresponding distribution in the treated sample. The selected subsample is usually called matched (control) sample. The analysis then proceeds, as in a controlled experiment, by estimating the average treatment effect, \( E[T_{a}] \), using the difference between the average outcomes in the treated and matched samples. Those control subjects not included in the matched sample are simply ignored. Alternative matching procedures differ to the extent that they use different criteria to select the matched sample.
sample. For reviews of this literature, see Anderson et al. (1980, Ch. 6), Cochran (1965, Ch. 5) and Rubin (1984).

Whichever matching procedure is used, as Rosenbaum and Rubin (1983, 1985) have emphasized, a balancing score will always be a very useful tool. In fact, to assure the desired concomitance of some set of covariates, it is not necessary to match them. Instead, we can potentially reduce the dimensionality of the problem by matching with any known balancing score. The veracity of this statement rests upon P1. According to this property, the concomitance of a balancing score implies the concomitance of the corresponding set of covariates. So, since matching with a balancing score ensures its own asymptotic concomitance, it also ensures, as a consequence, the asymptotic concomitance of the corresponding set of covariates.

To investigate the relationship between matching and $D$-estimators, consider the following stylized matching procedure: For each treated subject $i = 1, \ldots, n$

1. choose a neighborhood $C_i$ or $X_{ri}$;
2. identify the set $A_i = \{j : X_{cj}$ is in $C_i\}$ of potential matches;
3. choose an integer $K$ and randomly select $K$ elements of each set $A_i$ to form a matched sample with $n = K$ subjects;
4. estimate $E[T_a]$ by the difference between the average outcomes in the treated and matched samples.

Clearly, the criterion used to select the neighborhoods is the defining characteristic of the procedure. The investigation into which restrictions on this criterion are sufficient for consistency is one of the main objectives of the paper. We begin here, however, only by describing conditions which suffice to express this stylized procedure as a $D$-estimator. Later, in Section 7, we then complete the study by investigating conditions for the consistency of $D$-estimators. To ensure that this stylized matching procedure can be expressed as a $D$-estimator, assume that each $C_i$ satisfies the following two conditions:

(C1) **Condition 1:** $\#A_i \geq K$,

(C2) **Condition 2:** $C_i = C(X_{ri}; X_{c1}, \ldots, X_{cm})$

Condition 1 assures that the third step of the procedure can be always actually performed. Condition 2 excludes, in particular, the possibility of direct dependence between $C_i$ and $[Y_{cj} : 1 \leq j \leq m]$.

Notice that both collections $[C_i : 1 < i < n]$ and $[A_i : 1 < i < n]$ are not necessarily made of pairwise disjoint sets. So, for example, it is possible to have a control subject $j$ belonging to $A_i \cap A_k$ with $i = k$. In such case, subject $j$ may end up being matched twice. In fact, nothing in the procedure constrains each control
subject to be matched at most once. Let $I(i,j)$ be an indicator function that equals to one when the control subject $j$ has been matched to the treated subject $i$ and equals to zero otherwise. Moreover, let $M(j)$ denote the total number of treated subjects to which the control subject $j$ has been matched. So,

$$M(j) = \sum_{i=1}^{n} I(i,j)$$

and as already mentioned, $M(j)$ can actually take values strictly greater than one. It is correct and very useful to consider the matched sample to be made of exactly $M(j)$ replicas of each subject $j$ in the control sample. In particular, notice that

$$\sum_{j=1}^{m} M(j) = n.K$$

which is indeed the size of the matched sample.

When each neighborhood shrinks as the sample sizes get large, the distributions of the balancing score in the treated and matched samples should be expected to become similar. So, after matching, the balancing score and consequently the covariates should become approximately concomitant. Thus, as in a controlled experiment, the difference between the average outcomes in the treated and matched samples should be a consistent estimator for the average treatment effect, $E[T_a]$. So, we define the matching estimator as

$$M = \left( \frac{1}{n} \right) \cdot \sum_{i=1}^{n} Y_{ri} - \left( \frac{1}{nK} \right) \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} I(i,j)Y_{cj}$$

(C21)

Clearly, $M$ can be expressed as a $D$-estimator where

$$p(i,j) = \frac{I(i,j)}{K}$$

(C22)

and

$$q(j) = \frac{M(j)}{n.K}$$

(C23)

The conditions on $[p(i,j) : 1 \leq i \leq n, 1 \leq j \leq m]$ required by the definition of $D$-estimators are ensured by Condition 2. For a simple description of these conditions, see Section 1; for a formal version, see Section 6.

Consider now two examples of simple matching procedures which can be cast into the stylized framework just described. First, consider the following variant of the nearest available pair-matching procedure proposed by Rubin (1973a,b). Assume that $K = 1$ and $C_i = [x : S(X_{ri}, x) \leq B_i]$ where $E_i = Min[S(X_{ri}, X_{cj}) : 1 \leq j \leq m]$ and $S$ is an arbitrary metric defined on $R^b \times R^b$. Notice that, by
Nonparametric Estimation of Causal Effects in Observational Studies

construction, Cl-C2 are both satisfied. Moreover, each treated subject ends up indeed being matched to (one of) the nearest subject(s) available in the control sample.

As a second example, consider the following variant of stratified matching. Set $K = l$ and split $R^d$ into $L$ disjoint blocks $[B_k : l < k < L]$ such that each block contains at least one sampled control subject. Finally, for each treated subject $i = 1, \ldots, n$ a neighborhood $C_i$ is selected such that $C_i = B_k$ if and only if $X_{ri}$ belongs to $B_k$. As already emphasized, the matched sample may end up containing duplicates of some control subjects. By construction, each neighborhood $C_i$ satisfies Conditions 1 and 2 and consequently the procedure is a special case of the stylized method.

These examples as well as the stylized procedure itself have some peculiarities which are not shared by most procedures considered in the literature. First, the stylized procedure assumes that enough matches are always available for all subjects in the treated sample, $#A_i \geq K$. However, except for Rubin’s (1973a,b) nearest available procedure, most matching methods considered in the literature do not assure that matches are always available. In general, when matches cannot be found, the corresponding treated subject is removed from the sample. Such procedures cannot be expressed as $D$-estimators. For instance, consider caliper matching. Here, a fixed tolerance or window, $e$, is set and the neighborhoods are defined as $C_i = \{x : S(X_{ri}, x) < e\}$. So, for any finite sample, there will always be a positive probability that $#A_i < K$. Consequently, C1 is violated.

Another important difference is that while in our stylized procedure matching is performed WITH replacement, in the literature it is usually performed WITHOUT it. This difference derives from the explicit intention of the literature to assure that control subjects are matched at most once, i.e., $M(j) = 0, 1$ for each $j = l, \ldots, m$. Moreover, as a consequence of matching with replacement, in our stylized procedure the order used to organize the treated subjects is irrelevant. In contrast, this order plays an important role in the literature; see, for example, the three variants proposed by Rubin (1973a,b) for the nearest available procedure: random order, the highest to the lowest (HL) and the lowest to the highest (LH).

In summary, some matching procedures might not fit in the framework of $D$-estimators because they either discard unmatchable treated subjects or match them without replacement. Nevertheless, we conjecture that in most interesting circumstances these are all small sample differences. Under some regularity conditions, the probability of having unmatchable treated subjects ($#A_i < K$ for some $i = 1, \ldots, n$) should become negligible as $m - n$ gets large. Moreover, the choice of matching with or without replacement should also be asymptotically irrelevant. So, under minor conditions, those commonly used matching procedures should be asymptotically equivalent to appropriately chosen $D$-estimators. Therefore, to study the asymptotic properties of a particular matching procedure one can proceed in two steps. First, one must establish the asymptotic equivalence
between the procedure and a $D$-estimator so that those results in Section 7 about the consistency of $D$-estimators can be used. In particular, since our stylized procedure can be expressed as a $D$-estimator, its consistency follows directly from the corresponding result for $D$-estimators.

Nevertheless, there are circumstances in which matching does not become asymptotically equivalent to $D$-estimators. An example is the case in which the support of the distribution of $X_r, B_r$ is not contained in the corresponding support of $X_c, B_c$. Here, for sampled treated subjects in $B_r - B_c$ it might be impossible to find close matches, even in large samples. Such cases are not considered in this chapter. In fact, they are excluded by Assumption 7 in Section 6.

Notice that matching is only a very special case of a $D$-estimator. In matching, $p(i, j)$ can only be 0 or 1/$K$. We can generalize this weighting scheme in at least two directions. First, we can choose weights as smoother functions of the distance between treated and control subjects. Second, we can allow the number or matches to depend on $i$. Since the first extension is related to the regression estimator discussed in the following section, here we consider in detail only the second suggestion. Let $K(i)$ denote the number of matches which are sought for the treated subject $i$. Then a natural estimator to consider is the following

$$M = \left(\frac{1}{n}\right) \cdot \sum_{i=1}^{n} \left[ Y_{ri} - \left(\frac{1}{K(i)}\right) \sum_{j=1}^{m} I(i, j)Y_{cj} \right]$$

which is indeed a $D$-estimator with

$$p(i, j) = \left(\frac{I(i, j)}{K(i)}\right)$$

and

$$q(j) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \left(\frac{I(i, j)}{K(i)}\right)$$

If $K(i)$ is defined as

$$K(i) = \# \{j : S(X_{ri}, X_{cj}) < e\}$$

for some tolerance or window $e$, then $M$ defined by (24) is a kind of blocking or kernel estimator similar to the procedure proposed by Mahalanobis (1960). Since in this procedure there is no intention to assure the concomitance of the covariates, it certainly could not be adequately classified as a matching procedure.

To obtain an effective matching, that is, a matched sample with a distribution of covariates very similar to the corresponding distribution in the treated sample, the size of the control sample, $m$, is usually required to be much larger than the size of the treated sample, $n$. In this case $(m \gg n)$ matching tends to ignore a
large portion of the information available in the original control sample. Exactly because of this characteristic, matching has been criticized (McKinlay, 1975, 1977, e.g.,) for being inefficient. However, as Rosenbaum and Rubin (1985) notice, (1) the original size of the control sample, m, is really large; (2) the covariates have already been measured; and (3) the outcomes have not been measured yet and the measurement costs involved are significant.

Then, it may be optimal to work only with a subsample of control subjects in order to reduce measurement costs. In such cases, matching may offer an ingenious criterion for selecting an adequate subsample.

5. Nonparametric Regression

The regression and matching procedures control for non-concomitant covariates using remarkably distinct ideas. The regression procedure is based on the fact that if A1 and A2 are satisfied, then \( E[T_a] \), the average treatment effect among subjects in the target population, equals the average difference between the regression functions for the treated and control populations. This important result has been proved in Section 2, equation (18), which is repeated here for convenience,

\[
E[T_a] = E[F_r(Z_r) - F_c(Z_r)] \tag{18*}
\]

This equation suggests using the estimation of the regression functions \( F_r \) and \( F_c \) as a basic step towards the estimation of \( E[T_a] \). Since nonparametric regression estimation is always expensive, we would like to know if balancing scores could also be used here, as they have been in matching, to reduce the dimensionality of the problem. Fortunately, Property 2 ensures that balancing scores play the same role in regression analysis as they do in matching. More specifically, recall that if A1 is satisfied, then by P2,

\[
E[Y_r - T_r | X_r = x] = E[Y_c | X_c = x]
\]

Moreover, define, similarly to (12) and (13),

\[
G_r(x) = E[Y_r | X_r = x]
\]

and

\[
G_c(x) = E[Y_c | X_c = x]
\]

Then, by the same argument used to prove (15), (18) and (19), we obtain that

\[
E[T_r | X_r = x] = G_r(x) - G_c(x) \tag{25}
\]
and

\[ E[T_a] = E[G_r(X_r) - G_c(X_c)] = \int [G_r(x) - G_c(x)] P_{xr}(dx) \quad (26) \]

Comparing equations (18) and (26), we establish that in regression analysis, as in matching, in order to control for a set of covariates, it is sufficient to control for any balancing score constructed with them. Moreover, equation (26) shows exactly how \( E[T_a] \) can be evaluated once we know the regression functions \( G_r \) and \( G_c \) as well as the probability distribution of \( X_r, P_{xr} \). In the regression procedure, \( E[T_a] \) is then estimated by a sample analog of the right-hand side of (26). This sample analog is denoted by \( R \) and can be expressed as

\[ R = \int [g_r(x) - g_c(x)] P_{rn}(dx) \]

where \( P_{rn} \) denotes the empirical distribution of \( X_r \). \( g_r \) and \( g_c \) are consistent non-parametric estimators for the regression functions \( G_r \) and \( G_c \), respectively. Several examples of how such estimators can be constructed are available in the literature (Gyorfi (see 1981) and Devroye (1981) for reviews and recent developments in this literature). Moreover, by the Glivenko-Cantelli lemma, \( P_{rn} \) converges uniformly to \( P_{xr} \). Thus, since all three components \( g_r, g_c, \) and \( P_{rn} \) of \( R \) converge to their population counterparts, \( R \) should be a consistent estimator of \( E[T_a] \) under weak regularity conditions.

Moreover, by the consistency of \( g_r \) and \( g_c \) and (25), it follows that \( E[T_{ra}|X_r = x] \) could also be consistently estimated by \( g_r(x) - g_c(x) \). If, in addition to A1-A2, the relation between the treated and target populations is such that the following condition is satisfied

(A4) **Assumption 4**: \( E[T_a|X_a = x] = E[T_r|X_r = x] \)

Then, A4 and (25) together imply that

\[ E[T_a|X_a = x] = E[T_r|X_r = x] = G_r(x) - G_c(x) \]

Under these assumptions, \( E[T_a|X_a = x] \) is identified and could be consistently estimated by \( g_r(x) - g_c(x) \). It is then opportune to ask: Why should we constrain our investigation to consistent estimators for \( E[T_a] \) when \( E[T_{ra}|X_r = x] \) itself could also be consistently estimated? We offer some arguments for and against the estimation of the unconditional mean, \( E[T_a] \).

First consider the case in which the effects of a treatment on outcome are independent of the level of the covariates. In this case, the regression function \( G_r \) and \( G_c \) are parallel and consequently

\[ E[T_a|X_a = x] = E[T_a] \]
Clearly, in this circumstance, it would be adequate to focus attention on the estimation of the unconditional mean, $E[T_a]$. Unfortunately, however, in most circumstances, the treatment effects interact with the covariates and, consequently, the regression functions do not end up being parallel. In such cases, $E[T_a]$ can be very sensitive to the distribution one considers for the covariates. For example, when the regression functions cross each other, positive or negative measures for the average treatment impact could be obtained depending on which distribution is used. The dependence of $E[T_a]$ on the choice of the distribution for the covariates reflects, to some extent, an undesirable arbitrariness of $E[T_a]$ as a measure for the effect of a treatment.

Simplicity is also a commonly used argument in favor of the unconditional mean. In fact, when the regression functions do not cross each other, $E[T_a]$ provides a useful summary of the treatment effects.

Nevertheless, we actually decide to confine ourselves to the estimation of the unconditional mean, $E[T_a]$, because it has an additional important advantage: In this case, it is not necessary to estimate $G_r$. This follows directly from the facts that

$$E[Y_r] = E[G_r(X_r)],$$

and

$$E[T_a] = E[Y_r] - E[G_c(X_r)].$$

Hence, $E[G_r(X_r)]$ can be immediately estimated by the average outcome in the treated sample, rendering the estimation of $G_r$ superfluous.

We can then summarize this simplified version of the regression procedure as follows: First, $E[Y_r]$ is consistently estimated by the average outcome in the treated sample: To estimate $E[G_c(X_r)]$ consistently, we proceed in three steps:

1. we use the control sample to estimate $G_c$ nonparametrically using $g_c$;
2. we use the treated sample to form the empirical distribution of $X_r$, $P_{rn}$; finally
3. we estimate $E[G_c(X_r)]$ by integrating $g_c(x)$ with respect to $P_{rn}(x)$.

Thus, the nonparametric regression estimator $R$ can be expressed as

$$R = \left(\frac{1}{n}\right) \cdot \sum_{i=1}^{n} Y_{ri} - \int g_c(x) \cdot P_{rn}(dx)$$

or

$$R = \left(\frac{1}{n}\right) \sum_{i=1}^{n} Y_{ri} - \left(\frac{1}{n}\right) \cdot \sum_{i=1}^{n} g_c(X_{ri})$$

(27)
In the general procedure proposed by Stone (1977), $G_c(x)$ is estimated by a weighted average of the outcomes of subjects in the control group, where the weight attached to the subject $j, w(j, x)$, is assumed to be inversely related to the distance of his or her balancing score to $x, S(X_{cj}, x)$. So, generically

$$g_c(x) = \sum_{j=1}^{m} w(j, x) Y_{cj}$$

Equations (27) and (28) together permit $R$ to be expressed as

$$R = \left(\frac{1}{n}\right) \cdot \sum_{i=1}^{n} Y_{ri} - \left(\frac{1}{n}\right) \cdot \sum_{i=1}^{n} \sum_{i=1}^{m} w(j, X_{ri}) Y_{cj}$$

Hence, the regression estimator can indeed be expressed as a $D$-estimator with $p(i, j) = w(j, X_{ri})$ and $q(j) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} w(j, X_{ri})$.

In the nonparametric regression literature, $w(j, x)$ is assumed to be a measurable function of $[X_{cj} : 1 \leq k \leq m]$. So, the requirements imposed on $p(i, j)$ by the definition of $D$-estimators are trivially satisfied.

Let us now investigate to what extent $D$-estimators generalize the nonparametric regression procedure. Equation (27) is the cornerstone of this discussion. Notice that for consistency we need

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} g_c(X_{ri}) \rightarrow E[G_c(X_r)]$$ in probability as $n, m \rightarrow \infty$ (29)\n
The literature on nonparametric regression offers a menu of sufficient conditions for

$$g_c(x) \rightarrow G_c(x)$$ in probability as $m \rightarrow \infty$ (30)

Clearly, (30) should imply the convergence result stated in (29). In other words, the conditions presented in the literature for the consistency of nonparametric regression estimator also suffice for the consistency of $D$-estimators.

Nevertheless, as we prove in Section 7, these conditions are overwhelmingly strong. Notice that the convergence of $g_c(X_{ri})$ to $G_c(X_{ri})$ is not actually required. If each $g_c(X_{ri})$ is just asymptotically unbiased and the dependence among them is not too strong, then we should expect (29) to hold. In fact, as is well-known, a simple nearest neighbor procedure does not estimate the regression function consistently, but as we show in Section 7, it generates consistent $D$-estimators. To consistently estimate the regression function, a $K_n$-nearest neighbor procedure is required, where $K_n \rightarrow \infty$ as $n \rightarrow \infty$.

More precisely, the procedures considered in the literature on nonparametric regression require
whereas we prove that $D$-estimators are consistent under the weaker condition that
\[
\max\{g(j) : 1 \leq j \leq m\} \to 0 \text{ in probability as } n, m \to \infty \tag{32}
\]
Finally, notice that (31) is not satisfied in matching, where $p(i, j) = 1/K$ or 0, see equation (22). So, the extension we propose actually has some important practical implications.

We conjecture, however, that if a sequence of weights satisfies (32) but not (31), then the $D$-estimators they generate should be inefficient. If correct, this would imply that matching is an inefficient procedure. Therefore, we plan, in a subsequent paper, to study the rate of convergence or alternative procedures.

Rates of convergence for nonparametric regression estimators have already been extensively studied (e.g., Stone, 1982).

6. Basic Assumptions and Definitions

Let $[S; A; P]$ denote our underlying probability space and $[R^b; B], [R; C]$ be two additional measurable spaces. Then, let $X_r, [X_{rn} : n \geq l], X_c,$ and $[X_{cm} : m \geq l]$ be measurable mappings from $[S; A]$ to $[R^b; B]$. Similarly, let $Y_r, [Y_{rn} : n \geq 1], Y_c,$ and $[Y_{cm} : m \geq 1]$ be measurable mappings from $[S; A]$ to $[R; C]$. Then, assume

Assumption 5:

(i) $[(Y_r, X_r); (Y_{rn}, X_{rn}) : n \geq 1]$ is an i.i.d. sequence;
(ii) $[(Y_c, X_c); (Y_{cm}, X_{cm}) : m \geq 1]$ is an i.i.d. sequence;
(iii) $[(Y_r, X_r); (Y_{rn}, X_{rn}) : n \geq l]$ and $[(Y_c, X_c); (Y_{cm}, X_{cm}) : m \geq l]$ are independent.

In addition, let $B_{nm}$ denote the sample information, that is, the sigma-algebra generated by $[(Y_{ri}, X_{ri}) : l \leq i \leq n]$ and $[(Y_{cj}, X_{cj}) : l \leq j \leq m]$; let $C_{nm}$ denote the subsigma-algebra of $B_{nm}$ generated by $[(Y_{ri}, X_{ri}) : 1 \leq i \leq n]$ and $[X_{cj} : 1 \leq j \leq m]$; finally, let $E_{nm}$ (i) denote the subsigma-algebra of $C_{nm}$ generated by $X_r$ and $[X_{cj} : l \leq j \leq m]$.

The objective of the following section is to prove that
\[
D_{nm} \to E[Y_r] - E[G_c(X_r)] \text{ in probability as } n, m \to \infty
\]
under some regularity conditions. This section provides the necessary background. First, we present conditions to ensure that $E[Y_r] - E[G_c(X_r)]$ is well-defined. Then,
we formally introduce the concepts of weighting systems and $D$-estimators. To ensure that $E[Y_r] - E[G_c(X_r)]$ is well-defined, assume

**Assumption 6:** $E[Y_r] < \infty$ and $E|Y_c| < \infty$

This assumption is sufficient for the existence of a $B$-measurable function $G_c : R^b \to R$ such that $G_c(x) = E[Y_c|X_c = x]$, see Billingsley (1979, p. 395). This regression function, however, is only defined almost surely with respect to the probability measure induced by $X_c$, $P_{xc}$. To guarantee that $G_c(X_r)$ is a well-defined random variable, $G_c$ is also required to be defined a.s. ($P_{xr}$), where $P \times r$ denotes the probability measure induced by $X_r$. To meet this requirement, it is sufficient to assume

**Assumption 7:** $P_{xc}$ dominates $P_{xr}$

Given $A6$, $E[Y_r] - E[G_c(X_r)]$ is well-defined if and only if $E[G_c(X_r)] < \infty$. Nevertheless, a slightly stronger condition is required for the consistency of $D$-estimators. To describe this condition, let $Y_c(M) = Y_c I[|Y_c| > M]$ and $H(x, M) = E[|Y_c(M)||X_c = x]$.

**Assumption 8:** $E[H(X_r, M)] \to 0$ as $M \to \infty$.

Notice that $A8$ would be a consequence of $A6$ if $X_r$ and $X_c$ were identically distributed. Moreover, since

$$G_c(x) \leq H(x, M) + E[|Y_c - Y_c(M)||X_c = x] \leq H(x, M) + M$$

$E[G_c(X_r)] < \infty$ is indeed a consequence of $A8$.

Next, we introduce the concepts of weighting systems, REGULAR sequence of weighting systems and $D$-estimators.

**Definition 1:** $[W_{nm} : n, m > l]$ is a sequence of weighting systems when for each $n, m > l$:

(D1) $W_{nm} = [p_{nm}(i, j) : 1 \leq i \leq l, 1 \leq j \leq m]$;

(D2) for each $l \leq i \leq n$ and $l \leq j, k \leq m$, $p_{nm}(i, j)$ and $p_{nm}(i, k)$ are non-negative random variables defined in $(S; A; P)$ such that $E[p_{nm}(i, j)p_{nm}(i, k)|B_{nm}] = E[p_{nm}(i, j)p_{nm}(i, k)|C_{nm}]$;

(D3) $\sum_{j=1}^{m} P_{nm}(i, j) = 1$ for every $1 \leq i \leq n$.

(D4) for each $l \leq j \leq m$, the distribution of $P_{nm}(i, j)$ conditional on $E_{nm}(i)$ does not depend on $i$ for every $1 \leq i \leq n$.

In the literature on nonparametric regression, $P_{nm}(i, j)$ is assumed to be $C_{nm}$-measurable. Examples of weights that are $C_{nm}$-measurable and satisfy D4 are: $P_{nm}(i, j) = P_{nm}(X_{ri}, X_{cj})$ or more generally $P_{nm}(i, j) = P_{nm}(X_{ri}, j; |X_{ri} : l \leq$
In some applications to matching where ties are broken using a random device, \( P_{nm}(i, j) \) is not \( C_{nm} \)-measurable, but D2-D4 are satisfied.

Let \( F_i = F([P_{nm}(i, j) : 1 \leq j \leq m], X_{ri}, [X_{cj} : l \leq j \leq m]) \), where \( F \) is a measurable function. Then, it follows from A5 and D4 that \([F_i : l \leq i \leq n]\) is an identically distributed sequence. This is the main consequence of D4, which is used repeatedly in our proofs.

Now, we introduce the concept of a regular sequence of weighting systems.

**Definition 2**: \([W_{nm} : n, m \geq l] \) is a REGULAR sequence of weighting systems when, in addition to being a sequence or weighting systems, it satisfies the following two additional conditions:

\[(D5) \sum_{j=1}^{m} P_{nm}(i, j) \cdot I[|X_{ri} - X_{cj}| > a] \to 0 \text{ in probability as } m \to \infty \text{ for all } a > 0 \text{ and } l \leq i \leq n;\]

\[(D6) \max[q_{nm}(j) : l \leq j \leq m] \to 0 \text{ in probability as } n, m \to \infty,\]

where \( q_{nm}(j) \) is defined as the total weight attached to control subject \( j \), i.e.,

\[
q_{nm}(j) = \left( \frac{1}{n} \right) \sum_{i=1}^{n} P_{nm}(i, j).
\]

Notice that if D5 holds for some \( i \), then it has to hold for every \( i \) as a consequence of A5 and D4. Finally, \( D \)-estimators are defined as follows:

**Definition 3**: \( D_{nm} = D_{nm}(W_{nm}) \) is a \( D \)-estimator when \( W_{nm} \) is a weighting system, and \( D_{nm} \) can be expressed as

\[
D_{nm} = \left( \frac{1}{n} \right) \cdot \left[ \sum_{i=1}^{n} \left\{ Y_{ri} - \sum_{j=1}^{m} P_{nm}(i, j)Y_{cj} \right\} \right]
\]

It turns out to be convenient to decompose the second term of \( D_{nm} \) into two parts. Define

\[
g_{nm}(i) = \sum_{j=1}^{m} p_{nm}(i, j)G_e(X_{cj})
\]

and

\[
v_{nm}(i) = \sum_{j=1}^{m} p_{nm}(i, j)V_{cj}
\]

where

\[
V_{cj} = Y_{cj} - E[Y_{cj}|X_{cj}] = Y_{cj} - G_e(X_{cj})
\]
Based on these definitions, a D-estimator can be alternatively expressed as

\[ D_{nm} = \left( \frac{1}{n} \right) \cdot \sum_{i=1}^{n} [Y_{ri} - [g_{nm}(i) + v_{nm}(i)]] \]

Notice that as a consequence of A5 and D4, for every \( n, m \geq l, [g_{nm}(i) : 1 \leq i \leq n] \) is an identically distributed sequence.

7. Consistency: General Weighting System

Let \( F : \mathbb{R}^b \to \mathbb{R} \) be a non-negative \( B \)-measurable function such that \( E[F(X_r)] \to \infty \). Moreover, let

\[ L(F) = \frac{1}{E[F(X_r)]} \lim_{m \to \infty} \sup E \left[ \sum_{j=1}^{m} P_{nm}(i, j) \cdot F(X_{cj}) \right] \]

where we let \( L(F) = \infty \) if \( \lim_{m \to \infty} \sup E[\sum_{j=1}^{m} P_{nm}(i, j) \cdot F(X_{cj})] = \infty \).

By A6, A7, and A8, \( E[Q_t(X_r)] < \infty \). Therefore, there exists a sequence \( \{Q_t : t \geq 1\} \) such that for each \( t \geq 1 \), \( Q_t : \mathbb{R}^b \to \mathbb{R} \) is a continuous function having compact support that satisfies

\[ E[|Q_t(X_r) - G_c(X_r)|] < \frac{1}{l} \]

D-estimators are proved to be consistent under the following regularity conditions.

**Assumption 9:** \( 0 < \sup_{l \geq 1} L(|Q_t - G_c|) = L_0 < \infty \)

**Assumption 10:** \( 0 < \sup_{M \geq 0} L(H(., M)) = L_1 < \infty \)

**Theorem 1:** If Assumptions 5 to 10 are satisfied, and \( [W_{nm} : n, m \geq 1] \) is a regular sequence in weighting systems, then the D-estimator \( D_{nm}(W_{nm}) \) converges in probability to \( E[Y_r] - E[G_c(X_r)] \) as \( n, m \to \infty \).

**Proof:** Choose \( c, b > 0 \). We want to show that there exist no and mo such that

\[ P \{ |D_{nm} - E[Y_r] - E[G_c(X_r)]| > 5b \} < 5e \text{ for any } n > no \text{ and } m > mo \] \hspace{1cm} (33)

Write

\[ D_{nm} - E[Y_r] - E[G_c(X_r)] = H_1 - [H_2 + H_3 + H_4] \]

where
\[ H_1 = \left( \frac{1}{n} \right) \sum_{i=1}^{n} Y_{ri} - E[Y_r], \]
\[ H_2 = \left( \frac{1}{n} \right) \sum_{i=1}^{n} G_c(X_{ri}) - E[G_c(X_r)], \]
\[ H_3 = \left( \frac{1}{n} \right) \sum_{i=1}^{n} [g_{nm}(i) - G_c(X_{ri})], \]
\[ H_4 = \left( \frac{1}{n} \right) \sum_{i=1}^{n} v_{nm}(i) \]

By the implication rule

\[
P[|D_{nm} - [E[Y_r] - E[G_c(X_r)]]| > 5b] \leq P[|H_1| > b] + P[|H_2| > b] + P[|H_3| > 2b] + P[|H_4| > 2b]
\] (34)

By A5-A6 and the Weak Law of Large Numbers, there exists \( n_1 \) such that

\[
P[|H_1| > b] < e \text{ for } n > n_1
\] (35)

Since A8 implies \( E|G_c(X_r)| < \infty \), A5 and the Weak Law of Large Numbers, there exists \( n_2 \) such that

\[
P[|H_2| > b] < e \text{ for } n > n_2
\] (36)

Now consider \( H_3 \). BY Markov’s inequality

\[
P[|H_3| > b] \leq \frac{1}{b} E|H_3|
\]

By Minkowski’s inequality,

\[
E|H_3| \leq \left( \frac{1}{n} \right) \sum_{i=1}^{n} E|g_{nm}(i) - G_c(X_{ri})|
\]

By A5 and D4,

\[
E|g_{nm}(i) - G_c(X_{ri})| = E|g_{nm}(k) - G_c(X_{rk})| \text{ for all } 1 \leq i, k \leq n
\]

By Lemma 1,

\[
E|g_{nm}(i) - G_c(X_{ri})| \to 0 \text{ as } m \to 0
\]
Therefore, there exists an $m_3$ such that

$$E|g_{nm}(i) - G_c(X_{ri})| < e.b \text{ for } m > m_3$$

Consequently,

$$P[|H_3| > b] < e \text{ for } m > m_3 \quad (37)$$

Now consider $H_4$. Let

$$V_{cj}(M) = Y_{cj}(M) - E[Y_{cj}(M)|X_{cj}]$$

where, as before,

$$Y_{cj}(M) = Y_{cj}.I[|Y_{cj}| > M]$$

Write $H_4 = H_5 + H_6$ where

$$H_5 = \left( \frac{1}{n} \right) \sum_{i=1}^{n} \sum_{j=1}^{m} P_{nm}(i,j).V_{cj}(M),$$

$$H_6 = \left( \frac{1}{n} \right) \sum_{i=1}^{n} \sum_{j=1}^{m} P_{nm}(i,j).[V_{cj} - V_{cj}(M)]$$

Consider $H_5$. By Markov’s and Minkowski’s inequalities

$$P[|H_5| > b] \leq \left( \frac{1}{b.n} \right) \sum_{i=1}^{n} \sum_{j=1}^{m} E[P_{nm}(i,j).|V_{cj}(M)|]$$

$$= \left( \frac{1}{b} \right) \sum_{j=1}^{m} E[P_{nm}(i,j).|V_{cj}(m)|]$$

where the last equality follows from A5-D4. Next, we use the iterated expectation formula, D2 and the fact that $|V_{cj}(M)|$ is $B_{nm}$-measurable to obtain

$$E[P_{nm}(i,j).|V_{cj}(M)|] = E[P_{nm}(i,j).E[|V_{cj}(M)||C_{nm}]]$$

Moreover, by A5,

$$E[|V_{cj}(M)||C_{nm}] = E[|V_{cj}(M)||X_{cj}]$$

Hence,
\[ \sum_{j=1}^{m} E[P_{nm}(i,j) | V_{cj}(M)] = \sum_{j=1}^{m} E[P_{nm}(i,j) E[V_{cj}(M) | X_{cj}]] \]

\[ \leq 2 \sum_{j=1}^{m} E[P_{nm}(i,j) E[Y_{cj}(M) | X_{cj}]] \]

where the last step follows from

\[ E[V_{cj}(M) | X_{cj}] = E[Y_{cj}(M) - E[Y_{cj}(M) | X_{cj}] | X_{cj}] \]

\[ \leq E[Y_{cj}(M) | X_{cj}] + |E[Y_{cj}(M) | X_{cj}]| \]

\[ \leq 2 E[Y_{cj}(M) | X_{cj}] \]

Moreover, by A10, there exists an \( m_{5} \) such that for all \( m > m_{5} \)

\[ \sum_{j=1}^{m} E[P_{nm}(i,j) E[Y_{cj}(M) | X_{cj}]] < 2 L_{1} E[H(X_r, M)] \]

By A8, there exists an \( M_{o} \) such that

\[ E[H(X_r, M)] < \frac{eb}{4 L_{1}} \text{ for } M > M_{o} \]

Hence

\[ P[|H_{5}| > b] < e \text{ for } M > M_{o} \text{ and } m > m_{5} \]  (38)

Consider \( H_{6} \). By Chebyshev’s inequality

\[ P[|H_{6}| > b] \leq \frac{1}{b} E[|H_{6}|^2] \]

By the definition of \( qnm(j) \)

\[ H_{6} = \sum_{j=1}^{m} q_{nm}(j) [V_{cj} - V_{cj}(M)] \]

By A5, D4, and the fact that \( V_{cj} - V_{cj}(M) < 2M \)
\[
E \left[ |H_6|^2 | C_{nm} \right] = E \left[ \sum_{j=1}^{m} q_{nm}(j) \right]^2 \cdot E \left[ |V_{cj} - V_{cj}(M)|^2 | X_{cj}\right] \\
< E \left[ \max \left[ q_{nm}(j) : 1 \leq j \leq m \right], \sum_{j=1}^{m} q_{nm}(j) \right] E \left[ |V_{cj}| \right] \\
\quad - \quad V_{cj}(M)^2 | X_{cj} |] \\
< 4M^2 - E \left[ \max \left[ q_{nm}(j) : 1 \leq j \leq m \right] \right]
\]

Since \( \max \left[ q_{nm}(j) : 1 \leq j \leq m \right] \leq 1 \), by D6 and the Dominated Convergence Theorem, there exist \( n_6 \) and \( m_6 \) such that
\[
E \left[ |H_6|^2 \right] < e. b^2 \text{ for every } n > n_6 \text{ and } m > m_6
\]
Consequently,
\[
P \left( |H_6| > b \right) < e \text{ for every } n > n_6 \text{ and } m > m_6 \quad (39)
\]
By the implications rule, (38) and (39)
\[
P \left( |H_4| > 2b \right) < 2e \text{ for every } n > n_6 \text{ and } m > \max \left[ m_5, m_6 \right] \quad (40)
\]
Let
\[
n_o = \max \left[ n_1, n_2, n_6 \right]
\]
and
\[
m_o = \max \left[ m_3, m_5, m_6 \right]
\]
Then, by (35), (36), (37), (40) and (34)
\[
P \left( |D_{nm} - E[Y_r] - E[G_c(X_{r})]| > 5b \right) < 5e \text{ for any } n > n_o \text{ and } m > m_o
\]
as we would like to show.

Finally, consider Lemma 1. This lemma is very closely related to Proposition 1 in Stone (1977, p. 607). The following proof is presented just for completeness.

**Lemma 1:** If A5 to A9 are satisfied and \( W_{nm} \) is a regular sequence of weighting systems, then
\[
E \left[ \sum_{j=1}^{m} P_{nm}(i,j) |G_c(X_{cj}) - G_c(X_{ri})| \right] \to 0 \text{ as } m
\]
**Proof:** Given $\varepsilon > 0$. Choose $l$ large enough such that

$$E[Q_l(X_r) - G_c(X_r)] < \frac{\varepsilon}{2(1 + 2L_0)} \quad (41)$$

By the triangle inequality

$$E \left[ \sum_{j=1}^{m} P_{nm}(i, j)[G_c(X_{cj}) - G_c(X_{ri})] \right]$$

$$\leq E \left[ \sum_{j=1}^{m} P_{nm}(i, j)[Q_l(X_{cj}) - Q_l(X_{ri})] \right]$$

$$+ E \left[ \sum_{j=1}^{m} P_{nm}(i, j)[Q_l(X_{cj}) - G_c(X_{cj})] \right] + E[Q_l(X_{ri}) - G_c(X_{ri})] \quad (42)$$

by A9 and (41), there exists an $m_0$ such that for all $m > m_0$

$$E \left[ \sum_{j=1}^{m} P_{nm}(i, j)[Q_l(X_{cj}) - G_c(X_{cj})] \right]$$

$$\leq 2L_0.E[Q_l(X_r) - G_c(X_r)]$$

$$< 2L_0.\frac{\varepsilon}{2(2L_0 + 1)} \quad (43)$$

Moreover, let $\sup |Q_l| = M$. Because $Q_l$ is uniformly continuous, there exists $a > 0$ such that $|Q_l(r) - Q_l(s)| < \varepsilon/4$ if $r, s$ are in $R^b$ and $|r - s| < a$. Hence,

$$E \left[ \sum_{j=1}^{m} P_{nm}(i, j)[Q_l(X_{cj}) - Q_l(X_{ri})] \right]$$

$$\leq 2M.E \left[ \sum_{j=1}^{m} P_{nm}(i, j).I[|X_{ri} - X_{cj}| > a] \right] + \frac{\varepsilon}{4} \quad (44)$$

Finally, by D5 and the Dominated Convergence Theorem, there exists an $m_1$ such that
Therefore, by (41)-(45)

\[ E \left[ \sum_{j=1}^{m} P_{nm}(i,j) I[|X_{ri} - X_{cj}| > a|a] \right] < \frac{e}{8M} \text{ for every } m > m_1 \]  \hspace{1cm} (45)

as we would like to show.

8. Conclusion

In this paper, we investigate a nonparametric procedure which aims to estimate the causal effect of a treatment in the context of observational studies. The procedures we study are members of a class of estimators which can be expressed as the DIFFERENCE between the average outcome in the treated sample and an adequately chosen weighted average of the outcomes in the control group. We refer to estimators in this class as \( D \)-estimators. Moreover, it has been shown that both the matching and the nonparametric regression procedures as used to control for non-concomitant covariates in observational studies can generally be expressed as \( D \)-estimators.

The well-known problem of identification of causal effects in observational studies has also been reviewed. We show that \( D \)-estimators should only be used in circumstances in which the assignment of subjects to treatment is random given the vector of observed covariates. This condition has been referred to as the “strongly ignorable assumption” by Rosenbaum and Rubin. Heckman and Robb refer to this condition as the “selection on observable model.”

Following the work of Rosenbaum and Rubin, we prove that when a KNOWN balancing score is available, to control for this balancing score using either a matching or a regression procedure suffices to account for the non-concomitant covariates. The consistency of \( D \)-estimators is proved under some regularity conditions. They impose very weak restrictions on the distribution of the covariates. Nevertheless, for a better understanding of the asymptotic properties of \( D \)-estimators, it would be very useful to study, in the future, the rates of convergence associated with alternative regular sequences of weighting systems.

We also show that most commonly used matching procedures are expected to be only asymptotically equivalent to \( D \)-estimators. Hence, a detailed study of the appropriate conditions for this equivalence would be an important topic for future research.
Finally, a study of the robustness of $D$-estimators to the presence of measurement errors associated with the covariates would also be an important contribution to theory and practice of $D$-estimators.

References


