

Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review*

Guido W. Imbens
UC Berkeley, and NBER

First Draft: July 2002
This Draft: May 2003

Abstract

Recently there has been a surge in econometric work focusing on estimating average treatment effects under various sets of assumptions. One strand of this literature has developed methods for estimating average treatment effects for a binary treatment under assumptions of exogeneity or unconfoundedness, also known as selection on observables. This assumption requires that given a set of covariates systematic differences in outcomes between treated units and control units with the same values for these covariates are attributable to the treatment. The recent econometric literature has considered estimation and inference under weaker assumptions than typically considered in the earlier literature, in particular by avoiding distributional and functional form assumptions, and has established efficiency bounds. Various approaches to semiparametric estimation have been proposed, including estimating the unknown regression functions, matching, methods using the propensity score including weighting and blocking, and combinations of these approaches.

In this paper I will review the state of this literature and discuss some of the unanswered questions, focusing in particular on the practical implementation of these methods, the plausibility of this exogeneity assumption in economic applications, the relative performance of the various semiparametric estimators when the key assumptions (unconfoundedness and overlap) are satisfied, alternative estimands such as quantile treatment effects, and alternative methods such as Bayesian inference.

JEL Classification: C14, C21, C52

Keywords: *Average Treatment Effects, Unconfoundedness, Semiparametric Methods, Matching, Propensity Score, Exogeneity*

*Department of Economics, and Department of Agricultural and Resource Economics, University of California at Berkeley, 661 Evans Hall #3880, Berkeley, CA 94720-3880. I am grateful to Xiangyi Meng and Caroline Hoxby and two referees for comments and to a number of collaborators, Alberto Abadie, Joshua Angrist, Susan Athey, Gary Chamberlain, Keisuke Hirano, V. Joseph Hotz, Julie Mortimer, Jack Porter, Whitney Newey, Geert Ridder, Paul Rosenbaum, and Donald Rubin, for many discussions on the topics of this paper. Financial support for this research was generously provided through NSF grants SBR 9818644 and SES 0136789 and the Giannini Foundation. Electronic correspondence: imbens@econ.berkeley.edu, <http://elsa.berkeley.edu/users/imbens/>.

1 Introduction

Since the work by Ashenfelter and Card (1985), Card and Sullivan (1988), Heckman and Robb (1984), Lalonde (1986) and others there has been much interest in econometric methods for estimating the effects of active labor market programs such as job search assistance or classroom teaching programs. This interest has led to a surge in theoretical work focusing on estimating average treatment effects under various sets of assumptions. See for general surveys of this literature Angrist and Krueger (2000), Heckman, Lalonde and Smith (2000), and Blundell and Costa-Dias (2002).

One strand of this literature has developed methods for estimating the average effect of receiving or not receiving a binary treatment under the assumption that the treatment satisfies some form of exogeneity. Different forms of this assumption are referred to as unconfoundedness (Rosenbaum and Rubin, 1983a), selection on observables (Barnow, Cain, and Goldberger, 1980; Fitzgerald, Gottschalk, and Moffitt, 1998), or the conditional independence assumption (Lechner, 1998). In the remainder of this paper I will use the terms unconfoundedness and exogeneity interchangeably. The implication of these assumptions is that systematic (e.g., average or distributional) differences in outcomes between treated units and control units with the same values for these covariates are attributable to the treatment. Much of the recent literature has built on the work in the statistical literature by Rubin (1973ab, 1977, 1978), Rosenbaum and Rubin (1983ab, 1984), Holland (1986) and others. The recent literature considered estimation and inference without distributional and functional form assumptions. Hahn (1998) derived efficiency bounds assuming only unconfoundedness and some regularity conditions. Various estimators have been proposed under these conditions. These include (i) estimating the unknown regression functions of the outcome on the covariates (Hahn, 1998; Heckman, Ichimura, and Todd, 1997; Heckman, Ichimura, Smith and Todd, 1998), (ii) matching on covariates (Rosenbaum, 1995; Abadie and Imbens, 2002) (iii) methods based on the propensity score including blocking (Rosenbaum and Rubin, 1984) and weighting (Hirano, Imbens, and Ridder, 2001), and (iv) combinations of these approaches, for example weighting and regression (Robins and Rotnitzky, 1995) or matching and regression (Abadie and Imbens, 2002).

In this paper I will review the state of this literature, with a particular emphasis on implications for empirical work. In addition I will discuss some of the questions that are still outstanding. The organization of the paper is as follows. In Section 2 I will set up the notation and the basic issues. Here I will also discuss the difference between population versus sample average treatment effects. The recent econometric literature, in contrast to some of the original experimental literature (Fisher, 1925, Neyman, 1923), has largely focused on estimation of the population average treatment effect and its counterpart for the subpopulation of treated units. An alternative is to consider estimation of the average effect of the treatment for the units in the sample. Many of the estimators proposed can be interpreted as estimating either the average treatment effect for the sample at hand or the average treatment effect for the population. The focus, on either population or sample average treatment effects, matters, however, for the asymptotic variance, with the variance of the estimators for the sample average treatment effect in general smaller. This perspective also has implications for the efficiency bounds

and for the form of estimators for the asymptotic variance. In this section I will also discuss alternative estimands. Almost the entire literature has focused on average treatment effects. In many cases such measures of typical effects may mask important distributional changes. These can be captured more easily by focusing on quantiles of the distributions of outcomes in the presence and absence of the treatment (Gelbach and Hoynes, 2002; Firpo, 2002).

In Section 3 I will discuss in more detail some of the semiparametric estimators for the average treatment effect that have been proposed in the literature, including those based on regression, matching and the propensity score. I will focus particularly on implementation and compare the different decisions regarding choices for smoothing parameters faced by researchers using the various estimators.

In Section 4 I will discuss estimation of the variances of the various estimators. For most of the estimators proposed in the literature corresponding estimators for the variance have been proposed. These typically involve additional nonparametric regression. In practice researchers have often used bootstrap methods. Such methods have not been formally justified, and in fact in the case where one is interested in the average treatment effect for the sample, do not appear to be appropriate. Here I discuss in more detail a simple estimator for the variance based on matching developed by Abadie and Imbens (2002) that does not require additional nonparametric estimation.

Section 5 contains a discussion of different approaches to assessing the plausibility of the two key assumptions, exogeneity or unconfoundedness and overlap in the covariate distributions. The first of these assumptions is in principle untestable. Nevertheless a number of tests have been proposed that are useful for assessing the credibility of this assumption (Heckman and Hotz, 1989; Rosenbaum, 1984). One may also wish to assess the sensitivity of the results to this assumption in a sensitivity analysis (Rosenbaum and Rubin, 1983; Imbens, 2003), or, in its extreme form, a bounds analysis (Manski, 1990, 1995). I also discuss the choice of covariates in this section. The second assumption of overlap in the distribution of covariates in the treated and control subpopulations is an assumption on the joint distribution of observable variables. However, as it only involves inequality restrictions there are no direct tests. Nevertheless, in practice it is extremely important to assess whether there is sufficient overlap to draw credible inferences. If there is insufficient overlap in the full sample, one may wish to limit inferences to the average effect for the subset of the covariate space where there is overlap.

In Section 6 I discuss a number of applications of these methods. Some of the most interesting of these applications involve comparisons of the non-experimental methods to results based on randomized experiments, and thus direct assessments of the plausibility of the unconfoundedness assumption. In addition I will review a number of simulation studies designed to shed light on the applicability of the methods in various settings.

I will not address in this paper a number of methods for estimating average treatment effects that do not rely on exogeneity assumptions. This includes methods where observed covariates are not adjusted for, such as instrumental variables methods (Björklund and Moffit, 1987; Heckman, 1990; Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Ichimura and Taber, 2000; Abadie, 2002; Chernozhukov and Hansen, 2001). The methods not discussed in this paper also includes methods exploiting the presence of additional data such as difference-

in-differences methods in repeated cross-sections (Abadie, 2001; Blundell, Gosling, Ichimura, and Meghir, 2002; Athey and Imbens, 2002), and regression discontinuity methods where the overlap assumption is violated (VanderKlaauw, 2002; Hahn, Todd and VanderKlaauw, 2000; Angrist and Lavy, 1999; Black, 1999; Lee, 2001). I will also limit the discussion to binary treatments, excluding models with dynamic treatment regimes as in Ham and Lalonde (1996), and Gill and Robins (2002) and models with static multi-valued treatments as in Imbens (2000) and Lechner (2001). Reviews of many of these methods can be found in Angrist and Krueger (2000), Heckman, Lalonde and Smith (2000), and Blundell and Costa-Dias (2002).

2 Estimands, Identification and Efficiency Bounds

2.1 Definitions

In this paper we use the potential outcome notation. This dates back to the analysis of randomized experiments by Fisher (1935) and Neyman (1923). After being forcefully advocated by Rubin (1974, 1977, 1978), it is now standard in the literature on both experimental and non-experimental program evaluation. We have N units, indexed by $i = 1, \dots, N$, viewed as drawn randomly from a large population. Each unit is characterized by a pair of potential outcomes, $Y_i(0)$ for the outcome under the control treatment and $Y_i(1)$ for the outcome under the active treatment. In addition each unit has a vector of characteristics, often called covariates, pretreatment variables or exogenous variables, denoted by X_i .¹ These variables are assumed not to have been affected by the treatment, often by being measured before the unit was exposed to the treatment. Importantly these can include lagged outcomes. Finally, each unit is exposed to a single treatment, $W_i = 0$ if unit i is exposed to the control treatment and $W_i = 1$ if unit i is exposed to the active treatment. We observe for each unit in a random sample from a large population the triple (W_i, Y_i, X_i) , where Y_i is the realized outcome:

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

Distributions of (W, Y, X) refer to the distribution induced by the random sampling from the superpopulation.

A couple of additional pieces of notation will be useful in the following. The propensity score (Rosenbaum and Rubin, 1983a) is defined as the conditional probability of receiving the treatment,

$$e(x) \equiv \Pr(W = 1|X = x) = \mathbb{E}[W|X = x].$$

Define for $w \in \{0, 1\}$ the two conditional regression and variance functions:

$$\mu_w(x) \equiv \mathbb{E}[Y(w)|X = x], \quad \sigma_w^2(x) \equiv \mathbb{V}(Y(w)|X = x).$$

¹Calling such variables exogenous variables is somewhat at odds with some formal definitions of exogeneity (e.g., Engle, Hendry and Richard, 1974), as knowledge of their distribution can be informative about the average treatment effects. See, e.g., Frölich (2002) for additional discussion.

For completeness, let $\rho(x)$ be the conditional correlation coefficient of $Y(0)$ and $Y(1)$ given $X = x$. As one never observes $Y_i(0)$ and $Y_i(1)$ for the same unit i , the data only contain indirect and very limited information about this correlation coefficient.²

2.2 Estimands: Average Treatment Effects

We are primarily interested in a number of average treatment effects. This is less limiting than it may seem, as this will include averages of arbitrary transformations of the original outcomes.

The first estimand, and the most commonly studied, is the population average treatment effect (PATE):

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

Alternatively we may be interested in the population average treatment effect for the treated (PATT) :

$$\tau_T^P = \mathbb{E}[Y(1) - Y(0)|W = 1].$$

Rubin (1977) studies this estimand in the context of educational programs. Heckman and Robb (1984) and Heckman, Ichimura and Todd (1997) argue that the subpopulation of treated units is often of more interest than the overall population in the context of narrowly targeted labor market programs. For example, if a program is specifically targeted at individuals disadvantaged in the labor market, there is often little interest in the effect of such a program on individuals with histories of strong attachment to the labor market.

We will also look at sample average versions of these two population averages. These sample average estimands focus on the average of the treatment effect in the specific sample, rather than in the population at large. First, the sample average treatment effect (SATE):

$$\tau^S = \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)),$$

and the sample average treatment effect for the treated (SATTT):

$$\tau_T^S = \frac{1}{N_T} \sum_{i:W_i=1} (Y_i(1) - Y_i(0)),$$

where $N_T = \sum_{i=1}^N W_i$ is the number of treated units.

The difference between the sample and population average treatment effects is often ignored in the recent literature, although the distinction has a long tradition in the analysis of randomized experiments (e.g., Neyman, 1923). Without further assumptions the sample contains no information about the population average treatment effect beyond the sample average treatment effect. To see this consider the case where we observe the sample $(Y_i(0), Y_i(1), W_i, X_i)$,

²As Heckman, Smith, and Clemens (1997) point out, however, one can draw some limited inferences about the correlation coefficient from the shape of the two marginal distributions of $Y(0)$ and $Y(1)$.

$i = 1, \dots, N$ (that is, we observe for each unit both potential outcomes). In that case the sample average treatment effect $\tau^S = \sum_i (Y_i(1) - Y_i(0))/N$ is known. Obviously the best estimator for the population average effect τ^P is τ^S . Note that if our target is τ^S , we can estimate it without error in this case. However, we cannot estimate τ^P without error even with such a sample where all potential outcomes are observed. This simple argument will be seen to have two implications. First, one can estimate the sample average treatment effect at least as accurately as the population average treatment effect, and typically more accurately. In fact, the difference between the two variances is the variance of the treatment effect, and so this is zero only when the treatment effect is constant. Second, a good estimator for one is automatically a good estimator for the other. One can therefore interpret many of the estimators for PATE or PATT as estimators for SATE or SATT, with lower implied standard errors as discussed in more detail in Section 2.6.

2.3 Identification

We make the following key assumption about the treatment assignment:

Assumption 2.1 (UNCONFOUNDEDNESS)

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

This assumption was first articulated in this form in Rosenbaum and Rubin (1983). Lechner (1999) refers to the same assumption as the “conditional independence assumption.” Barnow, Cain and Goldberger (1971) describe a regression-based version of this as “selection on observables.”

To see the link with standard exogeneity assumptions, suppose that the treatment effect is constant, $\tau = Y_i(1) - Y_i(0)$ for all i . Then we can write

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i,$$

where $\alpha = \mathbb{E}[Y(0)]$ and $\varepsilon_i = Y_i(0) - \mathbb{E}[Y(0)]$. Given the constant treatment effect assumption, exogeneity of the treatment indicator (that is, independence of W_i and ε_i) is identical to unconfoundedness. Without the constant treatment effect assumption, however, unconfoundedness does not imply a linear relation with (mean-)independent errors.

Next, we make a second assumption regarding the joint distribution of treatments and covariates:

Assumption 2.2 (OVERLAP)

$$0 < Pr(W = 1|X) < 1.$$

For many of the formal results one will also need smoothness assumptions on the conditional regression functions $\mu_w(x)$ and the propensity score $e(x)$, and moment conditions on $Y(w)$. I will not discuss these regularity conditions here. Details can be found in the references for the specific estimators given below.

There has been some controversy about the plausibility of Assumptions 2.1 and 2.2 in economic settings. In this debate it has sometimes been argued that optimizing behavior by agents precludes their choices being independent of the potential outcomes, whether or not conditional on covariates. This appears an unduly narrow view of the unconfoundedness assumption. I will discuss three arguments for considering these assumptions. The first is a statistical, data descriptive motivation. A natural starting point in the evaluation of any treatment would appear to be a comparison of average outcomes for treated and control units. A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment). Such an analysis may not lead to the final word on the efficacy of the treatment, but the absence of such an analysis would seem difficult to rationalize in a serious attempt to understand the evidence of the data regarding the effect of the treatment.

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not receive the treatment. The question is typically not whether such a comparison should be made, rather which units should be compared, that is which units would have been comparable to treated units had those treated units not been treated. Economic theory can be helpful in classifying variables into those that need to be adjusted for versus those that do not need to be adjusted for on the basis of their role in the decision process (e.g., whether they enter the utility function or the constraints). Given that, the unconfoundedness assumption merely asserts that all those that need to be adjusted for are observed by the researcher. This is an empirical question, and not one that should be controversial as a general principle. It is clear that alternatives where some of the variables that need to be adjusted for are in fact not observed, will require strong assumptions to allow for identification of the effects of interest. Such assumptions include instrumental variables settings where some variables are assumed to be independent of the potential outcomes. Absent those assumptions only bounds can be identified (e.g., Manski, 1990, 1995).

A third, related, argument is that even when agents optimally choose their treatment, two agents with the same values for observed characteristics may differ in their treatment choices without invalidating the unconfoundedness assumption. The difference in their choices may be driven by differences in unobserved characteristics that are themselves unrelated to the outcomes of interest. This will depend critically on the exact nature of the optimization program faced by the agents. In particular it may be important that the objective of the decision maker is distinct from the outcome that is of interest to the evaluator. For example, suppose we are interested in estimating the effect of a binary input on the output of a firm. Production is a stochastic function of this input because other inputs (e.g., weather) are not under the control of the firm. Suppose also that a firm chooses a production level to maximize expected profits (e.g., production times a fixed price minus costs). If unobserved marginal costs differ between firms, independent of the errors in the firms' forecast of production given inputs,

then unconfoundedness will hold and the effect of the input on production can be identified. Note that under the same assumptions one cannot necessarily identify the effect of the input on profits. See for a related discussion in the context of instrumental variables Athey and Stern (1998). Heckman, Lalonde and Smith (2000) discuss different models where individuals do attempt to optimize the same outcome that is the variable of interest to the evaluator. They show that selection on observables assumptions can be justified by imposing restrictions on the way individuals form their expectations about the unknown potential outcomes.

In general, therefore, a researcher may wish to, either as a final analysis, or as part of a larger investigation, consider estimates based on the unconfoundedness assumption.

Given the two key assumptions, unconfoundedness and overlap, one can identify the average treatment effects. The key insight is that given unconfoundedness the following equalities holds:

$$\mathbb{E}[Y(w)|X = x] = \mathbb{E}[Y(w)|W = w, X = x] = \mathbb{E}[Y|W = w, X = x],$$

and hence both are equal to $\mu_w(x)$ by definition. Therefore, since the righthand side is directly estimable from the data, if we want to estimate the average treatment effect τ , we can first estimate the average treatment effect for a subpopulation with covariates $X = x$ as

$$\begin{aligned} \tau(x) &\equiv \mathbb{E}[Y(1) - Y(0)|X = x] = \mathbb{E}[Y(1)|X = x] - \mathbb{E}[Y(0)|X = x] \\ &= \mathbb{E}[Y(1)|X = x, W = 1] - \mathbb{E}[Y(0)|X = x, W = 0] \\ &= \mathbb{E}[Y|X, W = 1] - \mathbb{E}[Y|X, W = 0]. \end{aligned}$$

To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y|X = x, W = w]$ for all values of w and x in the support of these variables. This is where the second assumption comes in. If the overlap assumption is violated at $X = x$, it would be infeasible to estimate both $\mathbb{E}[Y|X = x, W = 1]$ and $\mathbb{E}[Y|X = x, W = 0]$ because at those values there would be either only treated, or only control units.

Some researchers use weaker versions of the unconfoundedness assumption (e.g., Heckman, Ichimura, and Todd, 1998). If the interest is in the population average treatment effect, it is sufficient to assume that

Assumption 2.3 (MEAN INDEPENDENCE)

$$\mathbb{E}[Y(w)|W, X] = \mathbb{E}[Y(w)|X],$$

for $w = 0, 1$.

In practice it is rare that a convincing case is made for the weaker assumption 2.1 without the case being equally strong for the stronger version 2.3. If the weaker assumption is made one cannot necessarily identify average treatment effects on all transformations of the original outcome.

One can weaken the unconfoundedness assumption in a different direction if one is interested in the average effect for the treated only (e.g., Heckman, Ichimura and Todd, 1997). In that case one only needs the assumption

Assumption 2.4 (UNCONFOUNDEDNESS FOR CONTROLS)

$$Y(0) \perp W \mid X.$$

This is sufficient for identification of PATT and SATT because moments of the distribution of $Y(1)$ for the treated are directly estimable.

An important result building on the unconfoundedness assumption shows that one does not need to condition simultaneously on all covariates. The following result is due to Rosenbaum and Rubin (1983a). With the propensity score defined as $e(X) \equiv \Pr(W = 1|X = x) = \mathbb{E}[W|X = x]$, the following result shows that all biases due to observable covariates can be removed by conditioning solely on the propensity score:

Lemma 2.1 (UNCONFOUNDEDNESS GIVEN THE PROPENSITY SCORE, ROSENBAUM AND RUBIN, 1983A)

Suppose that Assumption 2.1 holds. Then:

$$(Y(0), Y(1)) \perp W \mid e(X).$$

Proof: We will show that $\Pr(W = 1|Y(0), Y(1), e(X)) = \Pr(W = 1|e(X)) = e(X)$, implying independence of $(Y(0), Y(1))$ and W conditional on $e(X)$. First, note that

$$\begin{aligned} \Pr(W = 1|Y(0), Y(1), e(X)) &= \mathbb{E}[W = 1|Y(0), Y(1), e(X)] \\ &= \mathbb{E} \left[\mathbb{E}[W = 1|Y(0), Y(1), e(X), X] \mid Y(0), Y(1), e(X) \right] \\ &= \mathbb{E} \left[\mathbb{E}[W = 1|Y(0), Y(1), X] \mid Y(0), Y(1), e(X) \right] \\ &= \mathbb{E} \left[\mathbb{E}[W = 1|X] \mid Y(0), Y(1), e(X) \right] = \mathbb{E} [e(X)|Y(0), Y(1), e(X)] = e(X), \end{aligned}$$

where the last equality follows from unconfoundedness. The same argument shows that

$$\Pr(W = 1|e(X)) = \mathbb{E}[W = 1|e(X)] = \mathbb{E} \left[\mathbb{E}[W = 1|X] \mid e(X) \right] = \mathbb{E} [e(X)|e(X)] = e(X).$$

□

Extensions of this result to the multivalued treatment case are given in Imbens (2000) and Lechner (2001). To interpret the Rosenbaum-Rubin result, recall the textbook formula for omitted variable bias in the linear regression model. Suppose we have a regression model with two regressors:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2 \cdot X_i.$$

The bias of omitting X from the regression on the coefficient on W is equal to $\beta_2 \cdot \delta$, where δ is the coefficient on W in a regression of X on W . By conditioning on the propensity score we remove the correlation between X and W because $X \perp W|e(X)$. Hence omitting X no longer leads to any bias (although it may still lead to some efficiency loss as we will see later).

2.4 Distributional and Quantile Treatment Effects

Most of the literature has focused on estimating average treatment effects. There are, however, many cases where one may wish to estimate other features of the joint distribution of outcomes. Heckman, Smith and Clemens (1997) focus on estimation of bounds on the joint distribution of $(Y(0), Y(1))$. One cannot without strong untestable assumptions identify the full joint distribution, as one can never observe both potential outcomes together, but one can nevertheless derive bounds on functionals of the two distributions. In instrumental variables settings Abadie, Angrist and Imbens (2002) and Chernozhukov and Hansen (2001) investigate estimation of differences in quantiles of the potential outcome distributions, either for the entire population or for subpopulations.

The two assumptions 2.1 and 2.2 allow for identification of more than just the average treatment effect. In fact, they allow for estimation of the full marginal distributions of $Y(0)$ and $Y(1)$. To see this, first note that we can identify not just the average treatment effect, but also the averages of the two potential outcomes. Second, note that we can similarly identify by these assumptions the average of any function of the basic outcome. Hence we can identify the average value of the indicator $1\{Y \leq y\}$, and thus the distribution function of the potential outcomes at y .

Given identification of the two distribution functions it is clear that one can also identify quantiles of the two potential outcome distributions. Gelbach and Hoynes (2002) focus on estimating quantiles in a randomized experiment. Firpo (2002) considers the nonexperimental case and develops an estimator under unconfoundedness.

2.5 Efficiency Bounds and Asymptotic Variances for Population Average Treatment Effects

Next I review some results on the efficiency bound for estimators for the average treatment effect τ^P and τ_T^P under the two assumptions unconfoundedness and overlap (Assumption 2.1 and 2.2), combined with some smoothness assumptions on the conditional expectations of potential outcomes and treatment indicator. These are calculated in Hahn (1998). Formally, the first result says that for any regular estimator for τ^P , denoted by $\hat{\tau}$ with

$$\sqrt{N} \cdot (\hat{\tau} - \tau^P) \xrightarrow{d} \mathcal{N}(0, V),$$

we have

$$V \geq \mathbb{E} \left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} + (\tau(X) - \tau)^2 \right].$$

Knowledge of the propensity score does not affect this efficiency bound.

Hahn also shows that asymptotically linear estimators exist with such variance, and hence such efficient estimators can be approximated as

$$\hat{\tau} = \tau^P + \frac{1}{N} \sum_{i=1}^N \psi(Y_i, W_i, X_i, \tau^P) + o_p(N^{-1/2}),$$

where $\psi(\cdot)$ is the efficient score:

$$\psi(y, w, x, \tau^P) = \left(\frac{wy}{e(x)} - \frac{(1-w)y}{1-e(x)} \right) - \tau^P - \left(\frac{\mu_1(x)}{e(x)} + \frac{\mu_0(x)}{1-e(x)} \right) \cdot (w - e(x)).$$

Hahn (1998) also reports the efficiency bound for τ_T^P , both with and without knowledge of the propensity score. For τ_T^P the efficiency bound given knowledge of $e(X)$ is

$$\mathbb{E} \left[\frac{e(X)\text{Var}(Y(1)|X)}{\mathbb{E}[e(X)]} + \frac{e(X)^2\text{Var}(Y(0)|X)}{\mathbb{E}[e(X)]^2(1-e(X))} + (\tau(X) - \tau_T^P)^2 \frac{e(X)^2}{\mathbb{E}[e(X)]^2} \right].$$

If the propensity score is not known, the efficiency bound for τ_T^P is affected, unlike the bound for τ^P . For τ_T^P the bound without knowledge of the propensity score is

$$\mathbb{E} \left[\frac{e(X)\text{Var}(Y(1)|X)}{\mathbb{E}[e(X)]} + \frac{e(X)^2\text{Var}(Y(0)|X)}{\mathbb{E}[e(X)]^2(1-e(X))} + (\tau(X) - \tau_T^P)^2 \frac{e(X)}{\mathbb{E}[e(X)]^2} \right],$$

which is higher by

$$\mathbb{E} \left[(\tau(X) - \tau_T^P)^2 \cdot \frac{e(X)(1-e(X))}{\mathbb{E}[e(X)]^2} \right].$$

The intuition that knowledge of the propensity score affects the efficiency bound for the average effect for the treated (PATT), but not for the overall average effect (PATE) goes as follows. Both are weighted averages of the treatment effect conditional on the covariates, $\tau(x)$. For PATE the weight is proportional to the density of the covariates. For PATT the weight gets multiplied by the propensity score (e.g., Hirano, Imbens, and Ridder, 2002). For PATT one can therefore estimate this weight function, but this increases the variance. Knowledge of the propensity score implies one does not need to estimate the weight function.

2.6 Efficiency Bounds and Asymptotic Variances for Sample Average Treatment Effects

Consider the leading term of the efficient estimator for PATE, $\tilde{\tau} = \tau^P + \bar{\psi}$, where $\bar{\psi} = (1/N) \sum \psi(Y_i, W_i, X_i, \tau^P)$, and let us view this as an estimator for the sample average treatment effect SATE instead of as an estimator for the population average PATE. I will show that, first, this estimator is unbiased conditional on the covariates and the potential outcomes, and second, it has lower variance than when viewed as an estimator of PATE. To see that the estimator is unbiased note that

$$\mathbb{E}[\psi(Y, W, X, \tau^P | Y(0), Y(1), X)] = Y(1) - Y(0) - \tau^P.$$

Hence

$$\begin{aligned} & \mathbb{E}[\tilde{\tau} - \tau^S | (Y_i(0), Y_i(1), X_i)_{i=1}^N] \\ &= \mathbb{E} \left[\tau^P + \frac{1}{N} \sum_{i=1}^N \psi(Y_i, W_i, X_i, \tau^P) \middle| (Y_i(0), Y_i(1), X_i)_{i=1}^N \right] - \tau^S \end{aligned}$$

$$= \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) - \tau^S = 0.$$

Next, consider the normalized variance:

$$V^P = N \cdot \mathbb{E} \left[(\tilde{\tau} - \tau^S)^2 \right] = N \cdot \mathbb{E} \left[(\bar{\psi} + \tau^P - \tau^S)^2 \right].$$

Note that the variance of $\tilde{\tau}$ as an estimator of τ^P is

$$N \cdot \mathbb{E} \left[(\tilde{\tau} - \tau^P)^2 \right] = N \cdot \mathbb{E} [(\bar{\psi})^2] = N \cdot \mathbb{E} \left[\left(\bar{\psi}(Y, W, X, \tau^P) + (\tau^P - \tau^S) - (\tau^P - \tau^S) \right)^2 \right].$$

Because

$$\mathbb{E} \left[\left(\bar{\psi}(Y, W, X, \tau^P) + (\tau^P - \tau^S) \right) \cdot (\tau^P - \tau^S) \right] = 0,$$

as follows by using iterated expectations, first conditioning on X , $Y(0)$ and $Y(1)$, it follows that

$$\begin{aligned} N \cdot \mathbb{E} \left[(\tilde{\tau} - \tau^P)^2 \right] &= N \cdot \mathbb{E} \left[(\tilde{\tau} - \tau^S)^2 \right] + N \cdot \mathbb{E} \left[(\tau^S - \tau^P)^2 \right] \\ &= N \cdot \mathbb{E} \left[(\tilde{\tau} - \tau^S)^2 \right] + N \cdot \mathbb{E} \left[(Y(1) - Y(0) - \tau^P)^2 \right] \end{aligned}$$

Thus, the same estimator that as an estimator of the population average treatment effect τ^P has a normalized variance equal to V^P , has the property, as an estimator of τ^S :

$$\sqrt{N}(\tilde{\tau} - \tau^S) \xrightarrow{d} \mathcal{N}(0, V^S),$$

with

$$V^S = V^P - \mathbb{E}[Y(1) - Y(0) - \tau]^2.$$

As an estimator of τ^S the variance of $\tilde{\tau}^P$ is lower than its variance as an estimator of τ^P , with the difference equal to the variance of the treatment effect.

This raises a number of issues. First of all, a researcher is forced to take a stance on what the quantity of interest is. For example, in a specific application one can legitimately reach the conclusion that there is no evidence at the 95% level that the population average treatment effect is different from zero, whereas there is compelling evidence that the average treatment effect in the sample is positive. Typically researchers in economics have focused on the population average treatment effect, but it can certainly be argued that it is of interest, in cases where one cannot ascertain the sign of the population average treatment effect, to know whether one can ascertain the sign of the effect for the sample. Especially in cases, which are all too common, where it is not clear whether the sample is representative of the population of interest, reporting results for the sample at hand may be entirely appropriate.

An example to illustrate this point may be helpful. Suppose that $X \in \{0, 1\}$, with $\Pr(X = 1) = p_x$, and $\Pr(W = 1|X) = 1/2$. Suppose that $\tau(x) = 2x - 1$, and $\sigma_w^2(x)$ is very small for all x and w . In that case the average treatment effect is $p_x \cdot 1 + (1 - p_x) \cdot (-1) = 2p_x - 1$.

The efficient estimator in this case, not assuming knowledge beyond unconfoundedness, is to estimate $\tau(x)$ for $x = 0, 1$ and average these two up by the empirical distribution of X . The variance of $\sqrt{N}(\hat{\tau} - \tau^S)$ will be small because $\sigma_w^2(x)$ is small. The variance of $\sqrt{N}(\tau - \tau^P)$ will be larger by $4p_x(1 - p_x)$. If p_x differs from $1/2$, and so PATE differs from zero, the confidence interval for PATE will in small samples still tend to include zero, whereas with N odd (and both N_0 and N_1 at least equal to 2 so one can actually estimate $\sigma_w^2(x)$) and $\sigma_w^2(x)$ small enough, the standard confidence interval for τ^S will exclude zero with probability one. The intuition is that τ^P is much more uncertain because it depends on the distribution of the covariates, whereas the uncertainty about τ^S depends on only the conditional outcome variances and the propensity score.

The second issue is how to estimate the variance of the sample average treatment effect. Specific estimators for the variance will be discussed in Section 4, but here some general issues will be addressed. Because the two potential outcomes are never observed together for the same unit, one cannot directly infer the variance of the treatment effect. This is the same issue as the non-identification of the correlation coefficient. One can, however, estimate a lower bound on the variance of the treatment effect, leading to an upper bound on the variance of the estimator of the sample average treatment effect. Decomposing the variance as

$$\begin{aligned} \mathbb{E}[(Y(1) - Y(0) - \tau^P)^2] &= \mathbb{V}(\mathbb{E}[Y(1) - Y(0) - \tau|X]) + \mathbb{E}[\mathbb{V}(Y(1) - Y(0) - \tau|X)], \\ &= \mathbb{V}(\tau(X) - \tau) + \mathbb{E}[\sigma_1^2(X) + \sigma_0^2(X) - 2\rho(X)\sigma_0(X)\sigma_1(X)], \end{aligned}$$

we can consistently estimate the first term, but generally say little about the second term other than that it is nonnegative. One can therefore bound the variance of $\tilde{\tau} - \tau^S$ from above by

$$\begin{aligned} &\mathbb{E}[\psi(Y, W, X, \tau^P)^2] - \mathbb{E}[(Y(1) - Y(0)) - \tau^P]^2 \\ &\leq \mathbb{E}[\psi(Y, W, X, \tau^P)^2] - \mathbb{E}[(\tau(X) - \tau^P)^2] = \mathbb{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)}\right], \end{aligned}$$

and use this conservative variance to construct confidence intervals that are guaranteed to be conservative. Note the connection with Neyman's (1923) discussion of conservative confidence intervals for average treatment effects in experimental settings. It should be noted that the difference between these variances is of the same order as the variance itself, and not a small sample issue. Only when the treatment effect is known to be constant can it be ignored. Depending on the correlation between the outcomes and the covariates, this may change the standard errors considerably. It should also be noted that bootstrapping methods in general lead to estimation of $\mathbb{E}[(\tilde{\tau} - \tau^P)^2]$, rather than $\mathbb{E}[(\tilde{\tau} - \tau^S)^2]$.

3 Estimating Average Treatment Effects

There have been a number of estimators proposed for estimating the average treatment effects PATE and PATT. All of these are also appropriate estimators of the sample versions SATE and SATT. (The implications of the focus on SATE or SATT rather than PATE or PATT only arise when estimating the variance, and so I will return to this distinction in Section

4. In the current section all discussion applies to both pairs of estimands.) Here I review some of these estimators. They are organized in five groups. The first set, referred to as “regression estimators”, consists of estimators that rely on consistent estimation of the two conditional regression functions, $\mu_0(x)$ and $\mu_1(x)$. They differ in the way they estimate these functions, but all rely on estimators that are consistent for these regression functions. The second set of “matching” estimators forms matches between treated and control units. Each unit is matched to a fixed number of units with the opposite treatment. Within these paired sets treatment effects are estimated by differencing average outcomes. The bias of these within-pair estimates for the average treatment effect disappears as the sample size increases, although their variance does not go to zero since the number of matches remains fixed. The third set of “propensity score” estimators is characterized by a central role for the propensity score. This can be the estimated or the true propensity score, the latter obviously only feasible in cases where the propensity score is known. Three leading approaches in this set are weighting by the inverse of the propensity score, blocking on the propensity score, or regression on the propensity score. The fourth category consists of estimators that rely on a combination of these methods. Typically regression is combined with one of the other methods, e.g., matching with additional regression adjustment, or weighting or blocking with additional regression. The motivation for these methods is that although in principle any one of these methods can remove all the bias associated with the covariates, combining two of the methods may lead to more robust inference. For example, matching leads to consistent estimators for average treatment effects under weak conditions, so matching and regression can combine some of the good variance properties of regression methods with the consistency of matching. Similarly a combination of weighting and regression, using parametric models for both the propensity score and the regression functions, may lead to an estimator that is consistent as long as only one of the models is correctly specified (doubly robust in the terminology of Robins and Ritov, 1997). Finally, in the fifth group I will discuss Bayesian approaches to inference for average treatment effects.

Some of the estimators that will be discussed below achieve the semiparametric efficiency bound, whereas others will not. This does not mean, however, that the former are necessarily to be preferred in practice, that is, in finite samples. More generally, the debate concerning the practical advantages of the various estimators, and the settings in which some have more attractive properties than others, is still ongoing, and no firm conclusions have been reached so far. Although all estimators implicitly or explicitly estimate the two unknown regression functions and the propensity score, they do so in very different ways. Differences in smoothness of the regression function or the propensity score, or relative discreteness of the covariates in specific applications may affect the relative ranking of the estimators.

In addition, even the appropriateness of the standard asymptotic distributions as a guide towards finite sample performance is still debated, see for example Robins and Ritov (1997), and Angrist and Hahn (2001). A key feature of the problem that casts doubt on the relevance of the asymptotic distributions is the fact that the root- N consistency is obtained by averaging a nonparametric estimator of a regression function, which has itself a slow nonparametric convergence rate, over the empirical distribution of its argument. The dimension of this argument

affects the rate of convergence for the unknown function, but not the rate of convergence for the average treatment effect. In practice, however, the resulting approximations can be poor if the argument is of high dimension. In this case the presence of information about the propensity score is of particular relevance. Hahn (1998) showed that for the standard asymptotic distributions knowledge of the propensity score is irrelevant, and conditioning only on the propensity score is in fact not as efficient as conditioning on all covariates. Nevertheless, conditioning on the propensity score involves only one dimensional nonparametric regression, suggesting that the asymptotic approximations may be more accurate in that case.

Another issue that is important in judging the various estimators is how well they account for non- or limited overlap in the covariate distributions in the two treatment groups. If there are regions in the covariate space with little overlap (the propensity score close to zero or one), average treatment effect estimators should have relatively high variance. This is not always the case for estimators based on tightly parametrized models for the regression functions where outliers in terms of covariate values can lead to spurious precision for regression parameters. Regions of limited overlap can also be difficult to detect directly in high-dimensional covariate spaces as it can be masked for any single covariate.

3.1 Regression

The first class of estimators relies on consistent estimation of $\mu_w(x)$ for $w = 0, 1$. Given estimators $\hat{\mu}_w(x)$ for these regression functions the PATE and SATE are estimated by averaging their difference over the empirical distribution of the covariates:

$$\hat{\tau}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^N \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right).$$

In most implementations the average of the predicted treated outcome for the treated is equal to the average observed outcome for the treated, and similarly for the controls, so that this estimator can also be written as

$$\frac{1}{N} \sum_{i=1}^N W_i \cdot \left(Y_i - \hat{\mu}_0(X_i) \right) + (1 - W_i) \cdot \left(\hat{\mu}_1(X_i) - Y_i \right).$$

For the PATT and SATT typically only the control regression function is estimated, with the averaging over the difference between the actual outcomes for the treated outcome and estimated outcomes under the control treatment:

$$\hat{\tau}_{\text{reg},T} = \frac{1}{N_T} \sum_{i=1}^N W_i \cdot \left(Y_i - \hat{\mu}_0(X_i) \right). \tag{3.1}$$

Early estimators for $\mu_w(x)$ included parametric regression functions, for example linear regression, e.g., Rubin (1977). Such parametric estimators include least squares estimators with the regression function specified as

$$\mu_w(x) = \beta'x + \tau \cdot w,$$

in which case the average treatment effect is equal to τ . More recently nonparametric estimators have been proposed. Hahn (1998) proposes estimating first the three conditional expectations $g_1(x) = \mathbb{E}[WY|X]$, $g_0(x) = \mathbb{E}[(1 - W)Y|X]$, and $e(x) = \mathbb{E}[W|X]$ nonparametrically using series methods. He then estimates $\mu_w(x)$ as

$$\hat{\mu}_1(x) = \frac{\hat{g}_1(x)}{\hat{e}(x)}, \quad \hat{\mu}_0(x) = \frac{\hat{g}_0(x)}{1 - \hat{e}(x)},$$

and shows that the estimators for both PATE and PATT achieve the semiparametric efficiency bounds discussed in Section 2.6 (the latter in the unknown propensity score case).

It is unnecessary in this series approach to estimate all three of these conditional expectations $\mathbb{E}[Y|X]$, $\mathbb{E}[Y(1 - W)|X]$, and $\mathbb{E}[W|X]$ and use those to estimate $\mu_w(x)$. Instead one can directly estimate the two regression functions $\mu_w(x)$ using series methods. This has the advantage of eliminating the need to estimate the propensity score.

Heckman, Ichimura and Todd (1997, 1998ab) consider kernel methods for estimating $\mu_w(x)$. In particular they focus on local linear methods. The simple kernel estimator has the form

$$\hat{\mu}_w(x) = \frac{\sum_{i:W_i=w} Y_i \cdot K((X_i - x)/h)}{\sum_{i:W_i=w} K((X_i - x)/h)},$$

with a kernel $K(\cdot)$ and bandwidth h . In the local linear kernel regression the regression function $\mu_w(x)$ is estimated as intercept β_0 in the minimization problem

$$\min_{\beta_0, \beta_1} \sum_{i:W_i=w} \left(Y_i - \beta_0 - \beta_1 \cdot (X_i - x) \right)^2 \cdot K \left(\frac{X_i - x}{h} \right).$$

In order to control the bias of their estimators Heckman, Ichimura and Todd (1998) require that the order of the kernel is at least as large as the dimension of the covariates (that is, they require the use of a kernel function $K(z)$ such that $\int_z z^r K(z) dz = 0$ for $r \leq \dim(X)$, so that the kernel must be negative on part of the range, and the implicit averaging involves negative weights). We shall see this role for the dimension of the covariate return later again in other estimators.

For the average treatment effect for the treated, PATT, it is important to note that with the propensity score known, the estimator given in (3.1) is generally not efficient irrespective of the nature of the estimator for $\mu_0(x)$. Intuitively, the reason is that with the propensity score known the average $\sum W_i Y_i / N_T$ is not efficient for the population expectation $\mathbb{E}[Y(1)|W = 1]$. An efficient estimator (e.g., Hahn, 1998) can be obtained by weighting all the estimated treatment effects $\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$ by the probability of receiving the treatment:

$$\tilde{\tau}_{\text{reg},T} = \frac{\sum_{i=1}^N e(X_i) \cdot (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))}{\sum_{i=1}^N e(X_i)}. \quad (3.2)$$

In other words, instead of estimating $\mathbb{E}[Y(1)|W = 1]$ as $\sum W_i Y_i / N_T$ using only the treated observations, it is estimated using all observations as $\sum \hat{\mu}_1(X_i) \cdot e(X_i) / \sum e(X_i)$. Knowledge of the propensity score improves the accuracy because it allows one to exploit the control observations to adjust for imbalances in the sampling of the covariates.

For all the estimators in this section an important issue is the choice of the smoothing parameter. In Hahn’s case that is the number of terms in the series, after having chosen the form of the series and the sequence. In the Heckman, Ichimura, Todd case this is, after choosing the form of the kernel, the bandwidth. The evaluation literature has been largely silent concerning the optimal choice of the smoothing parameters. The larger literature on nonparametric estimation of regression functions does provide some guidance, and offers data driven methods such as cross-validation criteria. The optimality properties of these criteria, however, are for estimation of the entire function, $\mu_w(x)$ in this case. Typically the focus is on mean-integrated-squared-error criteria of the form $\int_x (\hat{\mu}_w(x) - \mu_w(x))^2 f_X(x) dx$, with possibly an additional weight function. In the current problem one is interested specifically in the average treatment effect, and so such integrated mean-squared-error criteria are not necessarily optimal. In particular global smoothing parameter choices may be inappropriate because they may be driven by the shape of the regression function and distribution of covariates in regions where these are not important for the average treatment effect of interest. Lalonde’s (1986) data set is a well known example of this where much of probability mass of the distribution of non-experimental control group is in a region with moderate to high earnings where few of the treated group located. There is little evidence whether results for average treatment effects are more or less sensitive to the smoothing parameter choices than results for estimation of the regression functions themselves.

3.2 Matching

Regression estimators impute the missing potential outcomes using the estimated regression function. Thus, if $W_i = 1$, $Y_i(0)$ is missing and imputed with a consistent estimator for the conditional expectation $\mu_0(X_i)$. Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of the nearest neighbours. In that sense matching is similar to nonparametric kernel regression methods, with the number of neighbors playing the role of the bandwidth in the kernel regression. A formal difference is that the asymptotic distribution is derived conditional on the implicit bandwidth, that is, the number of neighbours, which is often fixed at one. Using such asymptotics, the implicit estimate $\hat{\mu}_w(x)$ is (close to) unbiased, but not consistent for $\mu_w(x)$. In contrast, the regression estimators discussed in the previous section relied on consistency of $\mu_w(x)$. Matching estimators have the attractive feature that given the matching metric, the researcher only has to choose the number of matches. For the regression estimators discussed in the previous section the researcher needs to choose smoothing parameters that are more difficult to interpret, either the number of terms in a series or the bandwidth in kernel regression. Obviously within the class of matching estimators using only a single match leads to the most credible inference with the least bias, at most giving up some efficiency. This can make the matching estimator easier to use than some of the other estimators that require more complex choices of smoothing parameters and may explain some of its popularity.

Matching estimators have been widely studied in practice and theory (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1989, 1995; Rubin, 1973b, Heckman, Ichimura and Todd, 1998; Dehejia and Wahba, 1999; Abadie and Imbens, 2002). Most often they have been applied in

settings with the following two characteristics: (i) the interest is in the average treatment effect for the treated, (ii), there is a large reservoir of potential controls. This allows the researcher to match each treated unit to one or more distinct controls (referred to as matching without replacement). Given the matched pairs, the treatment effect within a pair is then estimated as the difference in outcomes, with an estimator for the PATE for the treated obtained by averaging these within-pair differences. Since the estimator is essentially the difference in two sample means the variance is calculated using standard methods for differences in means or methods for paired randomized experiments. The remaining bias is typically ignored in these studies. The literature has studied fast algorithms for matching the units as fully efficient matching methods are computationally cumbersome (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1995). Note that in such matching schemes the order in which the units are matched is typically important.

Abadie and Imbens (2001) study both bias and variance in a more general setting where potentially both treated and control units are matched and matching is done with replacement, as in Dehejia and Wahba (1999). The Abadie-Imbens estimator is implemented in Matlab and STATA (see Abadie, Drukker, Herr, and Imbens, 2003). Formally, given a sample, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $\ell_m(i)$ be the index l that solves $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1 \left\{ \|X_j - X_i\| \leq \|X_{\ell_m(i)} - X_i\| \right\} = m,$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $\ell_m(i)$ is the index of the unit that is the m -th closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$, among the units with the treatment opposite to that of unit i . In particular, $\ell_1(i)$ is the nearest match for unit i . Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i : $\mathcal{J}_M(i) = \{\ell_1(i), \dots, \ell_M(i)\}$. Define the imputed potential outcomes as:

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

and

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

The simple matching estimator Abadie and Imbens study is

$$\hat{\tau}_M^{sm} = \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i(1) - \hat{Y}_i(0) \right). \quad (3.3)$$

They show that the bias of this estimator is of order $O(N^{-1/K})$, where K is the dimension of the covariates. Hence, if one studies the asymptotic distribution of the estimator by normalizing by \sqrt{N} , as can be justified by the fact that the variance of the estimator is of order $O(1/N)$, the bias does not disappear if the dimension of the covariates is equal to two, and will dominate, in large samples, the large sample variance if the dimension of the covariates is at least three.

Three caveats to the Abadie-Imbens result should be pointed out. First, it is only the continuous covariates that should be counted in this dimension. With discrete covariates the matching will be exact in large samples, and such covariates do not contribute to the order of the bias. Second, if one matches only the treated, and the number of potential controls is much larger than the number of treated units, one can justify ignoring the bias by appealing to an asymptotic sequence where the number of potential controls increases faster than the number of treated units. Specifically, if the number of controls N_0 and the number of treated N_1 satisfy $N_1/N_0^{4/K} \rightarrow 0$, then the bias disappears in large samples after normalization by $\sqrt{N_1}$. Third, even though the order of the bias may be high, the actual bias may still be small if the coefficients in the leading term are small. This is possible if the biases for different units are at least partially offsetting. For example, the leading term in the bias relies on the regression function being nonlinear, and the density of the covariates having a nonzero slope. If one of the two is at least close to being satisfied, the resulting bias may be fairly limited. To remove the bias Abadie and Imbens suggest combining the matching with regression adjustment. I will discuss this modification in Section 3.4.3

Another point made by Abadie and Imbens is that matching estimators are generally not efficient. Even in the case where the bias is of low enough order so it gets dominated by the variance, the estimators are not efficient given a fixed number of matches. To reach efficiency one would need to increase the number of matches with the sample size. If $M \rightarrow \infty$, with $M/N \rightarrow 0$, then the matching estimator is essentially like a regression estimator, with the imputed missing potential outcomes consistent for their conditional expectations. However, the efficiency gain of such estimators is of course somewhat artificial. If in a given data set one uses M matches, one can calculate the variance as if this number of matches increases at the appropriate rate with the sample size, in which case the estimator would be efficient, or one could calculate the variance conditional on the number of matches, in which case the same estimator would not be efficient. Little is known yet about the optimal number of matches, or about data-dependent ways of choosing the number of matches.

In the above discussion the distance metric in choosing the optimal matches was the standard Euclidan metric:

$$d_E(x, z) = (x - z)'(x - z).$$

In practice the metrics implicitly standardize the covariates in some manner. Abadie and Imbens use the diagonal matrix with the inverse of the covariate variances on the diagonal:

$$d_{AI}(x, z) = (x - z)' \text{diag}(\Sigma_X^{-1})(x - z),$$

where Σ_X is the covariance matrix of the covariates. The most common choice is the Mahalanobis metric (e.g., Rosenbaum and Rubin, 1985) which uses the inverse of the covariance matrix of the pretreatment variables:

$$d_M(x, z) = (x - z)' \Sigma_X^{-1} (x - z).$$

This metric has the attractive property that it reduces differences in covariates within matched pairs in all directions. See for more formal discussions Rubin and Thomas (1992).

Zhao (2002), in an interesting discussion of the choice of metrics, suggests some alternative metrics that depend on the correlation between covariates and treatment assignment and outcomes. He starts by assuming that the propensity score has a logistic form

$$e(x) = \frac{\exp(x'\gamma)}{1 + \exp(x'\gamma)},$$

and the regression functions are linear:

$$\mu_w(x) = \alpha_w + x'\beta.$$

He then considers two alternative metrics. The firsts weights absolute differences in the covariates by the coefficient in the propensity score:

$$d_{Z1}(x, z) = \sum_{k=1}^K |x_k - z_k| \cdot |\gamma_k|,$$

and the second weights them by the coefficients in the regression function:

$$d_{Z2}(x, z) = \sum_{k=1}^K |x_k - z_k| \cdot |\beta_k|,$$

where x_k and z_k are the k th elements of the K vectors x and z respectively.

In the light of this discussion, it is interesting to consider optimality of the metric. Suppose, following Zhao (2002), that the regression functions are linear with coefficients β_w . Now consider a treated unit with covariate vector x who will be matched to a control unit with covariate vector z . The bias resulting from such a match is $(z - x)'\beta_0$. If one is interested in minimizing for each match the squared bias, one should choose the first match by minimizing over the control observations $(z - x)'\beta_0\beta_0'(z - x)$. Typically one does not know the value of the regression coefficients. In that case one may wish to minimize expected squared bias. Using a normal distribution for the regression errors, and a flat prior on β_0 , the posterior distribution for β_0 is normal with mean $\hat{\beta}_0$ and variance $\Sigma_X^{-1}\sigma^2/N$. Hence the expected squared bias from a match is

$$\mathbb{E} [(z - x)'\beta_0\beta_0'(z - x)] = (z - x)' \left(\hat{\beta}_0\hat{\beta}_0' + \sigma^2\Sigma_X^{-1}/N \right) (z - x).$$

The optimal metric is in this argument a combination of the sample covariance matrix plus the outer product of the regression coefficients, with the former scaled down by a factor $1/N$:

$$d^*(z, x) = (z - x)' \left(\hat{\beta}_w\hat{\beta}_w' + \sigma_w^2\Sigma_{X,w}^{-1}/N \right) (z - x),$$

A clear problem with this approach is that when the regression function is misspecified, matching with this particular metric may not lead to a consistent estimator. On the other hand, when the regression function is correctly specified, it would be more efficient to use the regression estimators. In practice one may want to use a metric that combines some of the optimal weighting with some safeguards in case the regression function is misspecified.

So far there is little experience with any alternative metrics beyond the Mahalanobis metric. Zhao (2002) reports on the results of some simulations using his proposed metrics in a setting with two covariates and finds that in his specific designs there is no clear winner, although his findings suggest that using the outcomes in defining the metric is a promising approach.

3.3 Propensity Score Methods

Since the work by Rosenbaum and Rubin (1983a) there has been considerable interest in methods that avoid adjusting directly for all covariates and instead focus on adjusting for differences in the propensity score, the conditional probability of receiving the treatment. This can be implemented in a number of different ways. One can weight the observations to create balance between treated and control units in the weighted sample in terms of the propensity score (and indirectly also in terms of the covariates). Hirano, Imbens and Ridder show how such estimators can achieve the semiparametric efficiency bound. Alternatively one can divide the sample into subsamples with approximately the same value of the propensity score, a technique known as blocking. Finally, one can directly use the propensity score as a regressor in a regression approach.

In practice there are two important cases. First, suppose the researcher knows the propensity score. In that case all three of these methods are likely to be effective. Even if the resulting estimator is not fully efficient, one can easily modify them to capture most of the efficiency loss and the fact that the estimators do not rely on high-dimensional nonparametric regression suggests that their finite sample properties are likely to be relatively attractive.

If the propensity score is not known, it is less clear what the advantages are of the estimators discussed in this section. Although they avoid the high-dimensional nonparametric regression of the two conditional expectations $\mu_w(x)$, they require instead the equally high-dimensional nonparametric regression of the treatment indicator on the regressors. In practice the relative merits of these estimators will depend on whether the propensity score is more or less smooth than the regression functions $\mu_w(x)$, or whether additional information is available about some of them.

3.3.1 Weighting

Simply taking the difference in average outcomes for treated and controls,

$$\hat{\tau} = \frac{\sum W_i Y_i}{\sum W_i} - \frac{\sum (1 - W_i) Y_i}{\sum 1 - W_i},$$

is not unbiased for $\tau^P = \mathbb{E}[Y(1) - Y(0)]$ because conditional on the treatment indicator the distributions of the covariates differ. By weighting the units by the inverse of the probability of receiving the treatment they got, one can undo this imbalance. Formally, weighting estimators rely on the equality:

$$\mathbb{E} \left[\frac{WY}{e(X)} \right] = \mathbb{E} \left[\frac{WY(1)}{e(X)} \right] = \mathbb{E} \left[\mathbb{E} \left[\frac{WY(1)}{e(X)} \middle| X \right] \right] = \mathbb{E} \left[\mathbb{E} \left[\frac{e(X)Y(1)}{e(X)} \right] \right] = \mathbb{E}[Y(1)],$$

and similarly

$$\mathbb{E} \left[\frac{(1 - W)Y}{1 - e(X)} \right] = \mathbb{E}[Y(0)],$$

Hence

$$\tau^P = \mathbb{E} \left[\frac{W \cdot Y}{e(X)} - \frac{(1 - W) \cdot Y}{1 - e(X)} \right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N} \sum_{i=1}^N \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right). \quad (3.4)$$

In this particular form this is not necessarily a very attractive estimator. The main reason is that although the estimator can be written as the difference between a weighted average of the outcomes for the treated units and a weighted average of the outcomes for the controls, the weights do not necessarily add up to one. Specifically, in (3.4), the weights for the treated units add up to $(\sum W_i / e(X_i)) / N$. In expectation this is equal to one, but since its variance is positive, in any given sample some of the weights are likely to deviate from one. The estimator can be improved simply by normalizing the weights to unity. This idea of normalizing the weights can be taken further, by normalizing the weights to unity within subpopulations defined by the covariates. In the limit this leads to an estimator proposed by Hirano, Imbens and Ridder (2002) who suggest using a nonparametric series estimator for $e(x)$. More precisely, they specify a sequence of functions of the covariates, e.g., a power series, $h_l(x)$, $l = 1, \dots, \infty$. Next, they choose a number of terms $L(N)$ as a function of the sample size, and then estimate the L -dimensional vector γ_L in

$$\Pr(W = 1 | X = x) = \frac{\exp((h_1(x), \dots, h_L(x))\gamma_L)}{1 + \exp((h_1(x), \dots, h_L(x))\gamma_L)},$$

by maximizing the associated likelihood function. Let γ_L be the maximum likelihood estimates. Third, the estimated propensity score is calculated as:

$$\hat{e}(x) = \frac{\exp((h_1(x), \dots, h_L(x))\hat{\gamma}_L)}{1 + \exp((h_1(x), \dots, h_L(x))\hat{\gamma}_L)}.$$

Finally they estimate the average treatment effect as:

$$\hat{\tau}_{\text{weight}} = \sum_{i=1}^N \frac{W_i \cdot Y_i}{\hat{e}(X_i)} - \sum_{i=1}^N \frac{(1 - W_i) \cdot Y_i}{1 - \hat{e}(X_i)}.$$

They show that with a nonparametric estimator for $e(x)$ this estimator is efficient, whereas with the true propensity score the estimator would not be fully efficient, and in fact not even very attractive.

This estimator highlights one of interesting features of the problem of efficiently estimating average treatment effects. One solution is to estimate the two regression functions $\mu_w(x)$ nonparametrically, as discussed in Section 3.1. That approach completely ignores the propensity score. A second approach is to estimate the propensity score nonparametrically, ignoring entirely the two regression functions. If appropriately implemented both approaches lead to fully efficient estimators, but clearly their finite sample properties may well be very different, depending for example, on the smoothness of the regression functions versus the smoothness of the propensity score. If there is only a single binary covariate, or more generally with only discrete covariates, the weighting estimator with a fully nonparametric estimator for the propensity

score is numerically identical to the regression estimator with a fully nonparametric estimator for the two regression functions.

To estimate the average treatment effect for the treated, one should weight both components by the propensity score. Here a difference again arises between the case where the propensity score is known and the case where it is unknown. In the unknown case an efficient estimator can be written as:

$$\hat{\tau}_{\text{weight,tr}} = \frac{1}{N_1} \sum_{i:W_i=1} Y_i - \sum_{i:W_i=0} Y_i \cdot \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} / \sum_{i:W_i=0} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)}.$$

In the known propensity score case the weights should be equal to the true propensity score:

$$\begin{aligned} \hat{\tau}_{\text{weight,tr}} &= \sum_{i=1}^N W_i \cdot Y_i \cdot \frac{e(X_i)}{\hat{e}(X_i)} / \sum_{i=1}^N W_i \frac{e(X_i)}{\hat{e}(X_i)} \\ &\quad - \sum_{i=1}^N (1 - W_i) \cdot Y_i \cdot \frac{e(X_i)}{1 - \hat{e}(X_i)} / \sum_{i=1}^N (1 - W_i) \frac{e(X_i)}{1 - \hat{e}(X_i)}. \end{aligned}$$

One difficulty with the weighting estimators that use the estimated propensity score is the problem of choosing the smoothing parameters, which also plagues the regression estimators in section 3.1. Hirano, Imbens and Ridder (2002) uses series estimators which requires choosing the number of terms in the series. Ichimura and Linton (2001) consider a kernel version of this which involves choosing a bandwidth. There is currently one of the few studies considering optimal choices for smoothing parameters focusing specifically on estimating average treatment effects.

A difference with standard problems of choosing smoothing parameters is that here one wants to use nonparametric regression methods even if the propensity score is known. For example, if the propensity score is constant, standard optimality results would suggest using a high degree of smoothing, as this would lead to the most accurate estimator for the function to be estimated. However, this would not necessarily lead to an efficient estimator for the average treatment effect of interest.

3.3.2 Blocking on the Propensity Score

In their original propensity score paper Rosenbaum and Rubin (1983a) suggest the following estimator. Take the propensity score and divide the sample into M blocks by the value of the (estimated) propensity score. Let J_{im} be an indicator for unit i being in block m . One way of implementing this is by dividing the unit interval into M blocks with boundary values equal to m/M for $m = 1, \dots, M - 1$, so that

$$J_{im} = 1\{(m - 1)/M < e(X_i) \leq m/M\},$$

for $m = 1, \dots, M$. Within each block there are N_{wm} observations with treatment equal to w , $M_{wm} = \sum_i 1\{W_i = w, J_{im} = 1\}$. Estimate within each block the average treatment effect as if

random assignment holds,

$$\hat{\tau}_m = \frac{1}{N_{1m}} \sum_{i=1}^N J_{im} W_i Y_i - \frac{1}{N_{0m}} \sum_{i=1}^N J_{im} (1 - W_i) Y_i.$$

Then estimate the overall average treatment effect as:

$$\hat{\tau}_{\text{block}} = \sum_{m=1}^M \hat{\tau}_m \cdot \frac{N_{1m} + N_{0m}}{N}.$$

If one is interested in the average effect for the treated, one would weight the within-block average treatment effects by the number of treated units:

$$\hat{\tau}_{T,\text{block}} = \sum_{m=1}^M \hat{\tau}_m \cdot \frac{N_{1m}}{N_T}.$$

Blocking can be interpreted as a crude form of nonparametric regression where the unknown function is approximated by a step function with fixed jump points. To establish asymptotic properties for this estimator would require establishing conditions on rate at which the number of blocks increases with the sample size. With the propensity score known these are easy to establish. No formal results have been established for the case with the unknown propensity score.

The question arises how many blocks to use in practice. Cochran (1968) analyses a case with a single covariate, and, assuming normality, shows that using five blocks removes at least than 95% of the bias associated with that covariate. Since all bias is under unconfoundedness associated with the propensity score, this suggests, under normality, that five blocks removes most of the bias associated with all the covariates. This has often been the starting point of empirical analyses using this estimator (e.g., Rosenbaum and Rubin, 1983b; Dehejia and Wahba, 1999). It has been implemented in STATA by Becker and Ichino (2002).³ Often, however, researchers subsequently check the balance of the covariates within the blocks. If within the blocks the true propensity score is constant, the distribution of the covariates among the treated and controls should be identical, or, in the evaluation terminology, the covariates should be balanced. Hence one can investigate the adequacy of the statistical model by comparing the distribution of the covariates among treated and controls within the blocks. If the distributions are different, one can either split the block into a number of subblocks, or one can generalize the specification of the propensity score. Often some informal version of the following algorithm is used: If within a block the propensity score itself is unbalanced, the blocks are too large and need to be split. If conditional on the propensity score being balanced the covariates are not balanced, the specification of the propensity score is not adequate. No formal algorithm exists for implementing these methods exists.

An alternative approach to finding the optimal number of blocks is to relate this estimator to the weighting estimator. One can view the blocking estimator as identical to a weighting

³Becker and Ichino also implement estimators that match on the propensity score.

estimator with a modified estimator for the propensity score. Specifically, given the original estimator for the propensity score, $\hat{e}(x)$, the estimator for the propensity score is discretized to

$$\tilde{e}(x) = \frac{1}{M} \sum_{m=1}^M 1\{(m/M) \leq \hat{e}(x)\}.$$

Using $\tilde{e}(x)$ as the propensity score in the weighting estimator leads to an estimator for the average treatment effect identical to that obtained by using the blocking estimator with $\hat{e}(x)$ as the propensity score and M blocks. With a sufficiently large number of blocks, the blocking estimator is sufficiently close to the original weighting estimator that it shares its first order asymptotic properties, including its efficiency. It suggests that in general there is little harm in choosing a relatively large number of blocks at least in terms of asymptotic properties, although again the relevance of this for finite samples has not been established.

3.3.3 Regression on the Propensity Score

The third method of using the propensity score is to estimate the conditional expectation of Y given W and $e(X)$. Define

$$\nu_w(e) = \mathbb{E}[Y(w)|e(X) = e].$$

By unconfoundedness this is equal to $\mathbb{E}[Y|W = w, e(X) = e]$. Given an estimator $\hat{\nu}_w(e)$, one can estimate the average treatment effect as

$$\hat{\tau}_{regprop} = \frac{1}{N} \sum_{i=1}^N (\hat{\nu}_1(e(X_i)) - \hat{\nu}_0(e(X_i))).$$

Hahn (1998) considers a series version of this estimator and shows that it is not as efficient as the regression estimator based on adjustment for all covariates. Heckman, Ichimura and Todd (1998) consider a local linear version of this for estimating the average treatment effect for the treated.

3.4 Mixed Methods

A number of methods have been proposed that combine two of the three methods described in the previous sections, typically regression with one of the others. The reason is that although one of the methods alone is often sufficient to obtain consistent or even efficient estimators, combining with regression may eliminate remaining bias and improve precision. This is particularly useful because neither matching nor the propensity score methods directly address the correlation between the covariates and the outcome. The robustness argument is made explicit in the notion developed by Robins and Ritov (1997) of double robustness. They propose a combination of weighting and regression where as long as either the parametric model for the propensity score or the parametric model for the regression functions is specified correctly the resulting estimator for the average treatment effect is consistent. Similarly, matching leads to consistency without additional assumptions, so methods that combine matching and regressions are robust against misspecification of the regression function.

3.4.1 Weighting and Regression

One can rewrite the weighting estimator discussed before as estimating the following regression function by weighted least squares,

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i,$$

with weights equal to

$$\lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1 - W_i}{1 - e(X_i)}}.$$

Without the weights the least squares estimator would not be consistent for the average treatment effect, but the weights ensure that the covariates are uncorrelated with the treatment indicator and hence the weighted estimator is consistent.

This weighted-least-squares representation suggests that one may add covariates to the regression function to improve precision, for example as

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i,$$

with the same weights λ_i .

Such an estimator, using a more general semiparametric regression model, is suggested in Robins and Rotnitzky (1995), Robins and Ritov (1997), and implemented in Hirano and Imbens (2002). In the parametric context Robins and Ritov argue that the estimator is consistent as long as either the regression model is specified correctly, or the propensity score (and thus the weights) is specified correctly. That is, in the Robins-Ritov terminology the estimator is doubly robust.

3.4.2 Blocking and Regression

Rosenbaum and Rubin (1983b) suggest modifying the basic blocking estimator by using least squares regression within the blocks. Without the additional regression adjustment the estimated treatment effect within the blocks can be written as a least squares estimator of τ_m for the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \varepsilon_i,$$

using only the units in block m . This representation suggests adding covariates by changing the regression function to

$$Y_i = \alpha_m + \beta'_m X_i + \tau_m \cdot W_i + \varepsilon_i,$$

again estimated on the units in block m .

3.4.3 Matching and Regression

Since Abadie and Imbens (2002) show that the bias of the simple matching estimator can dominate the variance if the dimension of the covariates is too high, additional bias corrections through regression can be particularly relevant in this case. A number of such corrections have been proposed, first by Rubin (1973b) in a parametric setting. Following the notation of Section 3.2, let $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ be the observed or imputed potential outcomes for unit i . These estimated potential outcomes equal observed outcomes for some units (for unit i and its match $\ell(i)$). The bias in their comparison, $\mathbb{E}[\hat{Y}_i(1) - \hat{Y}_i(0)] - (Y_i(1) - Y_i(0))$ arises from the fact that the covariates for units i and $\ell(i)$, X_i and $X_{\ell(i)}$ are not equal, although because of the matching, they will be close. To further explore this, define for each unit (focusing on the single match case):

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell(i)} & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{X}_i(1) = \begin{cases} X_{\ell(i)} & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

If the matching is exact, for each unit one has $\hat{X}_i(0) = \hat{X}_i(1)$. If, however, the matching is not exact there will be some discrepancies that lead to potential bias. The difference $\hat{X}_i(1) - \hat{X}_i(0)$ will therefore be used to reduce the bias of the simple matching estimator.

Suppose unit i is a treated unit with $W_i = 1$. In that case $\hat{Y}_i(0)$ is imputed value for $Y_i(0)$ for this unit. This imputed value is unbiased for $\mu_0(X_{\ell(i)})$, not necessarily for $\mu_0(X_i)$. One therefore may wish to adjust $\hat{Y}_i(0)$ by an estimate of $\mu_0(X_i) - \mu_0(X_{\ell(i)})$. Typically these corrections are taken to be linear in the covariates, that is, of the form $\beta_0'(X_i - X_{\ell(i)})$. Rubin (1973b) proposed three corrections which differ in the way β_0 is estimated. All three agree in the case where all the matching is exact, and no correction is required.

To introduce the first correction, note that one can write the matching estimator as the least squares estimator for the regression function

$$\hat{Y}_i(1) - \hat{Y}_i(0) = \tau + \varepsilon_i.$$

This representation suggests modifying the regression function to

$$\hat{Y}_i(1) - \hat{Y}_i(0) = \tau + (\hat{X}_i(1) - \hat{X}_i(0))'\beta + \varepsilon_i,$$

and again estimating τ by least squares.

The second correction is to take all control units, and estimate a linear regression of the form

$$Y_i = \alpha_0 + \beta_0'X_i + \varepsilon_i,$$

by least squares. Abadie and Imbens (2002) show that if this last correction is done nonparametrically, the resulting matching estimator is consistent and asymptotically normal, with its bias dominated by the variance.

The third method is to estimate the same regression function using only the controls that are used as matches, with weights corresponding to the number of times a control observations is used as a match. Compared to the second method this approach is potentially less efficient as it discards some control observations and weights some more than others. It has the advantage of only using the most relevant controls. The controls that are discarded in the matching are likely to be relative outliers, and they may therefore unduly affect the least squares estimates. If the regression function is in fact linear this may be an attractive feature, but if there is uncertainty about this, one may not wish to have such observations have considerable influence. This is one of the matching estimators considered by Abadie and Imbens (2002).

3.5 Bayesian Approaches

Little has been done using Bayesian methods, both in terms of methodology and in terms of applications. Rubin (1978) lays out the general approach from a Bayesian perspective. Dehejia (1997) studies the decision problem faced by a policy maker who has to assign heterogeneous individuals to various training programs with uncertain and heterogeneous effects, using a Bayesian approach.

There are no applications, however, that I am aware of which focus on estimating the average treatment effect either for the population or for the subpopulation of the treated under unconfoundedness using a Bayesian approach. Neither are there simulation studies comparing operating characteristics of Bayesian methods to the frequentist methods discussed in the earlier sections in this paper. Such a Bayesian approach can be easily implemented with the regression methods discussed in Section 3.1. Interestingly it is less clear how Bayesian methods would be used with pairwise matching, which does not appear to have a natural likelihood interpretation.

A Bayesian approach to the regression estimators may be useful for a number of reasons. One is that a leading problem is the presence of many covariates, relative to the number of observations. Standard frequentist methods tend to either include those covariates without any restriction, or exclude them entirely. Bayesian methods would allow researchers to include covariates with more or less informative prior distributions. For example, if the researcher has a number of lagged outcomes, one may expect recent lags to be more important in predicting future outcomes than long lags, which can be reflected in tighter prior distributions around zero for the longer lags. Alternatively with a number of similar covariates one may wish to use hierarchical models that avoid problems with large dimensional parameter spaces.

A second argument for considering Bayesian methods is that in a closely related area, that of missing data with the Missing At Random (MAR) assumption Bayesian methods have found widespread applicability. As advocated by Rubin (1987, 2001), multiple imputation methods often rely on Bayesian methods for imputing the missing data, taking account of the parameter heterogeneity in a manner consistent with the uncertainty in the missing data model itself. The same methods could be used with little modification for causal models, with the main complication that a relatively large proportion, namely 50% of the total potential outcomes, is missing.

A more formal argument is that estimating average causal effects is at heart a prediction problem. A number of potential outcomes are missing and they need to be predicted based

on the available information. As such this is a clear decision problem, and admissible decision rules should be close to Bayesian answers.

4 Estimating Variances

The variances of the estimators considered so far typically involve unknown functions. For example, the variance of efficient estimators of the population average treatment effect PATE is equal to

$$V^P = \mathbb{E} \left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} + (\mu_1(X) - \mu_0(X) - \tau)^2 \right],$$

involving the two regression functions, the two conditional variances and the propensity score.

There are a number of ways we can estimate this asymptotic variance. The first is essentially by brute force. All five components of the variance, $\sigma_0^2(x)$, $\sigma_1^2(x)$, $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, are consistently estimable using kernel methods or series, and hence the asymptotic variance itself is consistently estimable. However, if one estimates the average treatment effect by estimating only the two regression functions, it is an additional burden to estimate the conditional variances and the propensity score. Similarly, if one estimates the average treatment effect efficiently using weighting by the estimated propensity score, it is a considerable additional burden to estimate the first two moments of the conditional outcome distributions just to estimate the asymptotic variance.

A second method applies to the case where either the regression functions or the propensity score is estimated using series or sieves. In that case one can interpret the estimators, given the number of terms in the series, as parametric estimators, and calculate the variance that way. Under some conditions that will lead to valid standard errors and confidence intervals.

A third approach is to use bootstrapping. Given that the estimators are asymptotically linear, it is likely that the bootstrap will lead to valid standard errors and confidence intervals, although there is little formal evidence specific for these estimators. Certainly subsampling methods are likely to work. Bootstrapping may be more complicated for matching estimators, as the bootstrapping introduces discreteness in the distribution that will lead to ties in the matching algorithm. Care has to be applied in dealing with the ties to ensure the validity of the bootstrap.

These first three methods are all for estimating the variance for estimators of τ^P . As argued before, one may wish to estimate τ^S instead. In that case the appropriate (conservative) variance is

$$V^S = \mathbb{E} \left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1 - e(X)} \right].$$

This variance can be estimated again by estimating the conditional moments of the outcome distributions, with the same difficulties as before. It cannot be estimated by bootstrapping since the estimand itself changes from bootstrap sample to bootstrap sample. There is, however, an alternative method for estimating this variance that does not require additional nonparametric estimation. This method was developed in Abadie and Imbens (2002).

The idea behind the Abadie-Imbens matching variance estimator is that even though the asymptotic variance depends on the conditional variance $\sigma_w^2(x)$, one need not actually estimate this variance consistently at all values of x . Rather, one needs the average of this variance over the distribution of the covariates, weighted by the inverse of either $e(x)$ or its complement $1 - e(x)$. The key is therefore to obtain a close-to-unbiased estimator for the variance $\sigma_w^2(x)$. Suppose we can find two treated units with $X = x$, say units i and j . In that case an unbiased estimator for $\sigma_1^2(x)$ is

$$\hat{\sigma}_1^2(x) = (Y_i - Y_j)^2 / 2.$$

In general it is again going to be difficult to find exact matches. Nevertheless, this is again not necessary. Instead one uses the closest match, within the set of units with the same treatment indicator. Let $v_m(i)$ be the m th closest unit with the same treatment indicator, the solution to $W_{v_m(i)} = W_i$, and

$$\sum_{l|W_l=W_i} 1 \left\{ \|X_l - x\| \leq \|X_{v_m(i)} - x\| \right\} = m.$$

Given a fixed number of matches M this gives us M units with the same treatment indicator and approximately the same values for the covariates. The sample variance of these M units is used to estimate $\sigma_1^2(x)$.

We can do the same for the control variance function $\sigma_0^2(x)$. This gives us estimates of $\sigma_w^2(x)$ at all values of the covariates and, for $w = 0, 1$. Note that these are not consistent estimators. As the sample size increases, the bias of these estimators will disappear, the same way the bias of the matching estimator for the average treatment effect will disappear. The rate at which this bias disappears depends on the dimension of the covariates, but this is not important for the consistency of the estimator of the variance of the average treatment effect. The variance of the estimators for $\sigma_w^2(X_i)$ will not go to zero, however. This is not important, however, as we are interested not in the variances at specific values of the covariates, but in the variance of the average treatment effect. This is finally estimated as:

$$\hat{V}^S = \frac{1}{N} \sum_{i=1}^N \left(\frac{\hat{\sigma}_1^2(X_i)}{\hat{e}(X_i)} + \frac{\hat{\sigma}_0^2(X_i)}{1 - \hat{e}(X_i)} \right).$$

Under standard regularity conditions this is consistent for the asymptotic variance of the average treatment effect estimator.

5 Assessing the Assumptions

5.1 Indirect Tests of the Unconfoundedness Assumption

The unconfoundedness assumption is not directly testable. It states that the conditional distribution of the outcome under the control treatment, $Y(0)$, given receipt of the active treatment and given covariates, is identical to the distribution of the control outcome given receipt of the

control treatment and given covariates, and the same for the outcome given the active treatment $Y(1)$. Since the data are completely uninformative about the distribution of the control treatment outcome for those who received the active treatment, and about the distribution of the active treatment outcome given receipt of the control treatment, the data cannot directly reject the unconfoundedness assumption. Nevertheless, there are often indirect ways of assessing the unconfoundedness assumption in applications. A number of such tests are developed in Heckman and Hotz (1989) and Rosenbaum (1987). These methods typically rely on estimating a causal effect that is known to be equal to zero. If the test suggests the causal effect is not zero, the unconfoundedness assumption is viewed as less plausible. The tests can be divided into two broad groups.

The first set of tests focuses on estimating the causal effect of the treatment on a variable that is known not to be affected by the treatment, typically because its value is determined prior to the treatment itself. Such a variable can be a time-invariant covariate, but the most interesting case is where this is a lagged outcome. In this case one takes all the covariates minus the single covariate that is being tested, say the lagged outcome. One estimates the average treatment effect on the lagged outcome. If it is zero, it is more plausible that the unconfoundedness assumption holds. Of course the test is not directly testing the unconfoundedness assumption, and so if the null hypothesis of no effect is not rejected, this does not directly reflect on the hypothesis of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test has arguably more power. For these tests it is clearly helpful to have a number of lagged outcomes.

The second set of tests focuses on estimating the causal effect of a treatment that is known not to have an effect. It relies on the presence of multiple control groups (Rosenbaum, 1987). Suppose one has two potential control groups, for example, as in Heckman, Ichimura and Todd (1997), eligible nonparticipants and ineligibles. One interpretation of the test is that one compares average treatment effects estimated using one control with average treatment effects estimated using the other control group. This can also be interpreted as estimating an average treatment effect using only the two control groups with the treatment indicator being the dummy for being a member of the first control group. In that case the treatment effect is known to be zero, and statistical evidence of a non-zero treatment effect implies that at least one of the control groups is not valid. Again, not rejecting the test does not mean the unconfoundedness assumption is valid, as it could be that both control groups have the same bias, but non-rejection in the case where the two control groups are a priori likely to have different biases makes it more plausible that the unconfoundedness assumption holds. The key for the power of this test is to have control groups that are likely to have different biases if at all. Comparing ineligibles and eligible nonparticipants as in Heckman, Ichimura and Todd (1997) is a particularly attractive comparison. Alternatively one may use different geographic control groups, for example on either side of the treatment group.

5.2 Choosing the Covariates

The discussion so far has focused on the case where the set of covariates is known a priori. In practice there can be two issues with the choice of covariates. First, there may be some variables

that should not be adjusted for. Second, even with variables that should be adjusted for in large samples, expected mean squared error may be reduced by ignoring some of the covariates that have only weak correlation with the treatment indicator and the outcomes. This second issue is essentially a statistical one. Including a covariate in the adjustment procedure, through regression, matching or otherwise, will not lower asymptotic precision of the average treatment effect if the assumptions are correct. In finite samples, however, a covariate that is not or only weakly correlated with outcomes and treatment indicators may reduce precision. There are few procedures currently available for optimally choosing the set of covariates to be included in matching or regression adjustments taking into account such finite sample properties.

The first issue is a substantive one. The unconfoundedness assumption may apply with one set of covariates but it need not apply with an expanded set. A particular concern is the inclusion of covariates that are themselves affected by the treatment such as intermediate outcomes. Suppose, for example, that the primary outcome of interest for the evaluation of a job training program is earnings two years later. In that case employment status prior to the training program is unaffected by the treatment and a valid element of the set of covariates to be used for adjustment. In contrast, employment status one year after the program is an intermediate outcome that should not be controlled for. It could itself be an outcome of interest, and therefore never a covariate in an analysis of the effect of the training program. One guarantee that a covariate is not affected by the treatment is that it was measured before the treatment was applied. In practice, however, often the covariates are recorded at the same time as the outcomes, subsequently to applying the treatment. In that case one has to assess on a case-by-case basis whether a particular covariate should be used in adjusting outcomes or not. See Rosenbaum (1984b) for more discussion on this.

5.3 Assessing the Overlap Assumption

The second of the key assumptions requires that the propensity score is strictly between zero and one. Although in principle this is testable, as it restricts the joint distribution of observables, formal tests are not necessarily the main concern. In practice, this assumption raises a number of issues. The first question is how to detect lack of overlap. A second question is how to deal with it once it is determined that there is indeed a lack of overlap in the covariate distributions. A third issue is how the methods discussed in Section 3 deal with lack of overlap. Ideally such lack of overlap would show up in large standard errors for the average treatment effects.

The first method to detect lack of overlap is to plot distributions of covariates by treatment groups. In the case with one or two covariates one can do this directly. In high dimensional cases this is more difficult. One can inspect pairs of marginal distributions by treatment status, but these are not necessarily informative about lack of overlap. It is possible that for each covariate the distribution for the treatment and control group are identical, even though there are areas where the propensity score is zero or one.

A more useful method is therefore to inspect the distribution of the propensity score in both treatment groups. This directly reveals lack of overlap in the covariate distributions. Its implementation requires nonparametric estimation of the propensity score, however, and misspecification may lead one to failure to detect lack of overlap in the same way that inspecting

various marginal distributions may not be sufficient. In practice one may wish to undersmooth in the estimation of the propensity score, either by choosing the bandwidth smaller than optimal for nonparametric estimation or by including higher order terms in a series expansion.

A third way to detect lack of overlap is to inspect the quality of the worst matches in a matching procedure. Given a set of matches, one can for each component of the vector of covariates inspect $\max_i |x_{i,k} - x_{\ell_1(i),k}|$, the maximum over all observations of the matching discrepancy. If this difference is large relative to the sample standard deviation of the k th component of the covariates, there would be reason for concern. The advantage of this method is that it does not require additional nonparametric estimation.

Once one determines that there is a lack of overlap one can decide that the average treatment effect of interest cannot be estimated with sufficient precision, and/or decide to focus on an average treatment effect that is estimable with greater accuracy. To do the latter it can be useful to discard some of the observations on the basis of their covariates. For example one may decide to discard control (treated) observations with propensity scores below (above) a cutoff level. The desired cutoff level may be sample size dependent. In a very large sample one may not be concerned with a propensity score value of 0.01, whereas in small sample such a value may make it difficult to find reasonable comparisons. To judge such tradeoffs, it is useful to understand the relation between the propensity score and the implicit weight a unit has in the average treatment effect. Using the weighting estimator the average outcome under the treatment is estimated by summing up outcomes for the control units with weight equal to 1 over the propensity score (and 1 over one minus the propensity score for treated units. Hence with N units the weight of unit i is $1/(N \cdot e(X_i))$ if this is a control unit and $1/(N \cdot (1 - e(X_i)))$ if it is a treated unit. One may wish to limit this weight to some fraction, for example, 0.05, so that no unit with have a weight of more than 5% in the average. Under that approach the limit on the propensity score in a sample with 200 units is 0.1, and units with a propensity score less than 0.1 or greater than 0.9 would be discarded. In a sample with 1000 units, only units with a propensity score outside the range $[0.02, 0.98]$ would be discarded.

In matching procedures one can need not rely entirely on the propensity score distribution comparisons. One may wish to discard the observations with an insufficient match quality. Rosenbaum and Rubin (1984) suggest accepting only matches where the difference in propensity score values between matches is below a cutoff point. Alternatively one may also wish to drop matches where other covariates are severely mismatched.

Finally, let us consider the three approaches to inference, regression, matching and propensity score methods, and see how they deal with lack of overlap. Suppose one is interested in estimating the average effect on the treated, and one has a data set with sufficient overlap. Now suppose one adds a few treated or control observations with covariate values rarely seen in the alternative treatment group. Adding the treated observations with outlying values implies one cannot estimate the average treatment effect for the treated very precisely, so with methods appropriately dealing with limited overlap one would see the variance estimates go up. Adding the control observations should not affect results very much since additional control observations with outlying covariate values is irrelevant for the average treatment effect for the treated, and so methods appropriately dealing with limited overlap would show estimates approximately

unchanged in terms of bias and precision.

Consider first the regression approach. Conditional on a particular parametric specification for the regression function adding observations with outlying values of the regressors leads to considerably more precise parameter estimates. Such observations are likely to be influential because of their outlying values. If the added observations are treated observations, the precision of the estimated control regression function at those values will be lower, and so the variance will go up as it should, although there is concern that the estimates may be sensitive to the specification with such observations. Adding control observations will lead to a spurious increase in precision. Regression methods can therefore be misleading in cases with limited overlap.

Next, consider matching. If we focus on estimating the average treatment effect for the treated. Adding control observations with outlying covariate values is unlikely to affect the results very much. Such observations would be unlikely to be used as matches and without that they do not affect the results at all. The results would be more sensitive to adding treated observations with outlying covariate values. These observations would be matched to inappropriate controls, leading to possibly biased estimates. The standard errors would not be affected much.

Finally, consider propensity score estimates. The propensity score estimates would be now include values close to zero and one. The values close to zero for the control observations would not lead to any problems because these observations would get close to zero weight in the estimation. The observations with the propensity score close to one would lead to large weights for some of the control observations, and thus to an increase in the variance of the average treatment effect estimator, correctly inferring that one cannot estimate the average treatment effect very precisely. Blocking on the propensity score would lead to similar conclusions.

Overall, matching (and similarly kernel-based regression methods) and propensity score methods are better designed to cope with limited overlap in the covariate distributions than parametric or semiparametric (series) regression models. In all cases it is useful to inspect histograms of the estimated propensity score in both groups to assess whether limited overlap is an issue.

6 Applications and Simulations

There are many studies using some form of unconfoundedness or selection on observables, ranging from simple least squares analyses to matching on the propensity score (e.g., Ashenfelter and Card, 1978; Lalonde, 1986; Card and Sullivan, 1988; Heckman, Ichimura and Todd, 1997; Angrist, 1998; Dehejia and Wahba, 1999, Lechner, 1998; Friedlander and Robins, 1995, and many others) Here I mainly focus on two sets of analyses that can be helpful in assessing the value of the methods surveyed in this paper. First, studies attempting to assess the plausibility of the assumptions, often using randomized experiments as a yard stick. Second, simulation studies focusing on the performance of the various techniques in settings where the assumptions are known to hold.

6.1 Applications: Randomized Experiments as Checks on Unconfoundedness

The basic idea behind these studies is simple. Given a randomized experiment one can obtain unbiased estimates of the average effect of a program. Then put aside the experimental control group, and attempt to replicate the results using a non-experimental control group. If one is successful, that suggests the assumptions and methods are more plausible than if one is not. Such investigations are of course not conclusive in general, but they are invaluable tools to assess the plausibility of the approach. The first such study, and one that made an enormous impact, was by Lalonde (1986). Fraker and Maynard (1986) conducted a similar investigation, and subsequently many more have followed.

Lalonde (1986) took the National Supported Work program, a fairly small program aimed at people particularly disadvantaged in the labor market, that is individuals with poor labor market histories and skills. He set aside the experimental control group and constructed alternative control groups from the Panel Study of Income Dynamics (PSID) and Current Population Survey (CPS), with various selection criteria depending on prior labor market histories. He then used a number of methods, ranging from a simple difference, least squares adjustment, Heckman selection correction methods to difference-in-differences techniques. His general conclusion was the results were very variable and that no method could consistently replicate the experimental results with the six non-experimental control groups he had constructed. A number of researchers have subsequently used the same data. Heckman and Hotz (1989) focused on testing the various models and argued that the testing procedures they developed would have eliminated many of the particularly inappropriate estimates. Dehejia and Wahba (1999) used some of the semiparametric methods based on the unconfoundedness assumption discussed in the current paper, and found that for the subsample of the Lalonde data they used (with two years of prior earnings), these methods replicated the experimental results more accurately, both overall and within subpopulations. Smith and Todd (2001) investigate the robustness of the Dehejia-Wahba results to the sample selected.

Others have used different experiments to carry out the same or similar analyses, often with different sets of estimators, and with different alternative control groups. Friedlander and Robins (1995) use data from the experiments which were conducted in a number of states and use control groups from other counties in the same state as well as from different states. They focus on least squares adjustment. Hotz, Imbens and Mortimer (2001) use the same data and consider matching methods with various sets of covariates, using single alternative states or multiple alternative states as non-experimental control groups.

Heckman, Ichimura and Todd (1997, 1998ab) study the national (JPTA) program, using data from different locations to investigate the nature of the biases, and the importance of overlap and labor market histories. Their conclusions provide the type of specific guidance that should be the aim of these studies. They give clear and generalizable conditions that make the assumptions, at least according to their study of a large training program, more plausible, such as detailed earnings histories, and control groups that are geographically close to the treatment group, preferably groups of ineligibles, or eligible nonparticipants from the same location. In contrast, control groups from very different locations are found to be poor non-experimental

control groups. Although such conclusions are obviously not generalizable to evaluations other than those of social programs, they are potentially very useful in providing applied evaluators of such programs with concrete guidance as to the applicability of these assumptions.

Dehejia (1997) uses the Greater Avenues to INdependence (GAIN) data, using different counties as well as different offices within the same county as nonexperimental control groups. Hotz, Imbens and Klerman (2001) use the basic GAIN data set supplemented with administrative data with long-term quarterly earnings data both prior and subsequent to the randomization date to investigate the importance of detailed earnings histories. Such detailed histories can also provide more evidence on the plausibility of non-experimental evaluations for long-term outcomes.

Two difficulties make this literature difficult to evaluate. One is the differences in covariates used. It is rare that the variables are measured consistently across different studies. Some studies have yearly earnings data, others quarterly, whereas some only have earnings indicators on a monthly or quarterly basis, making it difficult to consistently investigate, for example, the level of earnings history detail necessary for the unconfoundedness assumption. A second issue is that generally different estimators are used, so that any differences found in results can be attributed to either estimators or assumptions. This is partly the result of few of the estimators being sufficiently standardized that they can be implemented easily by empirical researchers.

All these studies took data from actual randomized experiments. Temporarily putting aside the experimental control group these studies attempt to replicate the experiment results using non-experimental control groups constructed either from experimental control groups from other locations, or from public use surveys such as the PSID and CPS. To some extent such experimental data are not required. The question of interest is whether an alternative control group is an adequate proxy for a randomized control group in a particular setting. First of all note that this question does not require data on the treatment group. Although these questions have typically been studied by comparing experimental results to non-experimental results, all that is relevant for this question is whether the non-experimental control group can predict the average outcomes for the experimental control group. As in the Heckman, Ichimura, Smith and Todd (1998) analysis of the Job Training Partnership Act (JTPA) data one can take two groups, neither subject to the treatment, and ask the question whether, using data on the covariates for the first control group, in combination outcome and covariate information on the second control group, one can predict the average outcome in the first control group. If this question is answered affirmatively, it implies that had there been an experiment on the population from which the first control group is drawn, the second control group would have been an acceptable non-experimental control group. From this perspective one can use data from many different surveys. In particular one can more systematically investigate whether control groups from different counties, states, regions or even different time periods make acceptable non-experimental control groups.

6.2 Simulations

A second question that is often confounded with that of the validity of the assumptions, is that of the relative performance of the various estimators. Suppose one is willing to make the

unconfoundedness and overlap assumptions. Which method is most appropriate in a particular setting? In many of the studies where non-experimental evaluations are compared to experimental ones, researchers use a number of the techniques described here. However, although it is useful to compare these techniques in realistic settings, it is also important to compare them in settings where one is confident that the underlying assumptions are valid.

There are a couple of studies that specifically set out to do so. Frölich (2000) compares a number of matching estimators as well as local linear regression methods, carefully formalizing fully data-driven procedures for the estimators he considers. He considers a large number of data generating processes, based on eight different regression functions (including some highly nonlinear and multimodal ones) and three different density functions for the covariate, and two different sample sizes. One important limitation is that he restricts the investigation to a single covariate. Frölich considers matching estimators with a single match, with replacement. For the local linear regression estimators he uses data-driven optimal bandwidth choices based on minimizing mean-squared-error of the average treatment effect. The first local linear estimator Frölich considers is the standard one. At x the regression function $\mu(x)$ is estimated as β_0 in the minimization problem

$$\min_{\beta_0, \beta_1} \sum_{i=1}^N (Y_i - \beta_0 - \beta_1 \cdot (X_i - x))^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

with an Epanechnikov kernel. He finds that this has computational problems, as well as poor small sample properties and so he also considers a modification suggested by Seifert and Gasser (1996, 2000). For given x , define $\bar{x} = \sum X_i K((X_i - x)/h) / \sum K((X_i - x)/h)$, so that one can write the standard local linear estimator as

$$\hat{\mu}(x) = \frac{T_0}{S_0} + \frac{T_1}{S_2} \cdot (x - \bar{x}),$$

where, for $r = 0, 1, 2$, $S_r = \sum K((X_i - x)/h)(X_i - x)^r$ and $T_r = \sum K((X_i - x)/h)(X_i - x)^r Y_i$. The Seifert-Gasser modification is to use instead

$$\hat{\mu}(x) = \frac{T_0}{S_0} + \frac{T_1}{S_2 + R} \cdot (x - \bar{x}),$$

where the choice for the ridge parameter R recommended by Seifert and Gasser is $R = |x - \bar{x}|(5/(16h))$, given the Epanechnikov kernel $k(u) = (3/4)(1 - u^2)1\{|u| < 1\}$. Note that with high-dimensional covariates such a nonnegative kernel would lead to biases that do not vanish fast enough to be dominated by the variance, as pointed out in Heckman, Ichimura and Todd (1998). This is not a problem in Frölich's simulations as he considers only cases with a single covariate. Frölich finds that the local linear estimator, with the modification by Seifert and Gasser performs better than the matching estimator or the standard local linear estimator.

Zhao (2002) compares parametric regression estimators with matching estimators. He uses metrics based on the propensity score, the covariates and estimated regression functions. Using designs with two covariates and linear regression functions Zhao finds there is no clear winner among the different estimators, although he notes that using the outcome data in choosing the metric appears a promising strategy.

Abadie and Imbens (2002) study their matching estimator by simulation methods using a data generating process inspired by the Lalonde study. The data are generated to allow for substantial nonlinearity by fitting a separate binary response model to the zeros in the earnings outcome and a log linear model for the positive part. The regression estimators include linear and quadratic models (the latter with a full set of interactions), with seven covariates. They find that the matching estimators, and in particular the bias-adjusted matching estimator, outperform the linear (with 7 covariates) and quadratic regression (with 35 covariates after dropping squares and interactions that lead to perfect collinearity) estimators. Their simulations also suggest that with relatively few matches (between one and four) matching estimators are not sensitive to the number of matches, and that their confidence intervals have close to nominal coverage.

The results from these simulation studies are overall somewhat inconclusive. It is clear that more work is required here. Future simulations may usefully focus on some of the following issues. First of all it is important obviously to closely model the data generating process on actual data sets to ensure that the results have some relevance for practice. Ideally one would build the simulations around a number of specific data sets to get a range of data generating processes that capture settings that are relevant in practice.

Second, it is important to understand which features of the data generating process are important for the properties of the various estimators. For example, do some estimators deteriorate more rapidly than others when there are many covariates and few observations? Are some estimators more robust against high correlations between covariates and outcomes or high correlations between covariates and treatment indicators? Which estimators are more likely to give conservative answers in terms of precision? What is important here is to isolate salient features of the data generating processes. It is clear that no estimator is always going to dominate all others. If the regression functions are linear with few covariates, it is clear that regression estimators will perform well. However, when the regression functions are highly nonlinear, and there is little overlap, such estimators are less likely to do well. It would be helpful to have descriptive statistics that summarize the features of the data that provide guidance to choosing estimators that are likely to perform well in a given situation.

Finally, it is important to have fully data-driven procedures that define an estimator as a function of $(Y_i, W_i, X_i)_{i=1}^N$, as for example in Frölich (2000). For the matching estimators this is relatively straightforward, but for some of the other estimators this requires more care. This allows other researchers to do meaningful comparisons of the various estimators.

7 Conclusion

In this paper I have attempted to give a review of the current state of the literature on inference for average treatment effects under the assumption of unconfoundedness. This has recently been a very active area of research where many new semi- and non-parametric econometric methods are being applied and developed. The research has moved a long way from relying on simple least squares methods for estimating average treatment effects.

Efficiency bounds have been established for a number of average treatment effects of interest,

and a variety of estimators relying on the weakest assumptions that allow point identification. These include propensity score methods and pairwise matching, as well as nonparametric regression methods. There is as of yet no consensus what are the best methods to apply in practice. Nevertheless, the applied researcher has now a large number of new estimators at her disposal.

For estimating variances a number of methods are available. First, one can estimate the asymptotic variance for the population average treatment effect and its counterpart for the subpopulation of treated units either through estimating the components of this variance nonparametrically. This is fairly cumbersome. In practice the second method of estimating the variance through bootstrapping methods is more common. Alternatively if one focuses on the average treatment effect for the sample (or conditional on the covariates), one can use the variance estimator developed by Abadie and Imbens (2002) which does not require nonparametric estimation.

Challenges remain in making the new tools more easily applicable. Although software has been made available to implement some of the estimators (see Becker and Ichino, 2002; Abadie, Drukker, Herr and Imbens, 2003), many of the estimators remain difficult to apply. A particularly urgent task is therefore to provide fully implementable versions of the various estimators that do not require the applied researcher to choose bandwidths or other smoothing parameters. This is less of a concern for matching methods and probably explains a large part of their popularity. Another outstanding question is the relative performance of these methods in realistic settings with large numbers of covariates and varying degrees of smoothness in the conditional means of the potential outcomes and the propensity score.

REFERENCES

- ABADIE, ALBERTO, (2001): "Semiparametric Difference-in-Differences Estimators," forthcoming, *Review of Economic Studies*.
- ABADIE, A. (2002), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics* (forthcoming).
- ABADIE, A., AND G. IMBENS, (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," unpublished manuscript, Kennedy School of Government, Harvard University.
- ABADIE, A., D. DRUKKER, H. HERR, AND G. IMBENS, (2003), "Implementing Matching Estimators for Average Treatment Effects in STATA," unpublished manuscript.
- ABADIE, A., J. ANGRIST, AND G. IMBENS, (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica*.
- ANGRIST, J., (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica*, Vol. 66, No. 2, 249–288.
- ANGRIST, J.D., G.W. IMBENS AND D.B. RUBIN (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.
- ANGRIST, J. D. AND A. B. KRUEGER (2000), "Empirical Strategies in Labor Economics," in A. Ashenfelter and D. Card eds. *Handbook of Labor Economics*, vol. 3. New York: Elsevier Science.
- ANGRIST, J. D., AND J. HAHN, (1999) "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER Technical Working Paper 241.
- ANGRIST, J., AND V. LAVY (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, Vol. CXIV, 1243.
- ASHENFELTER, O., AND D. CARD, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648-660.
- ATHEY, S., AND G. IMBENS (2002), "Identification and Inference in Nonlinear Difference-In-Differences Models," unpublished manuscript, Department of Economics, Stanford University.
- ATHEY, S., AND S. STERN, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", NBER working paper 6600.
- BARNOW, B.S., G.G. CAIN AND A.S. GOLDBERGER (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies*, vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.
- BECKER, S., AND A. ICHINO, (2002), "Estimation of Average Treatment Effects Based on Propensity Scores," forthcoming, *The Stata Journal*
- BITLER, M., J. GELBACH, AND H. HOYNES (2002) "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," unpublished paper, Department of Economics, University of Maryland.
- BJÖRKLUND, AND R. MOFFIT, (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models", *Review of Economics and Statistics*, Vol. LXIX, 42–49.
- BLACK, S., (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, Vol. CXIV, 577.

- BLUNDELL, R. AND COSTA-DIAS (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02.
- BLUNDELL, R., A. GOSLING, H. ICHIMURA, AND C. MEGHIR, (2002) "Changes in the Distribution of Male and Female Wages Accounting for the Employment Composition," unpublished paper, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, United Kingdom.
- CARD, D., AND SULLIVAN, (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica*, vol. 56, no. 3 497-530.
- CHERNOZHUKOV, V., AND C. HANSEN, (2001), "An IV Model of Quantile Treatment Effects," unpublished working paper, Department of Economics, MIT.
- COCHRAN, W., (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies", *Biometrics* 24, 295-314.
- COCHRAN, W., AND D. RUBIN (1973) "Controlling Bias in Observational Studies: A Review" *Sankhya*, 35, 417-46.
- DEHEJIA, R., (2002) "A Decision-theoretic Approach to Program Evaluation", *Journal of Business and Economic Statistics*.
- DEHEJIA, R., AND S. WAHBA, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94, 1053-1062.
- ENGLE, R., D. HENDRY, AND J.-F. RICHARD, (1974) "Exogeneity," *Econometrica*.
- FIRPO, S. (2002), "Efficient Semiparametric Estimation of Quantile Treatment Effects," Unpublished paper, Department of Economics, University of California, Berkeley.
- FISHER, R. A., (1935), *The Design of Experiments*, Boyd, London.
- FITZGERALD, J., P. GOTTSCHALK, AND R. MOFFITT, (1998), "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics", *Journal of Human Resources* 33, 251-299.
- FRAKER, T., AND R. MAYNARD, (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *Journal of Human Resources*, Vol. 22, No. 2, p 194-227.
- FRIEDLANDER, D., AND P. ROBINS, (1995), "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods", *American Economic Review*, Vol. 85, p 923-937.
- FRÖLICH, M. (2000), "Treatment Evaluation: Matching versus Local Polynomial Regression," Discussion paper 2000-17, Department of Economics, University of St. Gallen.
- FRÖLICH, M. (2002), "What is the Value of knowing the propensity score for estimating average treatment effects", Department of Economics, University of St. Gallen.
- GILL, R., AND J. ROBINS, J., "Causal Inference for Complex Longitudinal Data: The Continuous Case," *Annals of Statistics*.
- GU, X., AND P. ROSENBAUM, (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms", *Journal of Computational and Graphical Statistics*, 2, 405-20.
- HAHN, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.
- HAHN, J., P. TODD, AND W. VANDERKLAUW, (2000), "Regression discontinuity," *Econometrica*.

- HAM, J., AND R. LALONDE, (1996) "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training, *Econometrica*, 64: 1.
- HECKMAN, J., AND J. HOTZ, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association*.
- HECKMAN, J., AND R. ROBB, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.
- HECKMAN, J., J. SMITH, AND N. CLEMENTS, (1997), "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts", *Review of Economic Studies*, Vol 64, 487-535.
- HECKMAN, J., H. ICHIMURA, AND P. TODD, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies* 64, 605-654.
- HECKMAN, J., H. ICHIMURA, AND P. TODD, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261-294.
- HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.
- HECKMAN, J.J., R.J. LALONDE, AND J.A. SMITH (2000), "The Economics and Econometrics of Active Labor Markets Programs," in A. Ashenfelter and D. Card eds. *Handbook of Labor Economics*, vol. 3. New York: Elsevier Science.
- HIRANO, K., AND G. IMBENS (2002), "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Hear Catherization," *Health Services anf Outcomes Research Methodology*, forthcoming.
- HIRANO, K., G. IMBENS, AND G. RIDDER, (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," forthcoming *Econometrica*.
- HOTZ J., G. IMBENS, AND J. MORTIMER (1999), "Predicting the Efficacy of Future Training Programs Using Past Experiences" NBER Working Paper.
- ICHIMURA, H., AND O. LINTON, (2001), "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." Institute for Fiscal Studies, cemmap working paper cwp04/01.
- ICHIMURA, H., AND C. TABER, (2000), "Direct Estimation of Policy Effects", unpublished manuscript, Department of Economics, Northwestern University.
- IMBENS, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*.
- IMBENS, G. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, Papers and Proceedings.
- IMBENS, G., AND J. ANGRIST (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*.
- LALONDE, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604-620.
- LECHNER, M, (1998), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics*.

- LECHNER, M., (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in Lechner and Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg, Physica.
- LEE, D, (2001), "The Electoral Advantage of Incumbency and the Voter's Valuation of Political Experience: A Regression Discontinuity Analysis of Close Elections," unpublished paper, Department of Economics, University of California.
- MANSKI, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.
- MANSKI, C., G. SANDEFUR, S. MCLANAHAN, AND D. POWERS (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.
- NEWBY, W.K., (1995) "Convergence Rates for Series Estimators," in G.S. Maddalla, P.C.B. Phillips and T.N. Srinivasan eds. *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao*. Cambridge: Blackwell.
- NEYMAN, J., (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science*, (with discussion), Vol 5, No 4, 465-480, 1990.
- QUADE, D., (1982), "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38, 597-611.
- ROBINS, J., AND R. GILL, *Annals of Statistics*.
- ROBINS, J., AND Y. RITOV, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16, 285-319.
- ROBINS, J.M., AND A. ROTNITZKY, (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122-129.
- ROBINS, J.M., ROTNITZKY, A., ZHAO, L-P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106-121.
- ROSENBAUM, P., (1984a), "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 79, 565-574.
- ROSENBAUM, P., (1984b), "The Consequences of Adjustment for a Concomitant Variable that has been Affected by the Treatment," *Journal of the Royal Statistical Society, Series A*, 147, 656-666.
- ROSENBAUM, P., (1989), "Optimal Matching in Observational Studies", *Journal of the American Statistical Association*, 84, 1024-1032.
- ROSENBAUM, P., (1987), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292-316.
- ROSENBAUM, P., (1995), *Observational Studies*, Springer Verlag, New York.
- ROSENBAUM, P., (2000), "Covariance Adjustment in Randomized Experiments and Observational Studies," forthcoming, *Statistical Science*.
- ROSENBAUM, P., AND D. RUBIN, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

- ROSENBAUM, P., AND D. RUBIN, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Ser. B*, 45, 212-218.
- ROSENBAUM, P., AND D. RUBIN, (1984), "Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of the American Statistical Association*, 79, 516-524.
- ROSENBAUM, P., AND D. RUBIN, (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", *American Statistician*, 39, 33-38.
- RUBIN, D., (1973a), "Matching to Remove Bias in Observational Studies", *Biometrics*, 29, 159-183.
- RUBIN, D., (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.
- RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.
- RUBIN, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1-26.
- RUBIN, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *Annals of Statistics*, 6:34-58.
- RUBIN, D., (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.
- RUBIN, D., AND N. THOMAS, (1992), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20 (2) 1079-1093.
- SEIFERT, B., AND T. GASSER (1996), "Finite-sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association*, 91, 267-275.
- SEIFERT, B., AND T. GASSER (2001), "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*.
- SMITH, J. A. AND P. E. TODD, (2001a), "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review, Papers and Proceedings*, 91:112-118.
- VAN DER KLAUW, W., (2002), "A Regression-discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment", *International Economic Review*, forthcoming
- ZHAO, Z., (2002), "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics and an Application", unpublished manuscript.