IRP Lectures **Madison, WI, August 2008**

Lecture 1, Monday, Aug 4th, 8.30-9.45am

Estimation of Average Treatment Effects Under Unconfoundedness, Part I

1. Introduction

In this lecture we look at several methods for estimating average effects of a program, treatment, or regime, under unconfoundedness. The setting is one with a binary program. The traditional example in economics is that of a labor market program where some individuals receive training and others do not, and interest is in some measure of the effectiveness of the training. Unconfoundedness, a term coined by Rubin (1990), refers to the case where (non-parametrically) adjusting for differences in a fixed set of covariates removes biases in comparisons between treated and control units, thus allowing for a causal interpretation of those adjusted differences. This is perhaps the most important special case for estimating average treatment effects in practice. Alternatives typically involves strong assumptions linking unobservables to observables in specific ways in order to allow adjusting for the relevant differences in unobserved variables. An example of such a strategy is instrumental variables, which will be discussed in Lecture 3. A second example that does not involve additional assumptions is the bounds approach developed by Manski (1990, 2003).

Under the specific assumptions we make in this setting, the population average treatment effect can be estimated at the standard parametric $\sqrt{N}$ rate without functional form assumptions. A variety of estimators, at first sight quite different, have been proposed for implementing this. The estimators include regression estimators, propensity score based estimators and matching estimators. Many of these are used in practice, although rarely is this choice motivated by principled arguments. In practice the differences between the estimators are relatively minor when applied appropriately, although matching in combination with regression is generally more robust and is probably the recommended choice. More important than the choice of estimator are two other issues. Both involve analyses of the data without the outcome variable. First, one should carefully check the extent of the overlap

in covariate distributions between the treatment and control groups. Often there is a need for some trimming based on the covariate values if the original sample is not well balanced. Without this, estimates of average treatment effects can be very sensitive to the choice of, and small changes in the implementation of, the estimators. In this part of the analysis the propensity score plays an important role. Second, it is useful to do some assessment of the appropriateness of the unconfoundedness assumption. Although this assumption is not directly testable, its plausibility can often be assessed using lagged values of the outcome as pseudo outcomes. Another issue is variance estimation. For matching estimators bootstrapping, although widely used, has been shown to be invalid. We discuss general methods for estimating the conditional variance that do not involve resampling.

In these notes we first set up the basic framework and state the critical assumptions in Section 2. In Section 3 we describe the leading estimators. In Section 4 we discuss variance estimation. In Section 5 we discuss assessing one of the critical assumptions, unconfoundedness. In Section 6 we discuss dealing with a major problem in practice, lack of overlap in the covariate distributions among treated and controls. In Section 7 we illustrate some of the methods using a well known data set in this literature, originally put together by Lalonde (1986).

In these notes we focus on estimation and inference for treatment effects. We do not discuss here a recent literature that has taken the next logical step in the evaluation literature, namely the optimal assignment of individuals to treatments based on limited (sample) information regarding the efficacy of the treatments. See Manski (2004, 2005, Dehejia (2004), Hirano and Porter (2005).

2. Framework

The modern set up in this literature is based on the potential outcome approach developed by Rubin (1974, 1977, 1978), which view causal effects as comparisons of potential outcomes defined on the same unit. In this section we lay out the basic framework.

2.1 Definitions

We observe N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population. We postulate the existence for each unit of 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, referred to as covariates, pretreatment variables or exogenous variables, and denoted by X_i.[1] It is important that these variables are not affected by the treatment. Often they take their values prior to the unit being exposed to the treatment, although this is not sufficient for the conditions they need to satisfy. Importantly, this vector of covariates can include lagged outcomes. Finally, each unit is exposed to a single treatment; $W_i = 0$ if unit i receives the control treatment and $W_i = 1$ if unit i receives the active treatment. We therefore observe for each unit 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_i, Y_i, X_i) refer to the distribution induced by the random sampling from the population.

Several additional pieces of notation will be useful in the remainder of these notes. First, the propensity score (Rosenbaum and Rubin, 1983) is defined as the conditional probability of receiving the treatment,

$$e(x) = \Pr(W_i = 1 | X_i = x) = \mathbb{E}[W_i | X_i = x].$$

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

$$\mu_w(x) = \mathbb{E}[Y_i(w) | X_i = x], \qquad \sigma_w^2(x) = \mathbb{V}(Y_i(w) | X_i = x).$$

2.2 Estimands: Average Treatment Effects

[1] Calling such variables exogenous is somewhat at odds with several formal definitions of exogeneity (e.g., Engle, Hendry and Richard, 1974), as knowledge of their distribution can be informative about the average treatment effects. It does, however, agree with common usage. See for example, Manski, Sandefur, McLanahan, and Powers (1992, p. 28).

In this discussion we will primarily focus on a number of average treatment effects (ATEs). For a discussion of testing for the presence of any treatment effects under unconfoundedness see Crump, Hotz, Imbens and Mitnik (2007). Focusing on average effects is less limiting than it may seem, however, as this includes averages of arbitrary transformations of the original outcomes.[2] The first estimand, and the most commonly studied in the econometric literature, is the population average treatment effect (PATE):

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

Alternatively we may be interested in the population average treatment effect for the treated (PATT, e.g., Rubin, 1977; Heckman and Robb, 1984):

$$\tau_{P,T} = \mathbb{E}[Y_i(1) - Y_i(0)|W = 1].$$

Most of the discussion in these notes will focus on τ_P, with extensions to $\tau_{P,T}$ available in the references.

We will also look at sample average versions of these two population measures. These estimands focus on the average of the treatment effect in the specific sample, rather than in the population at large. These include, the sample average treatment effect (SATE) and the sample average treatment effect for the treated (SATT):

$$\tau_S = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big), \qquad \text{and} \quad \tau_{S,T} = \frac{1}{N_T}\sum_{i:W_i=1}\Big(Y_i(1) - Y_i(0)\Big),$$

where $N_T = \sum_{i=1}^{N} W_i$ is the number of treated units. The sample average treatment effects have received little attention in the recent econometric literature, although it 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 ATE beyond the

[2]Lehman (1974) and Doksum (1974) introduce quantile treatment effects as the difference in quantiles between the two marginal treated and control outcome distributions. Bitler, Gelbach and Hoynes (2002) estimate these in a randomized evaluation of a social program. Firpo (2003) develops an estimator for such quantiles under unconfoundedness.

sample ATE. To see this, consider the case where we observe the sample $(Y_i(0), Y_i(1), W_i, X_i)$, $i = 1, \ldots, N$; that is, we observe for each unit both potential outcomes. In that case the sample average treatment effect, $\tau_S = \sum_i (Y_i(1) - Y_i(0))/N$, can be estimated without error. Obviously the best estimator for the population average effect, τ_P, is τ_S. However, we cannot estimate τ_P without error even with a sample where all potential outcomes are observed, because we lack the potential outcomes for those population members not included in the sample. This simple argument has two implications. First, one can estimate the sample ATE at least as accurately as the population ATE, and typically more so. In fact, the difference between the two variances is the variance of the treatment effect, which is zero only when the treatment effect is constant. Second, a good estimator for one average treatment effect 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.

The difference in asymptotic variances forces the researcher 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 PATE is different from zero, whereas there may be compelling evidence that the SATE is positive. Typically researchers in econometrics have focused on the PATE, but one can argue that it is of interest, when one cannot ascertain the sign of the population-level effect, to know whether one can determine 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, results for the sample at hand may be of considerable interest.

2.2 Identification

We make the following key assumption about the treatment assignment:

Assumption 1 (Unconfoundedness)

$$\Big(Y_i(0), Y_i(1)\Big) \perp\!\!\!\perp W_i \,\Big|\, X_i.$$

This assumption was first articulated in this form in Rosenbaum and Rubin (1983a). Lechner (1999, 2002) refers to this as the "conditional independence assumption," Following a parametric version of this in Heckman and Robb (1984) it is also referred to as "selection on observables." In the missing data literature the equivalent assumption is referred to as "missing at random."

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. Suppose also that the control outcome is linear in X_i:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i,$$

with $\varepsilon_i \perp\!\!\!\perp X_i$. Then we can write

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

Given the constant treatment effect assumption, unconfoundedness is equivalent to independence of W_i and ε_i conditional on X_i, which would also capture the idea that W_i is exogenous. Without this 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 (Overlap)

$$0 < \Pr(W_i = 1|X_i) < 1.$$

Rosenbaum and Rubin (1983a) refer to the combination of the two assumptions as "stongly ignorable treatment assignment." For many of the formal results one will also need smoothness assumptions on the conditional regression functions and the propensity score ($\mu_w(x)$ and $e(x)$), and moment conditions on $Y_i(w)$. 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 1 and 2 in economic settings and thus the relevance of the econometric literature that focuses on estimation and inference under these conditions for empirical work. In this debate it has been argued that agents' optimizing behavior precludes their choices being independent of the potential outcomes, whether or not conditional on covariates. This seems an unduly narrow view. In response I will offer three arguments for considering these assumptions. The first is a statistical, data descriptive motivation. A natural starting point in the evaluation of any program is 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 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. The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated. Economic theory can help in classifying variables into those that need to be adjusted for versus those that do not, 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 variables 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 settings where some of these covariates are not observed will require strong assumptions to allow for identification. Such assumptions include instrumental variables settings where some covariates are assumed to be independent of the potential outcomes. Absent those assumptions, typically 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 if the difference in their choices is driven by differencese in unobserved characteristics that are themselves unrelated to the outcomes of interest. The plausability of this will depend critically on the exact nature of the optimization process 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 average effect of a binary input (e.g., a new technology) on a firm's output. Assume production is a stochastic function of this input because other inputs (e.g., weather) are not under the firm's control, or $Y_i = g(W, \varepsilon_i)$. Suppose that profits are output minus costs, $\pi_i(w) = g(w, \varepsilon_i) - c_i \cdot w$, and also that a firm chooses a production level to maximize expected profits, equal to output minus costs:

$$W_i = \arg\max_w \mathbb{E}[\pi_i(w)|c_i] = \arg\max_w \mathbb{E}[g(w, \varepsilon_i) - c_i \cdot w|c_i],$$

implying

$$W_i = 1\{\mathbb{E}[g(1, \varepsilon_i) - g(0, \varepsilon_i) \geq c_i|c_i]\} = h(c_i).$$

If unobserved marginal costs c_i differ between firms, and these marginal costs are independent of the errors ε_i in the firms' forecast of production given inputs, then unconfoundedness will hold as

$$(g(0, \varepsilon_i), g(1, \varepsilon_i)) \perp\!\!\!\perp c_i.$$

Note that under the same assumptions one cannot necessarily identify the effect of the input on profits since $(\pi_i(0), \pi_i(1))$ are not independent of c_i. See for a related discussion, in the context of instrumental variables, Athey and Stern (1998). Heckman, Lalonde and Smith (2000) discuss alternative models that justify unconfoundedness. In these models 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 strongly ignorable treatment assignment one can identify the population average treatment effect. The key insight is that given unconfoundedness, the following equalities holds:

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x] = \mathbb{E}[Y_i(w)|W_i = w, X_i = x] = \mathbb{E}[Y_i|W_i = w, X_i = x],$$

and $\mu_w(x)$ is identified. Thus one can estimate the average treatment effect τ by first estimating the average treatment effect for a subpopulation with covariates $X = x$:

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

To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y_i|X_i = x, W_i = w]$ for all values of w and x in the support of these variables. This is where the second assumption enters. If the overlap assumption is violated at $X = x$, it would be infeasible to estimate both $\mathbb{E}[Y_i|X_i = x, W_i = 1]$ and $\mathbb{E}[Y_i|X_i = x, W_i = 0]$ because at those values of x 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 in fact sufficient to assume that

$$\mathbb{E}[Y_i(w)|W_i, X_i] = \mathbb{E}[Y_i(w)|X_i],$$

for $w = 0, 1$. Although this assumption is unquestionably weaker, in practice it is rare that a convincing case is made for the weaker assumption without the case being equally strong

for the stronger Assumption 1. The reason is that the weaker assumption is intrinsically tied to functional form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (e.g., logarithms) without the strong assumption.

One can weaken the unconfoundedness assumption in a different direction if one is only interested in the average effect for the treated (e.g., Heckman, Ichimura and Todd, 1997). In that case one need only assume $Y_i(0) \perp\!\!\!\perp W_i \,\Big|\, X_i$. and the weaker overlap assumption $\Pr(W_i = 1|X_i) < 1$. These two assumptions are sufficient for identification of PATT 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 need not condition simultaneously on all covariates. The following result shows that all biases due to observable covariates can be removed by conditioning solely on the propensity score:

Result 1 *Suppose that Assumption 1 holds. Then:*

$$\Big(Y_i(0), Y_i(1)\Big) \perp\!\!\!\perp W_i \,\Big|\, e(X_i).$$

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

$$\begin{aligned}
\Pr(W_i = 1|Y_i(0), Y_i(1), e(X_i)) &= \mathbb{E}[W_i = 1|Y_i(0), Y_i(1), e(X_i)] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|Y_i(0), Y_i(1), e(X), X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|Y_i(0), Y_i(1), X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] = \mathbb{E}\left[e(X_i)|Y_i(0), Y_i(1), e(X_i)\right] = e(X_i),
\end{aligned}$$

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

$$\Pr(W_i = 1|e(X_i)) = \mathbb{E}[W_i = 1|e(X_i)] = \mathbb{E}\left[\mathbb{E}[W_i = 1|X_i]\Big|e(X_i)\right] = \mathbb{E}\left[e(X_i)|e(X_i)\right] = e(X_i).$$

□

Extensions of this result to the multivalued treatment case are given in Imbens (2000) and Lechner (2001).

To provide intuition for 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:

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

The bias of omitting X_i from the regression on the coefficient on W_i is equal to $\beta_2'\delta$, where δ is the vector of coefficients on W_i in regressions of the elements of X_i on W_i. By conditioning on the propensity score we remove the correlation between X_i and W_i because $X_i \perp\!\!\!\perp W_i | e(X_i)$. Hence omitting X_i no longer leads to any bias (although it may still lead to some efficiency loss).

2.4 Efficiency Bounds and Asymptotic Variances for Population Average Treatment Effects

Next we review some results on the effiency bound for estimators of the average treatment effects τ_P. This requires strong ignorability and some smoothness assumptions on the conditional expectations of potential outcomes and the treatment indicator (for details, see Hahn, 1998). Formally, Hahn (1998) shows 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 can show that

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

Knowing the propensity score does not affect this efficiency bound.

Hahn also shows that asymptotically linear estimators exist that achieve the efficiency bound, 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:

$$\psi(y, w, x, \tau_P) = \left(\frac{wy}{e(x)} - \frac{(1-w)y}{1-e(x)}\right) - \tau_P - \left(\frac{\mu_1(x)}{e(x)} + \frac{\mu_0(x)}{1-e(x)}\right) \cdot (w - e(x)). \tag{2}$$

3. Estimating Average Treatment Effects

Here we discuss the leading estimators for average treatment effects under unconfoundedness. What is remarkable about this literature is the wide range of ostensibly quite different estimators, many of which are regularly used in empirical work. We first briefly describe a number of the estimators, and then discuss their relative merits.

3.1 Regression

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

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} \Big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\Big). \tag{3}$$

In most implementations the average of the predicted treated outcome for the treated is equal to the average observed outcome for the treated (so that $\sum_i W_i \cdot \hat{\mu}_1(X_i) = \sum_i W_i \cdot Y_i$), and similarly for the controls, implying that $\hat{\tau}_{\text{reg}}$ can also be written as

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} W_i \cdot \Big(Y_i - \hat{\mu}_0(X_i)\Big) + (1 - W_i) \cdot \Big(\hat{\mu}_1(X_i) - Y_i\Big).$$

Early estimators for $\mu_w(x)$ included parametric regression functions, for example linear regression (e.g., Rubin, 1977). Such parametric alternatives 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. In this case one can estimate τ simply by least squares estimation using the regression function

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

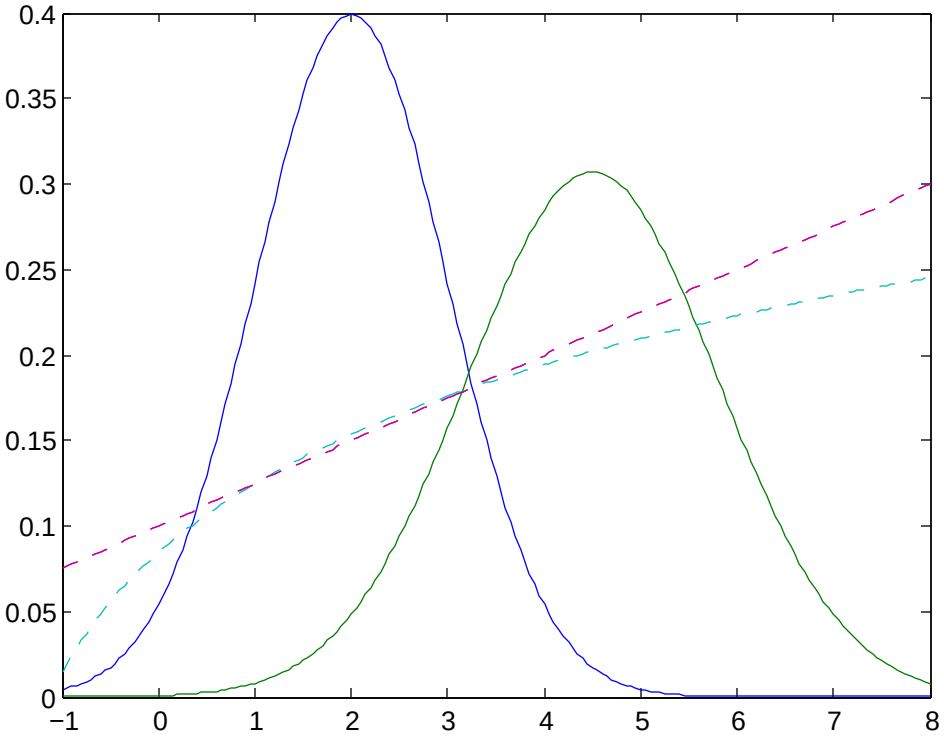
More generally, one can specify separate regression functions for the two regimes, $\mu_w(x) = \beta_w' x$. In that case one estimate the two regression functions separately on the two subsamples and then substitute the predicted values in (3).

These simple regression estimators can be sensitive to differences in the covariate distributions for treated and control units. The reason is that in that case the regression estimators rely heavily on extrapolation. To see this, note that the regression function for the controls, $\mu_0(x)$ is used to predict missing outcomes for the treated. Hence on average one wishes to use predict the control outcome at $\overline{X}_T = \sum_i W_i \cdot X_i / N_T$, the average covariate value for the treated. With a linear regression function, the average prediction can be written as $\bar{Y}_C + \hat{\beta}'(\overline{X}_T - \overline{X}_C)$. If $\overline{X}_T$ and the average covariate value for the controls, $\overline{X}_C$ are close, the precise specification of the regression function will not matter much for the average prediction. However, with the two averages very different, the prediction based on a linear regression function can be sensitive to changes in the specification.

More recently, nonparametric estimators have been proposed. Imbens, Newey and Ridder (2005) and Chen, Hong, and Tarozzi (2005) propose estimating $\mu_w(x)$ through series or sieve methods. A simple version of that with a scalar X would specify the regression function as

$$\mu_w(x) = \sum_{l=0}^{L_N} \beta_{w,l} \cdot x^k,$$

with L_N, the number of terms in the polynomial expansion, an increasing function of the sample size. They show that this estimator for τ_P achieves the semiparametric efficiency


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−1
0
1
2
3
4
5
6
7
8

bounds. Heckman, Ichimura and Todd (1997, 1998), and Heckman, Ichimura, Smith and Todd (1998) consider kernel methods for estimating $\mu_w(x)$, in particular focusing on local linear approaches. Given a kernel $K(\cdot)$, and a bandwidth h_N let

$$\left(\hat{\alpha}_{w,x}, \hat{\beta}_{w,x}\right) = \arg\min_{\alpha_{w,x},\beta_{w,x}} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h_N}\right) \cdot \left(Y_i - \alpha_{w,x} - \beta_{w,x} \cdot X_i\right)^2,$$

leading to the estimator

$$\hat{\mu}_w(x) = \hat{\alpha}_{w,x}.$$

3.2 Matching

Regression estimators impute the missing potential outcomes using the estimated regression function. Thus, if $W_i = 1$, $Y_i(1)$ is observed and $Y_i(0)$ is missing and imputed with a consistent estimator $\hat{\mu}_0(X_i)$ for the conditional expectation. Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of nearest neighbours of the opposite treatment group. 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. In fact, matching can be inrrepreted as a limiting version of the standard kernel estimator where the bandwidth goes to zero. This minimizes the bias among nonnegative kernels, but potentially increases the variance relative to kernel estimators. A formal difference with kernel estimators 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 earlier relied on the 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. In contrast, for the regression estimators discussed above, the researcher must choose smoothing parameters that are more difficult to interpret; either the number of terms in a series or the bandwidth in kernel regression.

Within the class of matching estimators, using only a single match leads to the most credible inference with the least bias, at most sacrificing some precision. This can make the matching estimator easier to use than those 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, 2002; Rubin, 1973b, 1979; Heckman, Ichimura and Todd, 1998; Dehejia and Wahba, 1999; Abadie and Imbens, 2002, AI). 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, and (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 PATT 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 potentially important.

Here we focus on matching estimators for PATE and SATE. In order to estimate these targes we need to match both treated and controls, and allow for matching with replacement. Formally, given a sample, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $\ell_m(i)$ be the index l that satisfies $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1\Big\{\|X_j - X_i\| \leq \|X_l - X_i\|\Big\} = m,$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the m-th closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$.

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: $\mathcal{J}_M(i) = \{\ell_1(i), \ldots, \ell_M(i)\}$. Define the imputed potential outcomes as:

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

The simple matching estimator is then

$$\hat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{Y}_i(1) - \hat{Y}_i(0)\Big). \tag{4}$$

AI 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 the large sample variance if K is at least three.

Let us make clear three caveats to the AI result. First, it is only the continuous covariates that should be counted in K. With discrete covariates the matching will be exact in large samples, therefore 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 appealling 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 $N_1/N_0^{4/K} \to 0$, 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 either the regression function is close to linear, or the density of the covariates close to constant, the resulting bias may be fairly limited. To remove the bias, AI suggest combining the matching process with a regression adjustment.

Another point made by AI is that matching estimators are generally not efficient. Even in the case where the bias is of low enough order to be 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, as done implicitly in kernel estimators. In practice the efficiency loss is limited though, with the gain of going from two matches to a large number of matches bounded as a fraction of the standard error by 0.16 (see AI).

In the above discussion the distance metric in choosing the optimal matches was the standard Euclidan metric $d_E(x,z) = (x-z)'(x-z)$. All of the distance metrics used in practice standardize the covariates in some manner. The most popular metrics are the Mahalanobis metric, where

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

where Σ is covariance matrix of the covairates, and the diagonal version of that

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

Note that depending on the correlation structure, using the Mahalanobis metric can lead to situations where a unit with $X_i = (5,5)$ is a closer match for a unith with $X_i = (0,0)$ than a unit with $X_i = (1,4)$, despite being further away in terms of each covariate separately.

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 in terms of the propensity score (and indirectly also in terms of the covariates) to create balance between treated and control units in the weighted sample. Hirano, Imbens and Ridder (2003) 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 or match on the propensity score.

If the researcher knows the propensity score all three of these methods are likely to be effective in eliminating bias. Even if the resulting estimator is not fully efficient, one can easily modify it by using a parametric estimate of the propensity score to capture most of the efficiency loss. Furthermore, since these estimators do not rely on high-dimensional nonparametric regression, this suggests that their finite sample properties would be attractive.

In practice the propensity score is rarely known, and in that case the advantages of the estimators discussed below are less clear. Although they avoid the high-dimensional nonparametric estimation of the two conditional expectations $\mu_w(x)$, they require instead the equally high-dimensional nonparametric estimation of the propensity score. In practice the relative merits of these estimators will depend on whether the propensity score is more or less smooth than the regression functions, or whether additional information is available about either the propensity score or the regression functions.

3.3.1 Weighting

The first set of "propensity score" estimators use the propensity score as weights to create a balanced sample of treated and control observations. 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 $\tau^P = \mathbb{E}[Y_i(1) - Y_i(0)]$ 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, one can undo this imbalance. Formally, weighting estimators rely on the equalities:

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

and similarly

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

implying

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

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

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right). \tag{5}$$

In this particular form this is not necessarily an 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 to one. Specifically, in (5), 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. One approach for improving this estimator is simply to normalize the weights to unity. One can further normalize the weights to unity within subpopulations as defined by the covariates. In the limit this leads to the estimator proposed by Hirano, Imbens and Ridder (2003) who suggest using a nonparametric series estimator for $e(x)$. More precisely, they first specify a sequence of functions of the covariates, e.g., a power series, $h_l(x)$, $l = 1,\ldots,\infty$. 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

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

by maximizing the associated likelihood function. Let $\hat{\gamma}_L$ be the maximum likelihood estimate. In the third step, the estimated propensity score is calculated as:

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

Finally they estimate the average treatment effect as:

$$\hat{\tau}_{\text{weight}} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1-W_i) \cdot Y_i}{1-\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{(1-W_i)}{1-\hat{e}(X_i)}. \tag{6}$$

Hirano, Imbens and Ridder (2003) show that this estimator is efficient, whereas with the true propensity score the estimator would not be fully efficient (and in fact not very attractive).

This estimator highlights one of the interesting features of the problem of efficiently estimating average treatment effects. One solution is to estimate the two regression functions $\mu_w(x)$ nonparametrically; that solution 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 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 approach with a fully nonparametric estimator for the propensity score is numerically identical to the regression approach with a fully nonparametric estimator for the two regression functions.

One difficulty with the weighting estimators that are based on the estimated propensity score is again the problem of choosing the smoothing parameters. Hirano, Imbens and Ridder (2003) use series estimators, which requires choosing the number of terms in the series. Ichimura and Linton (2001) consider a kernel version, which involves choosing a bandwidth. Theirs is currently one of the few studies considering optimal choices for smoothing parameters that focuses specifically on estimating average treatment effects. A departure from standard problems in choosing smoothing parameters is that here one wants to use nonparametric regression methods even if the propensity score is known. For example, if the probability of treatment is constant, standard optimality results would suggest using a high degree of smoothing, as this would lead to the most accurate estimator for the propensity score. 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 "blocking propensity score" estimator. Using the (estimated) propensity score, divide the sample into M blocks of units of approximately equal probability of treatment, letting 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 $m = 1, \ldots, M-1$, so that

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

for $m = 1, \ldots, M$. Within each block there are N_{wm} observations with treatment equal to w, $N_{wm} = \sum_i 1\{W_i = w, J_{im} = 1\}$. Given these subgroups, 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:

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

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 the rate at which the number of blocks increases with the sample size. With the propensity score known, these are easy to determine; no formal results have been established for the unknown case.

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 95% of the bias associated with that covariate. Since all bias, under unconfoudnedness, is

associated with the propensity score, this suggests that under normality 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), and has been implemented in STATA by Becker and Ichino (2002). Often, however, researchers subsequently check the balance of the covariates within each block. If the true propensity score per block 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 assess the adequacy of the statistical model by comparing the distribution of the covariates among treated and controls within blocks. If the distributions are found to be different, one can either split the blocks into a number of subblocks, or 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 unbalanced, the specification of the propensity score is not adequate. In the illustrations in the next lecture a particular algorithm is decribed for choosing the blocks.

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 $e(X)$ and average the difference. Although this method has been used in practice, there is no particular reason why this is an attractive method compared to the regression methods based on the covariates directly. In addition, the large sample properties have not been established.

3.3.4 Matching on the Propensity Score

The Rosenbaum-Rubin result implies that it is sufficient to adjust solely for differences in the propensity score between treated and control units. Since one of the ways in which one can adjust for differences in covariates is matching, another natural way to use the propensity score is through matching. Because the propensity score is a scalar function of the covariates,

the bias results in Abadie and Imbens (2002) imply that the bias term is of lower order than the variance term and matching leads to a $\sqrt{N}$-consistent, asymptotically normally distributed estimator. The variance for the case with matching on the true propensity score also follows directly from their results. More complicated is the case with matching on the estimated propensity score. We are not aware of any results that give the asymptotic variance for this case.

3.4. Mixed Methods

A number of approaches have been proposed that combine two of the three methods described earlier, typically regression with one of its alternatives. These methods appear to be the most attractive in practice. The motivation for these combinations is that, although one method alone is often sufficient to obtain consistent or even efficient estimates, incorporating 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 benefit associated with combining methods 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 the parametric model for either the propensity score or the regression functions is specified correctly, the resulting estimator for the average treatment effect is consistent. Similarly, because matching is consistent with few assumptions beyond strong ignorability, thus methods that combine matching and regressions are robust against misspecification of the regression function.

3.4.1 Weighting and Regression

One can rewrite the HIR weighting estimator discussed above 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; 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, Rotnitzky and Zhao (1995), Robins and Ritov (1997), and implemented in Hirano and Imbens (2001). In the parametric context Robins and Ritov argue that the estimator is consistent as long as either the regression model or the propensity score (and thus the weights) are 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 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. As above, one can also add covariates to the regression function

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

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 large, additional bias corrections through regression can be particularly relevant in this case. A number of such corrections have been proposed, first by Rubin (1973b) and Quade (1982) in a parametric setting. Let $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ be the observed or imputed potential outcomes for unit i; where these estimated potential outcomes equal observed outcomes for some 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 close because of the matching process.

To further explore this, focusing on the single match case, define for each unit:

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

If the matching is exact $\hat{X}_i(0) = \hat{X}_i(1)$ for each unit. If not, these discrepancies will 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 ($W_i = 1$), so that $\hat{Y}_i(1) = Y_i(1)$ and $\hat{Y}_i(0)$ is an imputed value for $Y_i(0)$. This imputed value is unbiased for $\mu_0(X_{\ell(i)})$ (since $\hat{Y}_i(0) = Y_{\ell(i)}$), but not necessarily for $\mu_0(X_i)$. One may therefore 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 difference in the covariates for units i and its match, that is, of the form $\beta_0'(\hat{X}_i(1) - \hat{X}_i(0) = \beta_0'(X_i - X_{\ell(i)})$. One proposed correction is to estimate $\mu_0(x)$ directly by taking the control units that are used as matches for the treated units, with weights corresponding to the number of times a control observations is used as a match, and estimate a linear regression of the form

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

on the weighted control observations by least squares. (If unit i is a control unit the correction would be done using an estimator for the regression function $\mu_1(x)$ based on a linear specification $Y_i = \alpha_1 + \beta_1' X_i$ estimated on the treated units.) AI show that if this correction is done nonparametrically, the resulting matching estimator is consistent and asymptotically normal, with its bias dominated by the variance.

4. Estimating Variances

The variances of the estimators considered so far typically involve unknown functions. For example, as discussed earlier, the variance of efficient estimators of PATE is equal to

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

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

4.1 Estimating The Variance of Efficient Estimators for τ_P

For efficient estimators for τ_P the asymptotic variance is equal to the efficiency bound V_P. There are a number of ways we can estimate this. 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 can be estimated consistently. However, if one estimates the average treatment effect using only the two regression functions, it is an additional burden to estimate the conditional variances and the propensity score in order to estimate V_P. Similarly, if one efficiently estimates the average treatment effect by weighting with 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 this way. Under some conditions that will lead to valid standard errors and confidence

intervals.

A third approach is to use bootstrapping (Efron and Tibshirani, 1993; Horowitz, 2002). Although there is little formal evidence specific for these estimators, given that the estimators are asymptotically linear, it is likely that bootstrapping will lead to valid standard errors and confidence intervals at least for the regression and propensity score methods. Bootstrapping is not valid for matching estimators, as shown by Abadie and Imbens (2007) Subsampling (Politis and Romano, 1999) will still work in this setting.

4.2 Estimating The Conditional Variance

Here we focus on estimation of the variance of estimators for τ_S, which is the conditional variance of the various estimators, conditional on the covariates $\mathbf{X}$ and the treatment indicators $\mathbf{W}$. All estimators used in practice are linear combinations of the outcomes,

$$\hat{\tau} = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W}) \cdot Y_i,$$

with the $\lambda(\mathbf{X}, \mathbf{W})$ known functions of the covariates and treatment indicators. Hence the conditional variance is

$$V\left(\hat{\tau}|\mathbf{X}, \mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W})^2 \cdot \sigma^2_{W_i}(X_i).$$

The only unknown component of this variance is $\sigma^2_w(x)$. Rather than estimating this through nonparametric regression, AI suggest using matching to estimate $\sigma^2_w(x)$. To estimate $\sigma^2_{W_i}(X_i)$ one uses the closest match within the set of units with the same treatment indicator. Let $v(i)$ be the closest unit to i with the same treatment indicator ($W_{v(i)} = W_i$). The sample variance of the outcome variable for these 2 units can then be used to estimate $\sigma^2_{W_i}(X_i)$:

$$\hat{\sigma}^2_{W_i}(X_i) = \left(Y_i - Y_{v(i)}\right)^2 /2.$$

Note that this estimator is not consistent estimators of the conditional variances. However this is not important, as we are interested not in the variances at specific points in the

covariates distribution, but in the variance of the average treatment effect. Following the process introduce above, this is estimated as:

$$\hat{V}\left(\hat{\tau}|\mathbf{X},\mathbf{W}\right)=\sum_{i=1}^{N}\lambda_i(\mathbf{X},\mathbf{W})^2\cdot\hat{\sigma}^2_{W_i}(X_i).$$

References

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics,"Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Chen, X., H. Hong, and Tarozzi, (2005), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects,"unpublished working paper, Department of Economics, New York University.

Dehejia, R. (2005) "Program Evaluation as a Decision Problem,"*Journal of Econometrics*, 125, 141-173.

Engle, R., D. Hendry, and J.-F. Richard, (1983) "Exogeneity,"*Econometrica*, 51(2): 277-304.

Firpo, S. (2003), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259-276.

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

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261–294.

Heckman, J., R. Lalonde, and J. 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 J. Porter, (2005), "Asymptotics for Statistical Decision Rules,"Working Paper, Dept of Economics, University of Wisconsin.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average

Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189. July

Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, Vol. 87, No. 3, 706-710.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and J. Wooldridge., (2007), "Recent Developments in the Econometrics of Program Evaluation," unpublished manuscript, department of economics, Harvard University.

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

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.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., (2004), "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, 72(4), 1221-1246.

Manski, C. (2005), *Social Choice with Partial Knowledge of Treatment Response*, Princeton University Press.

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.

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.

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

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

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", *A*nnals of Statistics, 6:34–58.

Rubin, D., (1990), "Formal Modes of Statistical Inference for Causal Effects", emphJournal of Statistical Planning and Inference, 25, 279-292.