The Assumptions Underlying Evaluation Estimators*

James J. Heckman**

Abstract

This paper reviews the basic principles underlying the identification of econometric evaluation estimators and their recent extensions. It considers the choice of the estimator in terms of the decision model used by agents and the information they have available. Building on Barros's pioneering work [Barros (1987), reprinted here] I apply the analysis to the method of matching.

Keywords: Evaluation Estimators, Policy Evaluation.

JEL Codes: C13.

Introduction

This paper reviews the basic principles underlying the identification of conventional econometric evaluation estimators and their recent extensions. It is in four parts. The first part discusses at an intuitive level general identification assumptions for leading econometric estimators. The second part presents a prototypical economic choice model. It is a framework that is useful for analyzing and motivating the economic assumptions underlying alternative estimators. The third part discusses the economic assumptions underlying the method of matching. The fourth part concludes.

---

*Submitted in November 2011. Revised in December 2011. University of Chicago, Department of Economics, 1126 E. 59th Street, Chicago IL 60637, USA. This research was supported by NICHD: R01-HD32058-03 and NIH: R01-HD054702, R01-HD065072, The J. B. and M. K. Pritzker Foundation, The American Bar Foundation, and a subcontract from UMich on the PSID: AG-029409. I thank Richard Blundell, Miriam Gensowski, Mohan Singh, Edward Vytlačil, and Ben Williams for useful comments on various drafts of this paper. For a more extensive discussion of these issues, see the discussion in Heckman and Vytlačil (2007a,b).

**The University of Chicago, University College Dublin, and the American Bar Foundation.

1. The Basic Principles Underlying the Identification of the Leading Econometric Evaluation Estimators

This paper reviews the main principles underlying the evaluation estimators commonly used in the econometrics literature. I assume two potential outcomes (Y_0, Y_1) . $D = 1$ if Y_1 is observed, and $D = 0$ corresponds to Y_0 being observed. The observed outcome is

$$Y = DY_1 + (1 - D)Y_0. \quad (1)$$

The *evaluation problem* arises because for each person we observe either Y_0 or Y_1 but not both. Thus in general it is not possible to identify the individual level treatment effect $Y_1 - Y_0$ for any person. The typical solution to this problem is to reformulate the problem at the population level rather than at the individual level and to identify certain mean outcomes, quantile outcomes, or various distributions of outcomes as described in Heckman and Vytlačil (2007a). For example, a commonly used approach focuses attention on average treatment effects, such as $ATE = E(Y_1 - Y_0)$.

If treatment is assigned or chosen on the basis of potential outcomes, so

$$(Y_0, Y_1) \not\perp D,$$

where $\not\perp$ denotes “is not independent” and $\perp$ denotes “independent,” we encounter the problem of selection bias. Suppose that we observe people in each treatment state $D = 0$ and $D = 1$. If $Y_j \not\perp D$, then the observed Y_j will be selectively different from randomly assigned Y_j , $j = 0, 1$. Thus $E(Y_0 | D = 0) \neq E(Y_0)$ and $E(Y_1 | D = 1) \neq E(Y_1)$. Using unadjusted data to construct $E(Y_1 - Y_0)$ will produce one source of evaluation bias:

$$E(Y_1 | D = 1) - E(Y_0 | D = 0) \neq E(Y_1 - Y_0).$$

The selection problem and the evaluation problem are conceptually distinct. Many methods have been proposed to solve both problems.

The method with the greatest intuitive appeal, which is sometimes called the “gold standard” in evaluation analysis, is the method of random assignment. Non-experimental methods can be organized by how they attempt to approximate what can be obtained by an ideal random assignment. If treatment is chosen at random with respect to (Y_0, Y_1) , or if treatments are randomly assigned and there is full compliance with the treatment assignment,

$$(R-1) (Y_0, Y_1) \perp D.$$

It is useful to distinguish several cases where (R-1) will be satisfied. The first is when agents (decision makers whose choices are being analyzed) pick outcomes

that are random with respect to (Y_0, Y_1) . Thus agents may not know (Y_0, Y_1) at the time they make their choices to participate in treatment or at least do not act on (Y_0, Y_1) , so that $\Pr(D = 1 \mid X, Y_0, Y_1) = \Pr(D = 1 \mid X)$ for all X . Matching assumes a version of (R-1) conditional on matching variables X : $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.

A second case arises when individuals are randomly assigned to treatment status even if they would choose to self-select into no-treatment status, and they comply with the randomization protocols. Let ξ be randomized assignment status. With full compliance, $\xi = 1$ implies that Y_1 is observed and $\xi = 0$ implies that Y_0 is observed. Then, under randomized assignment,

$$(R-2) (Y_0, Y_1) \perp\!\!\!\perp \xi,$$

even if in a regime of self-selection, $(Y_0, Y_1) \not\perp\!\!\!\perp D$. If randomization is performed conditional on X , we obtain $(Y_0, Y_1) \perp\!\!\!\perp \xi \mid X$.

Let A denote actual treatment status. If the randomization has full compliance among participants, $\xi = 1 \Rightarrow A = 1$; $\xi = 0 \Rightarrow A = 0$. This is entirely consistent with a regime in which a person would choose $D = 1$ in the absence of randomization, but would have no treatment ($A = 0$) if suitably randomized, even though the agent might desire treatment.

If treatment status is chosen by self-selection, $D = 1 \Rightarrow A = 1$ and $D = 0 \Rightarrow A = 0$. If there is imperfect compliance with randomization, $\xi = 1 \not\Rightarrow A = 1$ because of agent choices. In general, $A = \xi D$ so that $A = 1$ only if $\xi = 1$ and $D = 1$. If treatment status is randomly assigned, either through randomization or randomized self-selection,

$$(R-3) (Y_0, Y_1) \perp\!\!\!\perp A.$$

This version of randomization can also be defined conditional on X . Under (R-1), (R-2), or (R-3), the average treatment effect (ATE) is the same as the marginal treatment effect (MTE) of Björklund and Moffitt (1987) and Heckman and Vytlačil (1999, 2005, 2007a), and the same as the parameters treatment on the treated (TT) ($E(Y_1 - Y_0 \mid D = 1)$) and treatment on the untreated (TUT) ($E(Y_1 - Y_0 \mid D = 0)$).¹ These parameters can be identified from population means:

$$TT = MTE = TUT = ATE = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

Forming averages over populations of persons who are treated ($A = 1$) or untreated ($A = 0$) suffices to identify this parameter. If there are conditioning variables X , we can define the mean treatment parameters for all X where (R-1) or (R-2) or (R-3) hold.

Observe that even with random assignment of treatment status and full compli-

---

¹The marginal treatment effect is formally defined in the next section.

ance, one cannot, in general, identify the distribution of the treatment effects ($Y_1 - Y_0$), although one can identify the marginal distributions $F_1(Y_1 | A = 1, X = x) = F_1(Y_1 | X = x)$ and $F_0(Y_0 | A = 0, X = x) = F_0(Y_0 | X = x)$. One special assumption, common in the conventional econometrics literature, is that $Y_1 - Y_0 = \Delta(X)$, a constant given $X = x$. Since $\Delta(X)$ can be identified from $E(Y_1 | A = 1, X = x) - E(Y_0 | A = 0, X = x)$ if A is allocated by randomization, in this special case the analyst can identify the joint distribution of (Y_0, Y_1) .² This approach assumes that (Y_0, Y_1) have the same distribution up to a parameter $\Delta(X)$ (Y_0 and Y_1 are perfectly dependent). One can make other assumptions about the dependence across ranks from perfectly positive or negative ranking to independence.³ The joint distribution of (Y_0, Y_1) or of $(Y_1 - Y_0)$ is not identified unless the analyst can pin down the dependence across (Y_0, Y_1) . Thus, even with data from a randomized trial one cannot, without further assumptions, identify the proportion of people who benefit from treatment in the sense of gross gain ($\Pr(Y_1 \geq Y_0)$). This problem plagues all evaluation methods. Abbring and Heckman (2007) discuss methods for identifying joint distributions of outcomes.⁴

Assumption (R-1) is very strong. In many cases, it is thought that there is *selection bias* with respect to (Y_0, Y_1) , so persons who select into status 1 or 0 are selectively different from randomly sampled persons in the population. The assumption most commonly made to circumvent problems with (R-1) is that even though D is not random with respect to potential outcomes, the analyst has access to control variables X that effectively produce a randomization of D with respect to (Y_0, Y_1) given X . This is the method of matching, which is based on the following conditional independence assumption:

$$(M-1) (Y_0, Y_1) \perp\!\!\!\perp D \mid X.$$

Conditioning on X randomizes D with respect to (Y_0, Y_1) . (M-1) assumes that any selective sampling of (Y_0, Y_1) can be adjusted by conditioning on observed variables. (R-1) and (M-1) are different assumptions and neither implies the other. In a linear equation model, assumption (M-1), that D is independent from (Y_0, Y_1) given X , justifies application of least squares on D and X to eliminate selection bias in mean outcome parameters. For means, matching is just nonparametric regression.⁵ In order to be able to compare X -comparable people in the treatment regime one must assume

$$(M-2) 0 < \Pr(D = 1 \mid X = x) < 1.$$

²Heckman (1992), Heckman, Smith, and Clements (1997).

³Heckman, Smith, and Clements (1997).

⁴See also Aakvik, Heckman, and Vytlačil (2005); Carneiro, Hansen, and Heckman (2001, 2003); and Cunha, Heckman, and Navarro (2005).

⁵Barnow, Cain, and Goldberger (1980) present one application of matching in a regression setting.

Assumptions (M-1) and (M-2) justify matching. Assumption (M-2) is required for *any* evaluation estimator that compares treated and untreated persons. It is produced by random assignment if the randomization is conducted for all $X = x$ in the relevant support and there is full compliance.

Observe that from (M-1) and (M-2), it is possible to identify $F_1(Y_1 | X = x)$ from the observed data $F_1(Y_1 | D = 1, X = x)$ since we observe the left hand side of

$$F_1(Y_1 | D = 1, X = x) = F_1(Y_1 | X = x) = F_1(Y_1 | D = 0, X = x).$$

The first equality is a consequence of conditional independence assumption (M-1). The second equality comes from (M-1) and (M-2). By a similar argument, we observe the left hand side of

$$F_0(Y_0 | D = 0, X = x) = F_0(Y_0 | X = x) = F_0(Y_0 | D = 1, X = x),$$

and the equalities are a consequence of (M-1) and (M-2). Since the pair of outcomes (Y_0, Y_1) is not identified for anyone, as in the case of data from randomized trials, the joint distributions of (Y_0, Y_1) given X or of $Y_1 - Y_0$ given X are not identified without further information. This is a problem that plagues all evaluation estimators.

From the data on Y_1 given X and $D = 1$ and the data on Y_0 given X and $D = 0$, since $E(Y_1 | D = 1, X = x) = E(Y_1 | X = x) = E(Y_1 | D = 0, X = x)$ and $E(Y_0 | D = 0, X = x) = E(Y_0 | X = x) = E(Y_0 | D = 1, X = x)$ we obtain

$$E(Y_1 - Y_0 | X = x) = E(Y_1 - Y_0 | D = 1, X = x) = E(Y_1 - Y_0 | D = 0, X = x).$$

Effectively, we have a randomization for the subset of the support of X satisfying (M-2).

At values of X that fail to satisfy (M-2), there is no variation in D given X . One can define the residual variation in D not accounted for by X as

$$\mathcal{E}(x) = D - E(D | X = x) = D - \Pr(D = 1 | X = x).$$

If the variance of $\mathcal{E}(x)$ is zero, it is not possible to contrast outcomes by treatment status for those X values and (M-2) is violated. To see the consequences of this violation in a regression setting, use $Y = Y_0 + D(Y_1 - Y_0)$ and take conditional expectations, under (M-1), to obtain

$$E(Y | X, D) = E(Y_0 | X) + D[E(Y_1 - Y_0 | X)].^6$$

---

⁶This follows because $E(Y | X, D) = E(Y_0 | X, D) + DE(Y_1 - Y_0 | X, D)$ but from (M-1), $E(Y_0 | X, D) = E(Y_0 | X)$ and $E(Y_1 - Y_0 | X, D) = E(Y_1 - Y_0 | X)$.

If $\text{Var}(\mathcal{E}(x)) > 0$ for all x in the support of X , one can use nonparametric least squares to identify $E(Y_1 - Y_0 \mid X = x) = \text{ATE}(x)$ by regressing Y on D and X . The function identified from the coefficient on D is the average treatment effect.⁷ If $\text{Var}(\mathcal{E}(x)) = 0$, $\text{ATE}(x)$ is not identified at that x value because there is no variation in D that is not fully explained by X . A special case of matching is linear least squares where one can write

$$Y_0 = X\alpha + U_0 \quad \text{and} \quad Y_1 = X\alpha + \beta + U_1,$$

where $U_0 = U_1 = U$ and hence under (M-1),

$$E(Y \mid X, D) = \varphi(X) + \beta D,$$

where $\varphi(X) = X\alpha + E(U \mid X)$. If D is perfectly predictable by X , one cannot identify β because of a multicollinearity problem. (M-2) rules out perfect collinearity. Matching is a nonparametric version of least squares that does not impose functional form assumptions on outcome equations, and that imposes support condition (M-2). Matching identifies β but not, in general, α . But if the goal is to estimate treatment effects, this is all that is required.⁸ This is an application of Marschak's Maxim (Heckman, 2008, 2010). We do not need to estimate the full model to estimate β .

Conventional econometric choice models make a distinction between variables that appear in outcome equations (X) and variables that appear in choice equations (Z). The same variables may be in (X) and (Z) but more typically, there are some variables not in common. For example, the instrumental variable estimator is based on variables that are not in X but that are in Z . Matching makes no distinction between the X and the Z .⁹ It does not rely on exclusion restrictions. The conditioning variables used to achieve conditional independence can in principle be a set of variables Q distinct from the X variables (covariates for outcomes) or the Z variables (covariates for choices). I use X solely to simplify the notation. The key identifying assumption is the assumed existence of a random variable X with the properties satisfying (M-1) and (M-2).

Conditioning on a larger vector (X augmented with additional variables) or a smaller vector (X with some components removed) may or may not produce suitably modified versions of (M-1) and (M-2). Without invoking further assumptions there is no objective principle for determining what conditioning variables produce (M-1).

⁷Under the conditional independence assumption (M-1), it is also the effect of treatment on the treated $E(Y_1 - Y_0 \mid X, D = 1)$ and the marginal treatment effect formally defined in the next section.

⁸See Heckman and Robb (1985a).

⁹Heckman, Ichimura, Smith, and Todd (1998) distinguish X and Z in matching. They consider a case where conditioning on X may lead to failure of (M-1) and (M-2) but conditioning on (X, Z) satisfies a suitably modified version of this condition.

Assumption (M-1) is strong. Many economists do not have enough faith in their data to invoke it. Assumption (M-2) is testable and requires no act of faith. To justify (M-1), it is necessary to appeal to the quality of the data.

Using economic theory can help guide the choice of an evaluation estimator. A crucial distinction is the one between the information available to the analyst and the information available to the agent whose outcomes are being studied. Assumptions made about these information sets drive the properties of econometric estimators. Analysts using matching make strong informational assumptions in terms of the data available to them. In fact, all econometric estimators make assumptions about the presence or absence of informational asymmetries, and I exposit them in this paper.

To analyze the informational assumptions invoked in matching, and other econometric evaluation strategies, following the analysis of Barros (1987) it is helpful to introduce five distinct information sets and establish some relationships among them.¹⁰ (1) An information set $\sigma(I_{R^*})$ with an associated random variable that satisfies conditional independence (M-1) is defined as a *relevant* information set; (2) The minimal information set $\sigma(I_R)$ with associated random variable needed to satisfy conditional independence (M-1), the *minimal relevant* information set; (3) The information set $\sigma(I_A)$ available to the agent at the time decisions to participate are made; (4) The information available to the economist, $\sigma(I_{E^*})$; and (5) The information $\sigma(I_E)$ used by the economist in conducting an empirical analysis. I will denote the random variables generated by these sets as $I_{R^*}, I_R, I_A, I_{E^*}, I_E$, respectively.¹¹

Definition 1. Define $\sigma(I_{R^*})$ as a **relevant information set** if the information set is generated by the random variable I_{R^*} , possibly vector valued, and satisfies condition (M-1), so

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid I_{R^*}.$$

Definition 2. Define $\sigma(I_R)$ as a **minimal relevant information set** if it is the intersection of all sets $\sigma(I_{R^*})$ and satisfies $(Y_0, Y_1) \perp\!\!\!\perp D \mid I_R$. The associated random variable I_R is a minimum amount of information that guarantees that condition (M-1) is satisfied. There may be no such set.¹²

¹⁰See the discussion in Barros (1987), Heckman and Navarro (2004), and Gerfin and Lechner (2002).

¹¹I start with a primitive probability space (Ω, σ, P) with associated random variables I . I assume minimal σ -algebras and assume that the random variables I are measurable with respect to these σ -algebras. Obviously, strictly monotonic or affine transformations of the I preserve the information and can substitute for the I .

¹²Observe that the intersection of all sets $\sigma(I_{R^*})$ may be empty and hence may not be characterized by a (possibly vector valued) random variable I_R that guarantees $(Y_0, Y_1) \perp\!\!\!\perp D \mid I_R$. If the information sets that produce conditional independence are nested, then the intersection of all sets $\sigma(I_{R^*})$ producing conditional independence is well defined and has an associated random variable I_R with the required property, although it may not be unique (e.g., strictly monotonic

If one defines the relevant information set as one that produces conditional independence, it may not be unique. If the set $\sigma(I_{R^*})$ satisfies the conditional independence condition, then the set $\sigma(I_{R^*}, Q)$ such that $Q \perp\!\!\!\perp (Y_0, Y_1) \mid I_{R^*}$ would also guarantee conditional independence. For this reason, when it is possible to do so I define the relevant information set to be minimal, that is, to be the intersection of all relevant sets that still produce conditional independence between (Y_0, Y_1) and D . However, no minimal set may exist.

Definition 3. *The agent's information set, $\sigma(I_A)$, is defined by the information I_A used by the agent when choosing among treatments. Accordingly, I call I_A the agent's information.*

By the agent I mean the person making the treatment decision, not necessarily the person whose outcomes are being studied (e.g. the agent may be the parent; the person being studied may be a child).

Definition 4. *The econometrician's **full information set**, $\sigma(I_{E^*})$, is defined as all of the information available to the econometrician, I_{E^*} .*

Definition 5. *The **econometrician's information set**, $\sigma(I_E)$, is defined by the information **used** by the econometrician when analyzing the agent's choice of treatment, I_E , in conducting an analysis.*

For the case where a unique minimal relevant information set exists, only three restrictions are implied by the structure of these sets: $\sigma(I_R) \subseteq \sigma(I_{R^*})$, $\sigma(I_R) \subseteq \sigma(I_A)$, and $\sigma(I_E) \subseteq \sigma(I_{E^*})$.¹³ I have already discussed the first restriction. The second restriction requires that the minimal relevant information set must be part of the information the agent uses when deciding which treatment to take or assign. It is the information in $\sigma(I_A)$ that gives rise to the selection problem which in turn gives rise to the evaluation problem. The third restriction is trivial.

Aside from these orderings, the econometrician's information set may be different from the agent's or the relevant information set. The econometrician may know something the agent doesn't know, for typically he is observing events after the decision is made. At the same time, there may be private information known to the agent but not the econometrician. Matching assumption (M-1) implies that $\sigma(I_R) \subseteq \sigma(I_E)$, so that the econometrician uses at least the minimal rele-

---

transformations and affine transformations of I_R also preserve the property). In the more general case of non-nested information sets with the required property, it is possible that no uniquely defined minimal relevant set exists. Among collections of nested sets that possess the required property, there is a minimal set defined by intersection but there may be multiple minimal sets corresponding to each collection.

¹³This formulation assumes that the agent makes the treatment decision. The extension to the case where the decision maker and the agent are distinct is straightforward.

vant information set, but of course he or she may use more. However, using more information is not guaranteed to produce a model with conditional independence property (M-1) satisfied for the augmented model. Thus an analyst can “overdo” it. Heckman and Navarro (2004) present examples of the consequences of the asymmetry in agent and analyst information sets which I review in Section 3.

The possibility of asymmetry in information between the agent making participation decisions and the observing economist creates the potential for a major identification problem that is ruled out by assumption (M-1). The methods of control functions and instrumental variables estimators (and closely related regression discontinuity design methods) address this problem in different ways. Accounting for this possibility is a more conservative approach to the selection problem than the one taken by advocates of least squares, or its nonparametric counterpart, matching. Those advocates assume that they know the X that produces a relevant information set. Below, I show the biases that can result in matching when standard econometric model selection criteria are applied to pick the X that are used to satisfy (M-1). Conditional independence condition (M-1) cannot be tested without maintaining other assumptions.¹⁴ Choice of the appropriate conditioning variables is a problem that plagues *all* econometric estimators.

The methods of control functions, replacement functions, proxy variables, and instrumental variables all recognize the possibility of asymmetry in information between the agent being studied and the econometrician and recognize that even after conditioning on X (variables in the outcome equation) and Z (variables affecting treatment choices, which may include the X), analysts may fail to satisfy conditional independence condition (M-1).¹⁵ These methods postulate the existence of some unobservables θ , which may be vector valued, with the property that

$$(U-1) (Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta,$$

but allow for the possibility that

$$(U-2) (Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z.$$

In the event (U-2) holds, approaches based on (U-1) model the relationships of the unobservable θ with Y_1, Y_0 and D in various ways. The content in the

¹⁴These assumptions may or may not be testable. I discuss the required “exogeneity” conditions below in my discussion of matching in Section 3. Thus randomization of assignment of treatment status might be used to test (M-1), but this requires that there be full compliance and that the randomization be valid (no anticipation effects or general equilibrium effects).

¹⁵The term and concept of control function is due to Heckman and Robb (1985a,b, 1986a,b). See Blundell and Powell (2003) (who call a nonparametric version of the Heckman and Robb replacement functions “control functions”). A more recent nomenclature is “control variate”. Matzkin (2007) provides a comprehensive discussion of identification principles for econometric estimators.

control function principle is to specify the exact nature of the dependence on the relationship between observables and unobservables in a nontrivial fashion that is consistent with economic theory. I present examples of models that satisfy (U-1) and (U-2) in Section 3.

The early literature focused on mean outcomes conditional on covariates (Heckman and Robb, 1985a,b, 1986a,b) and assumes a weaker version of (U-1) based on conditional mean independence rather than full conditional independence. More recent work analyzes distributions of outcomes (e.g., Aakvik, Heckman, and Vytlacil, 2005, Carneiro, Hansen, and Heckman, 2001, 2003). This work is reviewed in Abbring and Heckman (2007).

The normal Roy (1951) selection model makes distributional assumptions and identifies the joint distribution of outcomes. A large literature surveyed by Matzkin (2007) makes alternative assumptions to satisfy (U-1) in nonparametric settings. Replacement functions (Heckman and Robb, 1985a) are methods that proxy θ . They substitute out for θ , or components of θ , using observables.¹⁶ Aakvik, Heckman, and Vytlacil (1999, 2005), Carneiro, Hansen, and Heckman (2001, 2003), Cunha, Heckman, and Navarro (2005), and Cunha, Heckman, and Schennach (2010, 2006) develop methods that integrate out θ from the model assuming $\theta \perp\!\!\!\perp (X, Z)$, or invoking weaker mean independence assumptions, and assuming access to proxy measurements for θ . They also consider methods for estimating the distributions of treatment effects. These are discussed in Abbring and Heckman (2007).

The normal selection model produces partial identification of a generalized Roy model and full identification of a Roy model under separability and normality. It models the conditional expectation of U_0 and U_1 given X, Z, D . In terms of (U-1), it models the conditional mean dependence of (Y_0, Y_1) on D and θ given X and Z . Powell (1994) and Matzkin (2007) survey methods for producing semiparametric versions of these models. Heckman and Vytlacil (2007a, Appendix B) present a prototypical identification proof for a general selection model that implements (U-1) by estimating the distribution of θ , assuming $\theta \perp\!\!\!\perp (X, Z)$, and invoking support conditions on (X, Z) .

Central to both the selection approach and the instrumental variable approach for a model with heterogeneous responses is the probability of selection. Let Z denote variables in the choice equation. Fixing Z at different values (denoted z) I define $D(z)$ as an indicator function that is “1” when treatment is selected at the fixed value of z and that is “0” otherwise. In terms of a separable index model $U_D = \mu_D(Z) - V$, for a fixed value of z ,

$$D(z) = \mathbf{1}[\mu_D(z) \geq V]$$

where $Z \perp\!\!\!\perp V \mid X$. Thus fixing $Z = z$, values of z do not affect the realizations of V

¹⁶This is the “control variate” of Blundell and Powell (2003). Heckman and Robb (1985a) and Olley and Pakes (1996) use a similar idea. Matzkin (2007) discusses replacement functions.

for any value of X . An alternative way of representing the independence between Z and V given X due to Imbens and Angrist (1994), writes that $D(z) \perp\!\!\!\perp Z$ for all $z \in \mathcal{Z}$, where $\mathcal{Z}$ is the support of Z . The Imbens and Angrist independence condition for IV is

$$\{D(z)\}_{z \in \mathcal{Z}} \perp\!\!\!\perp Z \mid X.$$

Thus the probabilities that $D(z) = 1$, $z \in \mathcal{Z}$, are not affected by the occurrence of Z . Vytlačil (2002) establishes the equivalence of these two formulations under general conditions.¹⁷

The method of instrumental variables (IV) postulates that

$$(IV-1) (Y_0, Y_1, \{D(z)\}_{z \in \mathcal{Z}}) \perp\!\!\!\perp Z \mid X. \text{ (*Independence*)}$$

One consequence of this assumption is that $E(D \mid X, Z) = P(X, Z)$, the propensity score, is random with respect to potential outcomes. Thus $(Y_0, Y_1) \perp\!\!\!\perp P(X, Z) \mid X$. So are all other functions of Z given X . The method of instrumental variables also assumes that

$$(IV-2) E(D \mid X, Z) = P(X, Z) \text{ is a nondegenerate function of } Z \text{ given } X. \text{ (*Rank Condition*)}$$

Comparing (IV-1) to (M-1), in the method of instrumental variables, Z is independent of (Y_0, Y_1) given X , whereas in matching D is independent of (Y_0, Y_1) given X . So in (IV-1), Z plays the role of D in matching condition (M-1). Comparing (IV-2) with (M-2), in the method of IV the choice probability $\Pr(D = 1 \mid X, Z)$ is assumed to vary with Z conditional on X , whereas in matching, D varies conditional on X . Unlike the method of control functions, no explicit model of the relationship between D and (Y_0, Y_1) is required in applying IV.

(IV-2) is a rank condition and can be empirically verified. (IV-1) is not testable as it involves assumptions about counterfactuals. In a conventional common coefficient regression model

$$Y = \alpha + \beta D + U,$$

where β is a constant and where $Cov(D, U) \neq 0$, (IV-1) and (IV-2) identify β .¹⁸ When β varies in the population and is correlated with D , additional assumptions must be invoked for IV to identify interpretable parameters. Heckman, Urzúa, and Vytlačil (2006) and Heckman and Vytlačil (2007b) discuss these conditions.

Assumptions (IV-1) and (IV-2), with additional assumptions in the case where β varies in the population, discussed in Heckman and Vytlačil (2007a,b) and in

¹⁷See Heckman and Vytlačil (2007b) for a discussion of these conditions.

¹⁸ $\beta = \frac{Cov(Z, Y)}{Cov(Z, D)}$.

this paper, can be used to identify mean treatment parameters. Replacing Y_1 with $\mathbf{1}(Y_1 \leq t)$ and Y_0 with $\mathbf{1}(Y_0 \leq t)$, where t is a constant, the IV approach allows us to identify marginal distributions $F_1(y_1 | X)$ or $F_0(y_0 | X)$.

In matching, the variation in D that arises after conditioning on X provides the source of randomness that switches people across treatment status. Nature is assumed to provide an experimental manipulation conditional on X that replaces the randomization assumed in (R-1)-(R-3). When D is perfectly predictable by X , there is no variation in it conditional on X , and the randomization by nature breaks down. Heuristically, matching assumes a residual $\mathcal{E}(X) = D - E(D | X)$ that is nondegenerate and is one manifestation of the randomness that causes persons to switch status.¹⁹

In the IV method, it is the choice probability $E(D | X, Z) = P(X, Z)$ that is random with respect to (Y_0, Y_1) , not components of D not predictable by (X, Z) . Variation in Z for a fixed X provides the required variation in D that switches treatment status and still produces the required conditional independence:

$$(Y_0, Y_1) \perp\!\!\!\perp P(X, Z) | X.$$

Variation in $P(X, Z)$ produces variations in D that switch treatment status. Components of variation in D not predictable by (X, Z) do not produce the required independence. Instead, the predicted component provides the required independence. It is just the opposite in matching. Versions of the method of control functions use measurements to proxy θ , or components of θ , in (U-1) and remove spurious dependence that gives rise to selection problems. These are called replacement functions (see Heckman and Robb, 1985a) or control variates (see Blundell and Powell, 2003).

The methods of replacement functions and proxy variables all start from characterizations (U-1) and (U-2). θ is not observed and (Y_0, Y_1) are not observed directly but Y is observed:

$$Y = DY_1 + (1 - D)Y_0.$$

Missing variables (θ) produce selection bias which creates a problem with using observational data to evaluate social programs. From (M-1), if one conditions on θ , condition (M-1) for matching would be satisfied, and hence one could identify the parameters and distributions that can be identified if the conditions required for matching are satisfied.

The most direct approach to controlling for θ is to assume access to a function $\tau(X, Z, Q)$ that perfectly proxies θ :

$$\theta = \tau(X, Z, Q). \quad (2)$$

¹⁹It is heuristically illuminating, but technically incorrect to replace $\mathcal{E}(X)$ with D in (R-1) or R in (R-2) or T in (R-3). In general $\mathcal{E}(X)$ is not independent of X even if it is mean independent.

This approach based on a perfect proxy is called the method of replacement functions by Heckman and Robb (1985a). In (U-1), one can substitute for θ in terms of observables (X, Z, Q) . Then

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, Q.$$

It is possible to condition nonparametrically on (X, Z, Q) and without having to know the exact functional form of τ . θ can be a vector and τ can be a vector of functions. This method has been used in the economics of education for decades (see the references in Heckman and Robb, 1985a). If θ is ability and τ is a test score, it is sometimes assumed that the test score is a perfect proxy (or replacement function) for θ and that one can enter it into the regressions of earnings on schooling to escape the problem of ability bias (typically assuming a linear relationship between τ and θ).²⁰ Heckman and Robb (1985a) discuss the literature that uses replacement functions in this way. Olley and Pakes (1996) apply this method and consider nonparametric identification of the τ function. The methods of Blundell and Powell can be interpreted as nonparametric replacement functions that generalize the principle. Matzkin (2007) provides a rigorous proof of identification for this approach in a general nonparametric setting.

The method of replacement functions assumes that (2) is a perfect proxy. In many applications, this assumption is far too strong. More often, θ is measured with error. This produces a factor model or measurement error model (Aigner, Hsiao, Kapteyn, and Wansbeek, 1984). Matzkin (2007) surveys this method. One can represent the factor model in a general way by a system of equations:

$$Y_j = g_j(X, Z, \theta, \varepsilon_j), \quad j = 0, 1. \quad (3)$$

A linear factor model separable in the unobservables writes

$$Y_j = g_j(X, Z) + \alpha_j \theta + \varepsilon_j, \quad j = 0, 1, \quad (4)$$

where

$$(X, Z) \perp\!\!\!\perp (\theta, \varepsilon_j), \varepsilon_j \perp\!\!\!\perp \theta, \quad j = 0, 1, \quad (5)$$

and the ε_j are mutually independent. Observe that under (3) and (4), Y_j controlling for X, Z only imperfectly proxies θ because of the presence of ε_j . θ is called a factor, α_j factor loadings and the ε_j “uniquenesses” (see, e.g. Aigner, 1985).

²⁰Thus if $\tau = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$, one can write

$$\theta = \tau - \alpha_0 - \alpha_1 X - \alpha_2 Q - \alpha_3 Z,$$

and use this as the proxy function. Controlling for τ, X, Q, Z controls for θ . Notice that one does not need to know the coefficients $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ to implement the method. One can condition on X, Q, Z .

A large literature, reviewed in Abbring and Heckman (2007) and Matzkin (2007) shows how to establish identification of econometric models under factor structure assumptions. Cunha, Heckman, and Matzkin (2003), Schennach (2004), Hu and Schennach (2008), and Cunha, Heckman, and Schennach (2010) establish identification in nonlinear and nonparametric models of the form (3).²¹ The key to identification is multiple, but imperfect (because of ε_j), measurements on θ from the Y_j , $j = 0, 1$ and X, Z, Q , and possibly other measurement systems that depend on θ . Carneiro, Hansen, and Heckman (2003), Cunha, Heckman, and Navarro (2005, 2006) and Cunha and Heckman (2008, 2007) apply and develop these methods. Under assumption (5), they show how to nonparametrically identify the econometric model and the distributions of the unobservables $F_\Theta(\theta)$ and $F_{\xi_j}(\varepsilon_j)$. In the context of classical simultaneous equations models, identification is secured by using covariance restrictions across equations exploiting the low dimensionality of vector θ compared to the high dimensional vector of (imperfect) measurements on it. The recent literature (Hu and Schennach, 2008, Cunha, Heckman, and Schennach, 2010) extends the linear model to a nonparametric setting.

The recent econometric literature applies, in special cases, the idea of the control function principle introduced in Heckman and Robb (1985a). This principle, versions of which can be traced back to Telser (1964), partitions θ in (U-1) into two or more components, $\theta = (\theta_1, \theta_2)$, where only one component of θ is the source of bias. Thus it is assumed that (U-1) is true, and (U-1)' is also true:

$$(U-1)' (Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta_1,$$

and (U-2) holds. For example, in a normal selection model with additive separability, one can break U_1 , the error term associated with Y_1 , into two components:

$$U_1 = E(U_1 \mid V) + \varepsilon,$$

where V plays the role of θ_1 and is associated with the choice equation. Under normality, ε is independent of $E(U_1 \mid V)$. Further,

$$E(U_1 \mid V) = \frac{\text{Cov}(U_1, V)}{\text{Var}(V)} V, \quad (6)$$

assuming $E(U_1) = 0$ and $E(V) = 0$. Heckman and Robb (1985a) show how to construct a control function in the context of the choice model

$$D = \mathbf{1}[\mu_D(Z) \geq V].$$

²¹Cunha, Heckman, and Schennach (2010, 2006) apply and extend this approach to a dynamic factor setting where the θ_t are time dependent.

Controlling for V controls for the component of θ_1 in $(U-1)'$ that gives rise to spurious dependence. The Blundell and Powell (2003, 2004) extension of the control function principle assumes functional form (6) but assumes that V can be perfectly proxied by a first stage equation. Thus they use a nonparametric version of a replacement function in their first stage. Their method does not work when one can only condition on D instead of directly measuring $D^* = \mu_D(Z) - V$.²² In the sample selection model, it is not necessary to estimate V . As developed in Heckman and Robb (1985a) and Heckman and Vytlacil (2007a,b), under additive separability for the outcome equation for Y_1 , one can write

$$E(Y_1 | X, Z, D = 1) = \mu_1(X) + \underbrace{E(U_1 | \mu_D(Z) \geq V)}_{\text{control function}},$$

so the analyst “expects out” rather than solves out the effect of the component of V on U_1 and thus controls for selection bias under the maintained assumptions. In terms of the propensity score, under the conditions specified in Heckman and Vytlacil (2007a), one may write the preceding expression in terms of $P(Z)$:

$$E(Y_1 | X, Z, D = 1) = \mu_1(X) + K_1(P(Z)),$$

where $K_1(P(Z)) = E(U_1 | X, Z, D = 1)$. It is not literally necessary to know V or be able to estimate it. The Blundell and Powell (2003, 2004) development of the control function principle assumes that the analyst can estimate and condition on V and that V is the source of endogeneity. The literature has begun to distinguish between the more general control function approach and the *control variate* approach that uses a first stage replacement function.

Matzkin (2003) develops the method of unobservable instruments which is a version of the replacement function approach applied to nonlinear models. Her unobservable instruments play the role of covariance restrictions used to identify classical simultaneous equations models (see Fisher, 1966). Her approach is distinct from and therefore complementary with linear factor models. Instead of assuming $(X, Z) \perp\!\!\!\perp (\theta, \varepsilon_j)$, she assumes in a two equation system where Y_1 and Y_0 are endogenous variables that $(\theta, \varepsilon_1) \perp\!\!\!\perp Y_1 | Y_0, X, Z$. See Matzkin (2007).

I do not discuss panel data methods. The most commonly used panel data method is difference-in-differences as discussed in Heckman and Robb (1985a), Blundell, Duncan, and Meghir (1998), Heckman, LaLonde, and Smith (1999), and Bertrand, Duflo, and Mullainathan (2004), to cite only a few of the key papers. Most of the estimators I have discussed can be adapted to a panel data setting. Heckman, Ichimura, Smith, and Todd (1998) develop difference-in-differences matching estimators. Abadie (2002) extends this work.²³ Separability between errors and observables is a key feature of the panel data approach in its standard

²²Imbens and Newey (2002) extend their approach. See the discussion in Matzkin (2007).

²³This work is applied in Athey and Imbens (2006).

application. Altonji and Matzkin (2005) and Matzkin (2003) present analyses of nonseparable panel data methods. Regression discontinuity estimators, which are versions of IV estimators, are discussed by Heckman and Vytlačil (2007b).

Table 1 summarizes some of the main lessons of this section. The stated conditions are necessary conditions. There are many versions of the IV and control functions principle and extensions of these ideas which refine these basic postulates. See Heckman and Vytlačil (2007b). Matzkin (2007) is an additional reference on sources of identification in econometric models.

I next turn to a detailed discussion of the method of matching which currently is “the method of choice” for many applied economists. Before analyzing matching, it is useful to have a prototypical choice model in hand. I introduce the generalized Roy model and the concept of the marginal treatment effect which helps to understand the strong assumptions made in the method of matching.

2. A Prototypical Policy Evaluation Problem

To motivate my discussion in the rest of this paper, consider the following prototypical policy problem. Suppose a policy is proposed for adoption in a country. It has been tried in other countries and we know outcomes there. We also know outcomes in countries where it was not adopted. From the historical record, what can we conclude about the likely effectiveness of the policy in countries that have not implemented it?

To answer questions of this sort, economists build models of counterfactuals. Consider the following model. Let Y_0 be the outcome of a country (e.g. GDP) under a no-policy regime. Y_1 is the outcome if the policy is implemented. $Y_1 - Y_0$ is the “treatment effect” of the policy. It may vary among countries. We observe characteristics X of various countries (e.g. level of democracy, level of population literacy, etc.). It is convenient to decompose Y_1 into its mean given X , $\mu_1(X)$ and deviation from mean U_1 . One can make a similar decomposition for Y_0 :

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0. \end{aligned} \tag{7}$$

Additive separability is not needed, but it is convenient to assume it, and I initially adopt it to simplify the exposition and establish a parallel regression notation that serves to link the statistical literature on treatment effects with the economic literature.

It may happen that controlling for the X , $Y_1 - Y_0$ is the same for all countries. This is the case of homogeneous treatment effects given X . More likely, countries vary in their responses to the policy even after controlling for X .

Figure 1 plots the distribution of $Y_1 - Y_0$ for a benchmark X . It also displays the various conventional treatment parameters. I use a special form of the generalized Roy model with constant cost C of adopting the policy. This is called the

“extended Roy model”. I use this model because it is simple and intuitive. (The precise parameterization of the extended Roy model used to generate the figure and the treatment effects is given at the base of Figure 1.) The special case of homogeneity in $Y_1 - Y_0$ arises when the distribution collapses to its mean. It would be ideal if one could estimate the distribution of $Y_1 - Y_0$ given X , and there is research that does this.

More often, economists focus on some mean of the distribution displayed in Figure 1 and use a regression framework to interpret the data. To turn (7) into a regression model, it is conventional to use the switching regression framework.²⁴ Define $D = 1$ if a country adopts a policy; $D = 0$ if it does not. The observed outcome Y is the switching regression model (1). Substituting (7) into this expression, and keeping X implicit, one obtains

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned} \quad (8)$$

Using conventional regression notation,

$$Y = \alpha + \beta D + \varepsilon, \quad (9)$$

where $\alpha = \mu_0$, $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$ and $\varepsilon = U_0$. I will also use the notation that $v = U_1 - U_0$, letting $\bar{\beta} = \mu_1 - \mu_0$ and $\beta = \bar{\beta} + v$. Throughout this paper I use treatment effect and regression notation interchangeably. The coefficient on D is the treatment effect. The case where β is the same for every country is the case conventionally assumed. More elaborate versions assume that β depends on X ($\beta(X)$) and estimates interactions of D with X . The case where β varies even after accounting for X is called the “random coefficient” or “heterogeneous treatment effect” case. The case where $v = U_1 - U_0$ depends on D is the case of essential heterogeneity or “correlated random coefficients” analyzed by Heckman, Urzúa, and Vytlačil (2006). This case arises when treatment choices depend at least in part on the idiosyncratic return to treatment. A great deal of attention has been focused on this case in recent decades, and I develop the implications of this model in this paper.

2.1 An Index Model of Choice and Treatment Effects: Definitions and Unifying Principles

I now present the model of treatment effects developed in Heckman and Vytlačil (1999, 2001, 2005, 2007a,b) and Heckman, Urzúa, and Vytlačil (2006), which relaxes the normality, separability and exogeneity assumptions invoked in the traditional economic selection models. It is rich enough to generate all of the treatment

²⁴Statisticians sometimes attribute this representation to Rubin (1974, 1978), but it is due to Quandt (1958, 1972). It is implicit in the Roy (1951) model. See the discussion of this basic model of counterfactuals in Heckman and Vytlačil (2007a). See also Heckman (2008, 2010).

effects displayed in Figure 1 as well as many other policy parameters. It does not require separability. It is a nonparametric generalized Roy model with testable restrictions that can be used to unify the treatment effect literature, identify different treatment effects, link the literature on treatment effects to the literature in structural econometrics and interpret the implicit economic assumptions underlying instrumental variables, regression discontinuity design methods, control functions and matching methods.

Y is the measured outcome variable. It is produced from the switching regression model (1). Outcomes are general nonlinear, nonseparable functions of observables and unobservables:

$$Y_1 = g_1(X, U_1) \quad (10)$$

$$Y_0 = g_0(X, U_0). \quad (11)$$

Examples of models that can be written in this form include conventional latent variable models for discrete choice that are generated by a latent variable crossing a threshold: $Y_i = \mathbf{1}(Y_i^* \geq 0)$, where $Y_i^* = \mu_i(X) + U_i$, $i = 0, 1$. Notice that in the general case, $g_i(X, U_i) - E(Y_i | X) \neq U_i$, $i = 0, 1$.

The individual treatment effect associated with moving an otherwise identical person from “0” to “1” is $Y_1 - Y_0 = \Delta$ and is defined as the causal effect on Y of a *ceteris paribus* move from “0” to “1”. To link this framework to the literature on economic choice models, I characterize the decision rule for program participation by an index model:

$$D^* = \mu_D(Z) - V; \quad D = 1 \quad \text{if} \quad D^* \geq 0; \quad D = 0 \quad \text{otherwise}, \quad (12)$$

where, from the point of view of the econometrician, (Z, X) is observed and (U_1, U_0, V) is unobserved. The random variable V may be a function of (U_1, U_0) . For example, in the original Roy Model, μ_1 and μ_0 are additively separable in U_1 and U_0 respectively, and $V = -[U_1 - U_0]$. In the original formulations of the generalized Roy model, outcome equations are separable and $V = -[U_1 - U_0 - U_C]$, where U_C arises from the cost function. Without loss of generality, I henceforth define Z so that it includes all of the elements of X as well as any additional variables unique to the choice equation.

I invoke the following assumptions that are weaker than those used in the conventional literature on structural econometrics or the recent literature on semi-parametric selection models and, at the same time, can be used both to define and to identify different treatment parameters.²⁵ The assumptions are:

²⁵A much weaker set of conditions is required to define the parameters than is required to identify them. See Heckman and Vytlačil (2007b, Appendix B).

(A-1) (U_0, U_1, V) are independent of Z conditional on X (**Independence**);

(A-2) $\mu_D(Z)$ is a nondegenerate random variable conditional on X (**Rank Condition**);

(A-3) The distribution of V is continuous;²⁶

(A-4) The values of $E|Y_1|$ and $E|Y_0|$ are finite (**Finite Means**);

(A-5) $0 < \Pr(D = 1 | X) < 1$.

(A-1) assumes that V is independent of Z given X and is used below to generate counterfactuals. For the definition of treatment effects one does not need either (A-1) or (A-2). The definitions of treatment effects and their unification do not require any elements of Z that are not elements of X or independence assumptions. However, an analysis of instrumental variables requires that Z contains at least one element not in X . Assumptions (A-1) or (A-2) justify application of instrumental variables methods and nonparametric selection or control function methods. Some parameters in the recent IV literature are defined by an instrument, therefore I make assumptions about instruments up front, noting where they are not needed. Assumption (A-4) is needed to satisfy standard integration conditions. It guarantees that the mean treatment parameters are well defined. Assumption (A-5) is the assumption in the population of both a treatment and a control group for each X . Observe that there are no exogeneity requirements for X . This is in contrast with the assumptions commonly made in the conventional structural literature and the semiparametric selection literature (see, e.g. Powell, 1994).

A counterfactual “no feedback” condition facilitates interpretability so that conditioning on X does not mask the effects of D . Letting X_d denote a value of X if D is set to d , a sufficient condition that rules out feedback from D to X is:

(A-6) Let X_0 denote the counterfactual value of X that would be observed if D is set to 0. X_1 is defined analogously. Assume $X_d = X$ for $d = 0, 1$. (The X_D are invariant to counterfactual manipulations.)

Condition (A-6) is not strictly required to formulate an evaluation model, but it enables an analyst who conditions on X to capture the “total” or “full effect” of D on Y (see Pearl, 2000). This assumption imposes the requirement that X is an external variable determined outside the model and is not affected by counterfactual manipulations of D . However, the assumption allows for X to be freely correlated with U_1 , U_0 and V , so it can be endogenous.

²⁶Absolutely continuous with respect to Lebesgue measure.

In this notation, $P(Z)$ is the probability of receiving treatment given Z , or the “propensity score” $P(Z) \equiv \Pr(D = 1 \mid Z) = F_{V|X}(\mu_D(Z))$, where $F_{V|X}(\cdot)$ denotes the distribution of V conditional on X .²⁷ I denote $P(Z)$ by P , suppressing the Z argument. I also work with U_D , a uniform random variable ($U_D \sim \text{Unif}[0, 1]$) defined by $U_D = F_{V|X}(V)$.²⁸ The separability between V and $\mu_D(Z)$ or $D(Z)$ and U_D is conventional. It plays a crucial role in justifying instrumental variable estimators in the general models analyzed in this paper.

Vytlacil (2002) establishes that assumptions (A-1)–(A-5) for selection model (10)–(12) are *equivalent* to the assumptions used to generate the LATE model of Imbens and Angrist (1994). Thus the nonparametric selection model for treatment effects developed by Heckman and Vytlacil is implied by the assumptions of the Imbens and Angrist instrumental variable model for treatment effects. Their approach is more general and links the IV literature to the literature on economic choice models. The latent variable model is a version of the standard sample selection bias model. This weaves together two strands of the literature often thought to be distinct (see e.g. Angrist and Krueger, 1999). Heckman, Urzúa, and Vytlacil (2006) develop this parallelism in detail.²⁹

2.2 Definitions of Treatment Effects in the Two Outcome Model

The difficulty of observing the same individual in both treated and untreated states leads to the use of various population level treatment effects widely used in the biostatistics literature and often applied in economics.³⁰ The most commonly

²⁷Throughout this paper, I will refer to the cumulative distribution function of a random vector A by $F_A(\cdot)$ and to the cumulative distribution function of a random vector A conditional on random vector B by $F_{A|B}(\cdot)$. I will write the cumulative distribution function of A conditional on $B = b$ by $F_{A|B}(\cdot \mid b)$.

²⁸This representation is valid whether or not (A-1) is true. However, (A-1) imposes restrictions on counterfactual choices. For example, if a change in government policy changes the distribution of Z by an external manipulation, under (A-1) the model can be used to generate the choice probability from $P(z)$ evaluated at the new arguments, i.e., the model is invariant with respect to the distribution Z .

²⁹The model of equations (10)–(12) and assumptions (A-1)–(A-5) impose two testable restrictions on the distribution of (Y, D, Z, X) . First, it imposes an index sufficiency restriction: for any set $\mathcal{A}$ and for $j = 0, 1$,

$$\Pr(Y_j \in \mathcal{A} \mid X, Z, D = j) = \Pr(Y_j \in \mathcal{A} \mid X, P(Z), D = j).$$

Z (given X) enters the model only through the propensity score $P(Z)$ (the sets of $\mathcal{A}$ are assumed to be measurable). This restriction has empirical content when Z contains two or more variables not in X . Second, the model also imposes monotonicity in p for $E(YD \mid X = x, P = p)$ and $E(Y(1 - D) \mid X = x, P = p)$. Heckman and Vytlacil (2005, Appendix A) develop this condition further, and show that it is testable. Even though this model of treatment effects is not the most general possible model, it has testable implications and hence empirical content. It unites various literatures and produces a nonparametric version of the selection model, and links the treatment literature to economic choice theory.

³⁰Heckman, LaLonde, and Smith (1999) discuss panel data cases where it is possible to observe both Y_0 and Y_1 for the same person.

invoked treatment effect is the Average Treatment Effect (ATE): $\Delta^{\text{ATE}}(x) \equiv E(\Delta \mid X = x)$ where $\Delta = Y_1 - Y_0$. This is the effect of assigning treatment randomly to everyone of type X assuming full compliance, and ignoring general equilibrium effects.³¹ The average impact of treatment on persons who actually take the treatment is Treatment on the Treated (TT): $\Delta^{\text{TT}}(x) \equiv E(\Delta \mid X = x, D = 1)$. This parameter can also be defined conditional on $P(Z)$: $\Delta^{\text{TT}}(x, p) \equiv E(\Delta \mid X = x, P(Z) = p, D = 1)$.³²

The mean effect of treatment on those for whom $X = x$ and $U_D = u_D$, the Marginal Treatment Effect (MTE), plays a fundamental role in the analysis of the next section:

$$\Delta^{\text{MTE}}(x, u_D) \equiv E(\Delta \mid X = x, U_D = u_D). \quad (13)$$

This parameter is defined independently of any instrument. I separate the definition of parameters from their identification. The MTE is the expected effect of treatment conditional on observed characteristics X and conditional on U_D , the unobservables from the first stage decision rule. For u_D evaluation points close to zero, $\Delta^{\text{MTE}}(x, u_D)$ is the expected effect of treatment on individuals with the value of unobservables that make them most likely to participate in treatment and who would participate even if the mean scale utility $\mu_D(Z)$ is small. If U_D is large, $\mu_D(Z)$ would have to be large to induce people to participate.

One can also interpret $E(\Delta \mid X = x, U_D = u_D)$ as the mean gain in terms of $Y_1 - Y_0$ for persons with observed characteristics X who would be indifferent between treatment or not if they were randomly assigned a value of Z , say z , such that $\mu_D(z) = u_D$. When Y_1 and Y_0 are value outcomes, MTE is a mean willingness-to-pay measure. MTE is a choice-theoretic building block that unites the treatment effect, selection, matching and control function literatures.

A third interpretation is that MTE conditions on X and the residual defined by subtracting the expectation of D^* from D^* : $\tilde{U}_D = D^* - E(D^* \mid Z, X)$. This is a “replacement function” interpretation in the sense of Heckman and Robb (1985a) and Matzkin (2007), or “control function” interpretation in the sense of Blundell and Powell (2003).³³ The additive separability of equation (12) in terms of observables and unobservables plays a crucial role in the justification of instrumental variable methods.

The LATE parameter of Imbens and Angrist (1994) is a version of MTE. I define LATE independently of any instrument after first presenting the Imbens–Angrist definition. Define $D(z)$ as a counterfactual choice variable, with

³¹See, e.g., Imbens (2004).

³²These two definitions of treatment on the treated are related by integrating out the conditioning p variable: $\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{TT}}(x, p) dF_{P(Z) \mid X, D}(p \mid x, 1)$ where $F_{P(Z) \mid X, D}(\cdot \mid x, 1)$ is the distribution of $P(Z)$ given $X = x$ and $D = 1$.

³³These three interpretations are equivalent under separability in D^* , i.e., when (12) characterizes the choice equation, but lead to three different definitions of MTE when a more general nonseparable model is developed. See Heckman and Vytlacil (2007b).

$D(z) = 1$ if D would have been chosen if Z had been set to z , and $D(z) = 0$ otherwise. Let $\mathcal{Z}(x)$ denote the support of the distribution of Z conditional on $X = x$. For any $(z, z') \in \mathcal{Z}(x) \times \mathcal{Z}(x)$ such that $P(z) > P(z')$, LATE is $E(\Delta \mid X = x, D(z) = 1, D(z') = 0) = E(Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0)$, the mean gain to persons who would be induced to switch from $D = 0$ to $D = 1$ if Z were manipulated externally from z' to z . In an example of the returns to education, z' could be the base level of tuition and z a reduced tuition level. Using the latent index model, Heckman and Vytlacil (1999, 2005) show that LATE can be written as

$$\begin{aligned} E(Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0) &= E(Y_1 - Y_0 \mid X = x, u'_D < U_D < u_D) \\ &= \Delta^{\text{LATE}}(x, u_D, u'_D) \end{aligned}$$

for $u_D = \Pr(D(z) = 1) = P(z)$, $u'_D = \Pr(D(z') = 1) = P(z')$, where assumption (A-1) implies that $\Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z)$ and $\Pr(D(z') = 1) = \Pr(D = 1 \mid Z = z')$.

Imbens and Angrist define the LATE parameter as the probability limit of an IV estimator. Their analysis conflates issues of definition of parameters with issues of identification. The representation of LATE given here allows analysts to separate these two conceptually distinct matters and to define the LATE parameter more generally. One can in principle evaluate the right hand side of the preceding equation at any u_D, u'_D points in the unit interval and not only at points in the support of the distribution of the propensity score $P(Z)$ conditional on $X = x$ where it is identified. From assumptions (A-1), (A-3), and (A-4), $\Delta^{\text{LATE}}(x, u_D, u'_D)$ is continuous in u_D and u'_D and $\lim_{u'_D \rightarrow u_D} \Delta^{\text{LATE}}(x, u_D, u'_D) = \Delta^{\text{MTE}}(x, u_D)$.³⁴

Heckman and Vytlacil (1999) use assumptions (A-1)–(A-5) and the latent index structure to develop the relationship between MTE and the various treatment effect parameters shown in the first three lines of Table 2A. They present the formal derivation of the parameters and associated weights and graphically illustrates the relationship between ATE and TT. All treatment parameters may be expressed as weighted averages of the MTE:

$$\text{Treatment Parameter}(j, x) = \int \Delta^{\text{MTE}}(x, u_D) \omega_j(x, u_D) du_D,$$

where $\omega_j(x, u_D)$ is the weighting function for the MTE, and the integral is defined over the full support of u_D . Except for the OLS weights, the weights in the table all integrate to one, although in some cases the weights for IV may be negative over a range of values (Heckman, Urzúa, and Vytlacil, 2006).

³⁴This follows from Lebesgue's theorem for the derivative of an integral and holds almost everywhere with respect to Lebesgue measure. The ideas of the marginal treatment effect and the limit form of LATE were first introduced in the context of a parametric normal generalized Roy model by Björklund and Moffitt (1987), and were analyzed more generally in Heckman (1997). Angrist, Graddy, and Imbens (2000) also define and develop a limit form of LATE.

In Table 2A, $\Delta^{\text{TT}}(x)$ is shown as a weighted average of Δ^{MTE} :

$$\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) \omega_{\text{TT}}(x, u_D) du_D,$$

where

$$\omega_{\text{TT}}(x, u_D) = \frac{1 - F_{P|X}(u_D | x)}{\int_0^1 (1 - F_{P|X}(t | x)) dt} = \frac{S_{P|X}(u_D | x)}{E(P(Z) | X = x)}, \quad (14)$$

and $S_{P|X}(u_D | x)$ is $\Pr(P(Z) > u_D | X = x)$ and $\omega_{\text{TT}}(x, u_D)$ is a weighted distribution. The parameter $\Delta^{\text{TT}}(x)$ oversamples $\Delta^{\text{MTE}}(x, u_D)$ for those individuals with low values of u_D that make them more likely to participate in the program being evaluated. Treatment on the untreated (TUT) is defined symmetrically with TT and oversamples those with u_D least likely to participate. The various weights are displayed in Table 2B. A central theme of the analysis of Heckman and Vytlacil (1999, 2005, 2007b) is that under their assumptions all estimators and estimands can be written as weighted averages of MTE. This allows them to unify the treatment effect literature using a common functional MTE (u_D).

Observe that if $E(Y_1 - Y_0 | X = x, U_D = u_D) = E(Y_1 - Y_0 | X = x)$, so $\Delta = Y_1 - Y_0$ is mean independent of U_D given $X = x$, then $\Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}} = \Delta^{\text{LATE}}$. Therefore in cases where there is no heterogeneity in terms of unobservables in MTE (Δ constant conditional on $X = x$) or agents do not act on it so that U_D drops out of the conditioning set, marginal treatment effects are average treatment effects, so that all of the evaluation parameters are the same. Otherwise, they are different. Only in the case where the marginal treatment effect is the average treatment effect will the “effect” of treatment be uniquely defined.

Figure 2A plots weights for a parametric normal generalized Roy model generated from the parameters shown at the base of Figure 2B. The model allows for costs to vary in the population and is more general than the extended Roy model used to construct Figure 1. The weights for IV depicted in Figure 2B are discussed in Heckman, Urzúa, and Vytlacil (2006) and the weights for OLS are discussed in the next section. A high u_D is associated with higher cost, relative to return, and less likelihood of choosing $D = 1$. The decline of MTE in terms of higher values of u_D means that people with higher u_D have lower gross returns. TT overweights low values of u_D (i.e., it oversamples U_D that make it likely to have $D = 1$). ATE samples U_D uniformly. Treatment on the Untreated ($E(Y_1 - Y_0 | X = x, D = 0)$), or TUT, oversamples the values of U_D which make it unlikely to have $D = 1$.

Table 3 shows the treatment parameters produced from the different weighting schemes for the model used to generate the weights in figures 2A and 2B. Given the decline of the MTE in u_D , it is not surprising that $\text{TT} > \text{ATE} > \text{TUT}$. This is the generalized Roy version of the principle of diminishing returns. Those most likely to self-select into the program benefit the most from it. The difference between

TT and ATE is a sorting gain: $E(Y_1 - Y_0 \mid X, D = 1) - E(Y_1 - Y_0 \mid X)$, the average gain experienced by people who sort into treatment compared to what the average person would experience. Purposive selection on the basis of gains should lead to positive sorting gains of the kind found in the table. If there is negative sorting on the gains, then $TUT \geq ATE \geq TT$.

2.3 The Weights for a Generalized Roy Model

For the case of continuous Z , I plot the weights associated with the MTE for IV. This analysis draws on Heckman, Urzúa, and Vytlacil (2006), who derive the weights. Figure 3 plots $E(Y \mid P(Z))$ and MTE for the extended Roy models generated by the parameters displayed at the base of the figure. In cases where $\beta \perp\!\!\!\perp D$, $\Delta^{\text{MTE}}(u_D)$ is constant in u_D . This is trivial when β is a constant (case IA). When β is random but selection into D does not depend on β , MTE is still flat (case IB). The more interesting case termed “essential heterogeneity” by Heckman and Vytlacil (1999, 2005, 2007a,b) has $\beta \not\perp\!\!\!\perp D$. The left hand side (Figure 3A) depicts $E(Y \mid P(Z))$ in the two cases. The first case makes $E(Y \mid P(Z))$ linear in $P(Z)$. The second case is nonlinear in $P(Z)$. This arises when $\beta \not\perp\!\!\!\perp D$. The derivative of $E(Y \mid P(Z))$ is presented in the right panel (Figure 3B). It is a constant for the first case (flat MTE) but declining in $U_D = P(Z)$ for the case with selection on the gain. A simple test for linearity in $P(Z)$ in the outcome equation reveals whether or not the analyst is in cases IA and IB ($\beta \perp\!\!\!\perp D$) or case II ($\beta \not\perp\!\!\!\perp D$).³⁵ These cases are the extended Roy counterparts to $E(Y \mid P(Z) = p)$ and MTE shown for the generalized Roy model in figures 4A and 4B.

MTE gives the mean marginal return for persons who have utility $P(Z) = u_D$. Thus, $P(Z) = u_D$ is the margin of indifference. Those with low u_D values have high returns. Those with high u_D values have low returns. Figure 3 highlights that in the general case MTE (and LATE) identify average returns for persons at the margin of indifference at different levels of the mean utility function ($P(Z)$).

Figure 5 plots MTE and LATE for different intervals of u_D using the model generating Figure 3. LATE is the chord of $E(Y \mid P(Z))$ evaluated at different points. The relationship between LATE and MTE is depicted in the right panel (B) of Figure 5. LATE is the integral under the MTE curve divided by the difference between the upper and lower limits.

Treatment parameters associated with the second case are plotted in Figure 6. The MTE is the same as that presented in Figure 3. ATE has the same value for all p . The effect of treatment on the treated for $P(Z) = p$, $\Delta^{\text{TT}}(p) = E(Y_1 - Y_0 \mid D = 1, P(Z) = p)$ declines in p (equivalently it declines in u_D). Treatment on the untreated given p , $TUT(p) = \Delta^{\text{TUT}}(p) = E(Y_1 - Y_0 \mid D = 0, P(Z) = p)$ also declines in p .

³⁵Recall that I keep the conditioning on X implicit.

$$LATE(p, p') = \frac{\Delta^{TT}(p')p' - \Delta^{TT}(p)p}{p' - p}, \quad p' \neq p$$

$$MTE = \frac{\partial[\Delta^{TT}(p)p]}{\partial p}.$$

One can generate all of the treatment parameters from $\Delta^{TT}(p)$.

Matching on $P = p$ (which is equivalent to nonparametric regression given $P = p$) produces a biased estimator of $TT(p)$. Matching assumes a flat MTE (average return equals marginal return) as I develop below. Therefore it is systematically biased for $\Delta^{TT}(p)$ in a model with essential heterogeneity, where $\beta \not\perp D$. Making observables alike makes the unobservables dissimilar. Holding p constant across treatment and control groups understates $TT(p)$ for low values of p and overstates it for high values of p . I develop this point further in the next section, where I discuss the method of matching.

3. Matching

The method of matching allows for selection of treatment based on potential outcomes

$$(Y_0, Y_1) \not\perp D,$$

so $\Pr(D = 1 \mid Y_0, Y_1)$ depends on Y_0, Y_1 . It assumes access to variables Q such that conditioning on Q removes the dependence:

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid Q. \quad (\text{Q-1})$$

I adopt a more general notation than is used in Section 1 allowing Q to include variables not in X . Thus,

$$\Pr(D = 1 \mid Q, Y_0, Y_1) = \Pr(D = 1 \mid Q).$$

Comparisons between treated and untreated can be made at all points in the support of Q such that

$$0 < \Pr(D = 1 \mid Q) < 1. \quad (\text{Q-2})$$

The method does not explicitly model choices of treatment or the subjective evaluations of participants, nor is there any distinction between the variables in the outcome equations (X) and the variables in the choice equations (Z) that is central to the IV method and the method of control functions. The conditioning variables do not have to be exogenous.

From condition (Q-1) one recovers the distributions of Y_0 and Y_1 given Q — $\Pr(Y_0 \leq y_0 \mid Q = q) = F_0(y_0 \mid Q = q)$ and $\Pr(Y_1 \leq y_1 \mid Q = q) = F_1$

$(y_1 | Q = q)$ — but not the joint distribution $F_{0,1}(y_0, y_1 | Q = q)$, because the analyst does not observe the same persons in the treated and untreated states. This is a standard evaluation problem common to all econometric estimators. Methods for determining which variables belong in Q rely on untested exogeneity assumptions which I discuss in this section.

OLS is a special case of matching that focuses on the identification of conditional means. In OLS linear functional forms are maintained as exact representations or valid approximations. Considering a common coefficient model, OLS writes

$$Y = \pi Q + D\beta + U, \quad (\text{Q-3})$$

where β is the treatment effect and

$$E(U | Q, D) = 0. \quad (\text{Q-4})$$

The assumption is made that the variance-covariance matrix of (Q, D) is of full rank:

$$\text{Var}(Q, D) \text{ full rank.} \quad (\text{Q-5})$$

Under these conditions, one can identify β even though D and U are dependent: $D \not\perp U$. Controlling for the observable Q eliminates any spurious mean dependence between D and U : $E(U | D) \neq 0$ but $E(U | D, Q) = 0$. (Q-3) is the linear regression counterpart to (Q-1). (Q-5) is the linear regression counterpart to (Q-4). Failure of (Q-5) would mean that using a nonparametric estimator one can perfectly predict D given Q over certain regions, and that in those regions $\Pr(D = 1 | Q = q) = 1$ or 0.³⁶ In a linear regression setting, matching improves on OLS by focusing attention solely on β and not π . It weakens Q-4 to

$$(\text{Q-4})' E(U | Q, D) = E(U | Q).$$

It still invokes (Q-5).

Matching can be implemented as a nonparametric method. When this is done, the procedure does not require specification of the functional form of the outcome equations. It enforces the requirement that (Q-2) be satisfied by estimating functions pointwise in the support of Q . Assume that $Q = (X, Z)$ and that X and Z are the same except where otherwise noted. Thus I invoke assumptions (M-1) and (M-2) presented in Section 1, even though in principle one can use a more general conditioning set.

³⁶This condition might be met only at certain values of $Q = q$. For certain parameterizations (e.g. the linear probability model), one may obtain predicted probabilities outside the unit interval.

Assumptions (M-1) and (M-2) or (Q-1) and (Q-2) rule out the possibility that after conditioning on X (or Q), agents possess more information about their choices than econometricians, and that the unobserved information helps to predict the potential outcomes. Put another way, the method allows for potential outcomes to affect choices but only through the observed variables, Q , that predict outcomes. This is the reason why Heckman and Robb (1985a, 1986b) call the method selection on observables.

This section establishes the following points.

- (1) Matching assumptions (M-1) and (M-2) generically imply a flat MTE in u_D , i.e. they assume that $E(Y_1 - Y_0 \mid X = x, U_D = u_D)$ does not depend on u_D . Thus the unobservables central to the Roy model and its extensions and the unobservables central to the modern IV literature are assumed to be absent once the analyst conditions on X . (M-1) implies that all mean treatment parameters are the same.
- (2) Even if one weakens (M-1) and (M-2) to mean independence instead of full independence, generically the MTE is flat in u_D under the assumptions of the nonparametric generalized Roy model developed in subsection 2.1, so again all mean treatment parameters are the same.
- (3) I show that IV and matching make distinct identifying assumptions even though they both invoke conditional independence assumptions.
- (4) I compare matching with IV and control function (sample selection) methods. Matching assumes that conditioning on observables eliminates the dependence between (Y_0, Y_1) and D . The control function principle models the dependence.
- (5) I present some examples that demonstrate that if the assumptions of the method of matching are violated, the method can produce substantially biased estimators of the parameters of interest.
- (6) I show that standard methods for selecting the conditioning variables used in matching assume exogeneity. Violations of the exogeneity assumption in selecting conditioning variables can produce biased estimators.

Nonparametric versions of matching embodying (M-2) avoid the problem of making inferences outside the support of the data. This problem is implicit in any application of least squares. Figure 7 shows the support problem that can arise in linear least squares when the linearity of the regression is used to extrapolate estimates determined in one empirical support to new supports. Careful attention to support problems is a virtue of any nonparametric method, including, but not unique to, nonparametric matching. Heckman, Ichimura, Smith, and Todd

(1998) show that the bias from neglecting the problem of limited support can be substantial. See also the discussion in Heckman, LaLonde, and Smith (1999).

I now show that matching implies that conditional on X , the marginal return is assumed to be the same as the average return (marginal = average). This is a strong behavioral assumption implicit in statistical conditional independence assumption (M-1). It says that the marginal participant has the same return as the average participant.

3.1 Matching Assumption (M-1) Implies a Flat MTE

An immediate consequence of (M-1) is that the MTE does not depend on U_D . This is so because $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$ implies that $(Y_0, Y_1) \perp\!\!\!\perp U_D \mid X$ and hence that

$$\Delta^{\text{MTE}}(x, u_D) = E(Y_1 - Y_0 \mid X = x, U_D = u_D) = E(Y_1 - Y_0 \mid X = x). \quad (15)$$

This, in turn, implies that Δ^{MTE} conditional on X is flat in u_D . Under the stated assumptions for the generalized Roy model, it assumes that $E(Y \mid X = x, P(Z) = p)$ is linear in p . Thus the method of matching assumes that mean marginal returns and average returns are the same and all mean treatment effects are the same given X . However, one can still distinguish marginal from average effects of the observables (X) using matching. See Carneiro (2002).

It is sometimes said that the matching assumptions are “for free” (See, e.g., Gill and Robins, 2001) because one can always replace unobserved $F_1(Y_1 \mid X = x, D = 0)$ with observed $F_1(Y_1 \mid X = x, D = 1)$ and unobserved $F_0(Y_0 \mid X = x, D = 1)$ with observed $F_0(Y_0 \mid X = x, D = 0)$. Such replacements do not contradict any observed data.

While the claim is true, it ignores the properties of the counterfactual states generated under the matching assumptions. The assumed absence of selection on unobservables is not a “for free” assumption, and produces fundamentally different counterfactual states for the same model under matching and selection assumptions. To explore these issues in depth, consider a nonparametric regression model.

Without assumption (M-1), a nonparametric regression of Y on D conditional on X identifies a nonparametric mean difference:

$$\begin{aligned} \Delta^{\text{OLS}}(X) &= E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 0) \\ &= E(Y_1 - Y_0 \mid X, D = 1) + \{E(Y_0 \mid X, D = 1) \\ &\quad - E(Y_0 \mid X, D = 0)\}. \end{aligned} \quad (16)$$

The term in braces in the second expression arises from selection on pre-treatment levels of the outcome. OLS identifies the parameter treatment on the

treated (the first term in the second line of (16)) plus a bias term in braces corresponding to selection on the levels.

The OLS estimator can be represented as a weighted average of Δ^{MTE} . The weight is given in Table 2B where U_1 and U_0 for the OLS model are defined as deviations from conditional expectations, $U_1 = Y_1 - E(Y_1 | X)$, $U_0 = Y_0 - E(Y_0 | X)$. Unlike the weights for Δ^{TT} and Δ^{ATE} , the OLS weights do not necessarily integrate to one and they are not necessarily nonnegative. Under its assumptions, application of IV eliminates the contribution of the second term of equation (16). The weights for the first term are the same as the weights for Δ^{TT} and hence they integrate to one.

The OLS weights for the generalized Roy model example are plotted in Figure 2B. The negative component of the OLS weight leads to a smaller OLS treatment estimate compared to the other treatment effects in Table 3. This table shows the estimated OLS treatment effect for a generalized Roy example. The large negative selection bias in this example is consistent with comparative advantage as emphasized by Roy (1951) and detected empirically by Willis and Rosen (1979) and Cunha, Heckman, and Navarro (2005). People who are good in sector 1 (i.e. receive treatment) may be very poor in sector 0 (those who receive no treatment). Hence the bias in OLS for the parameter treatment on the treated may be negative ($E(Y_0 | X, D = 1) - E(Y_0 | X, D = 0) < 0$). The differences among the conventional treatment effects and the OLS estimand are illustrated in Figure 8A and Table 3 for the generalized Roy model example. While IV eliminates the term in braces in (16), it reweights the MTE differently from what might be desired for many policy analyses.

If there is no selection on unobserved variables conditional on covariates, $U_D \perp \perp (Y_0, Y_1) | X$, then $E(U_1 | X, U_D) = E(U_1 | X) = 0$ and $E(U_0 | X, U_D) = E(U_0 | X) = 0$ so that the OLS weights are unity and OLS identifies both ATE and the parameter treatment on the treated (TT), which are the same under this assumption. This condition is an implication of matching condition (M-1). Given the assumed conditional independence in terms of X , we can identify ATE and TT without use of any instrument Z satisfying assumptions (M-1)–(M-1). The conditional independence assumption invoked in the method of matching has come into widespread use for much the same reason that OLS has come into widespread use. It is easy to implement with modern software and makes little demands of the data because it assumes the existence of X variables that satisfy the conditional independence assumptions. The crucial conditional independence assumption is not testable. As I note below, additional assumptions on the X are required to test the validity of the matching assumptions.

If the sole interest is to identify treatment on the treated, Δ^{TT} , it is apparent from representation (16) that one can weaken (M-1) to

$$(M-1)' \quad Y_0 \perp\!\!\!\perp D | X.$$

(The equivalent assumption for (Q-1) is

$$(Q-1)' \quad Y_0 \perp\!\!\!\perp D \mid Q.)$$

This is possible because we know $E(Y_1 \mid X, D = 1)$ from data on outcomes of the treated and only need to construct $E(Y_0 \mid X, D = 1)$. In this case MTE is not restricted to be flat in u_D and all treatment parameters are not the same. A straightforward implication of (M-1)' in the Roy model, where selection is made solely on the gain, is that persons must sort into treatment status positively in terms of levels of Y_1 . I now consider more generally the implications of assuming mean independence of the errors rather than full independence.

3.2 Matching and MTE Using Mean Independence Conditions: Matching Versus IV

To identify all mean treatment parameters, one can weaken the assumption (M-1) to the condition that Y_1 and Y_0 are mean independent of D conditional on X . However, (Y_0, Y_1) will be mean independent of D conditional on X without U_D being independent of (Y_0, Y_1) conditional on X only if fortuitous balancing occurs, with regions of positive dependence of (Y_0, Y_1) on U_D and regions of negative dependence of (Y_0, Y_1) on U_D just exactly offsetting each other.

IV and matching invoke very different assumptions. Matching requires no exclusion restrictions whereas IV is based on the existence of exclusion restrictions. Superficially, we can bridge these literatures by invoking matching with an exclusion condition: $(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X$ but $(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z$. This looks like an IV condition, but it is not.

Heckman and Vytlacil (2007b, Appendix L) explore the relationship between matching with exclusion and IV and demonstrate a fundamental contradiction between the two identifying conditions. For an additively separable representation of the outcome equations $U_1 = Y_1 - E(Y_1 \mid X)$ and $U_0 = Y_0 - E(Y_0 \mid X)$, if (U_0, U_1) is mean independent of D conditional on (X, Z) , as required by IV, but (U_0, U_1) is not mean independent of D conditional on X alone, then U_0 is dependent on Z conditional on X , contrary to all assumptions used to justify instrumental variables. I next consider how to implement matching.

3.3 Implementing the Method of Matching

I draw on Heckman, Ichimura, Smith, and Todd (1998) and Heckman, LaLonde, and Smith (1999) to describe the mechanics of matching. Todd (2007, 2008) presents a comprehensive treatment of the main issues and a guide to software.

To operationalize the method of matching, I assume two samples: “ t ” for treatment and “ c ” for comparison group. Treatment group members have $D = 1$ and control group members have $D = 0$. Unless otherwise noted, I assume that observations are statistically independent within and across groups. Simple

matching methods are based on the following idea. For each person i in the treatment group, find some group of “comparable” persons. The same individual may be in both treated and control groups if that person is treated at one time and untreated at another. Denote outcomes for person i in the treatment group by Y_i^t and match these outcomes to the outcomes of a subsample of persons in the comparison group to estimate a treatment effect. In principle, one can use a different subsample as a comparison group for each person.

In practice, one can construct matches on the basis of a neighborhood $\xi(X_i)$, where X_i is a vector of characteristics for person i . Neighbors to treated person i are persons in the companion sample whose characteristics are in neighborhood $\xi(X_i)$. Suppose that there are N_c persons in the comparison sample and N_t in the treatment sample. Thus the persons in the comparison sample who are neighbors to i , are persons j for whom $X_j \in \xi(X_i)$, i.e., the set of persons $A_i = \{j \mid X_j \in \xi(X_i)\}$. Let $W(i, j)$ be the weight placed on observation j in forming a comparison with observation i and further assume that the weights sum to one, $\sum_{j=1}^{N_c} W(i, j) = 1$, and that $0 \leq W(i, j) \leq 1$. Form a weighted comparison group mean for person i , given by

$$\bar{Y}_i^c = \sum_{j=1}^{N_c} W(i, j) Y_j^c. \quad (17)$$

The estimated treatment effect for person i is $Y_i - \bar{Y}_i^c$. This selects a set of comparison group members associated with i and the mean of their outcomes. Unlike IV or the control function approach, the method of matching identifies counterfactuals for each treated member.

Heckman, Ichimura, and Todd (1997) and Heckman, LaLonde, and Smith (1999) survey a variety of alternative matching schemes proposed in the literature. Todd (2007, 2008) provides a comprehensive survey. I briefly consider two widely-used methods. The nearest-neighbor matching estimator defines A_i such that only one j is selected so that it is closest to X_i in some metric:

$$A_i = \{j \mid \min_{j \in \{1, \dots, N_c\}} \|X_i - X_j\|\},$$

where “ $\| \ \|$ ” is a metric measuring distance in the X characteristics space. The Mahalanobis metric is one widely used metric for implementing the nearest neighbor matching estimator. This metric defines neighborhoods for i as

$$\| \ \| = (X_i - X_j)' \sum_c^{-1} (X_i - X_j),$$

where $\sum_c$ is the covariance matrix in the comparison sample. The weighting scheme for the nearest neighbor matching estimator is

$$W(i, j) = \begin{cases} 1 & \text{if } j \in A_i, \\ 0 & \text{otherwise.} \end{cases}$$

The nearest neighbor in the metric “ $\|\cdot\|$ ” is used in the match. A version of nearest-neighbor matching, called “caliper” matching (Cochran and Rubin, 1973), makes matches to person i only if

$$\|X_i - X_j\| < \varepsilon,$$

where ε is a pre-specified tolerance. Otherwise person i is bypassed and no match is made to him or her.

Kernel matching uses the entire comparison sample, so that $A_i = \{1, \dots, N_c\}$, and sets

$$W(i, j) = \frac{K(X_j - X_i)}{\sum_{j=1}^{N_c} K(X_j - X_i)},$$

where K is a kernel.³⁷ Kernels are typically a standard distribution function such as the normal cumulative distribution function. Kernel matching is a smooth method that reuses and weights the comparison group sample observations differently for each person i in the treatment group with a different X_i . Kernel matching can be defined pointwise at each sample point X_i or for broader intervals.

For example, the impact of treatment on the treated can be estimated by forming the mean difference across the i :

$$\hat{\Delta}^{TT} = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i^t - \bar{Y}_i^c) = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i^t - \sum_{j=1}^{N_c} W(i, j) Y_j^c). \quad (18)$$

One can define this mean for various subsets of the treatment sample defined in various ways. More efficient estimators weight the observations accounting for the variance (Heckman, Ichimura, and Todd, 1997, 1998, Heckman, 1998, Hirano, Imbens, and Ridder, 2003, Abadie and Imbens, 2006).³⁸

Matching assumes that conditioning on X eliminates selection bias. The method requires no functional form assumptions for outcome equations. If, however, a functional form assumption is maintained, as in the econometric procedure proposed by Barnow, Cain, and Goldberger (1980), it is possible to implement the

³⁷See, e.g., Härdle (1990) or Ichimura and Todd (2007) for a discussion of kernels and choices of bandwidths.

³⁸Regression-adjusted matching, proposed by Rubin (1979) and clarified in Heckman, Ichimura, and Todd (1997, 1998), uses regression-adjusted Y_i , denoted by $\tau(Y_i) = Y_i - X_i\beta$, in place of Y_i in the preceding calculations. (See the cited papers for the econometric details of the procedure).

matching assumption using standard regression analysis. Suppose, for example, that Y_0 is linearly related to observables X and an unobservable U_0 , so that

$$E(Y_0 | X, D = 0) = X\alpha + E(U_0 | X, D = 0),$$

and

$$E(U_0 | X, D = 0) = E(U_0 | X)$$

is linear in X ($E(U_0 | X) = X\psi$). Under these assumptions, controlling for X via linear regression allows one to identify $E(Y_0 | X, D = 1)$ from the data on nonparticipants. Setting $X = Q$, this approach justifies OLS equation (Q-3).³⁹ Such functional form assumptions are not strictly required to implement the method of matching. Moreover, in practice, users of the method of Barnow, Cain, and Goldberger (1980) do not impose the common support condition (M-2) for the distribution of X when generating estimates of the treatment effect. The distribution of X may be very different in the treatment group ($D = 1$) and comparison group ($D = 0$) samples, so that comparability is only achieved by imposing linearity in the parameters and extrapolating over different regions.

One advantage of the method of Barnow, Cain, and Goldberger (1980) is that it uses data parsimoniously. If the X are high dimensional, the number of observations in each cell when matching can get very small.

Another solution to this problem that reduces the dimension of the matching problem without imposing arbitrary linearity assumptions is based on the probability of participation or the “propensity score,” $P(X) = \Pr(D = 1 | X)$. Rosenbaum and Rubin (1983) demonstrate that under assumptions (M-1) and (M-2),

$$(Y_0, Y_1) \perp\!\!\!\perp D | P(X) \text{ for } X \in \chi_c, \quad (19)$$

for some set χ_c , where it is assumed that (M-2) holds in the set. Conditioning either on $P(X)$ or on X produces conditional independence.⁴⁰

Conditioning on $P(X)$ reduces the dimension of the matching problem down to matching on the scalar $P(X)$. The analysis of Rosenbaum and Rubin (1983) assumes that $P(X)$ is known rather than estimated. Heckman, Ichimura, and Todd (1998), Hahn (1998), and Hirano, Imbens, and Ridder (2003) present the asymptotic distribution theory for the kernel matching estimator in the cases in which $P(X)$ is known and in which it is estimated both parametrically and non-parametrically.

³⁹In equation (Q-3), this approach shows that π combines the estimate $U_0 - Q$ effect with the causal effect of Q on Y .

⁴⁰Their analysis is generalized to a multiple treatment setting in Lechner (2001) and Imbens (2003).

Conditioning on P identifies all treatment parameters but as has been shown, it imposes the assumption of a flat MTE. Marginal returns and average returns are the same. A consequence of (19) is that

$$\begin{aligned} E(Y_1|D=0, P(X)) &= E(Y_1|D=1, P(X)) = E(Y_1|P(X)), \\ E(Y_0|D=1, P(X)) &= E(Y_0|D=0, P(X)) = E(Y_0|P(X)). \end{aligned}$$

Support condition (M-2) has the unattractive feature that if the analyst has too much information about the decision of who takes treatment, so that $P(X) = 1$ or 0 , the method breaks down at such values of X because people cannot be compared at a common X . The method of matching assumes that, given X , some unspecified randomization in the economic environment allocates people to treatment. This produces assumption (Q-5) in the OLS example. The fact that the cases $P(X) = 1$ and $P(X) = 0$ must be eliminated suggests that methods for choosing X based on the fit of the model to data on D are potentially problematic, a point I discuss further below.

Offsetting these disadvantages, the method of matching with a known conditioning set that produces condition (M-2) does not require separability of outcome or choice equations, exogeneity of conditioning variables, exclusion restrictions, or adoption of specific functional forms of outcome equations. Such features are commonly used in conventional selection (control function) methods and conventional applications of IV, although recent work in semiparametric estimation relaxes these assumptions. As noted in subsection 3.2, the method of matching does not strictly require (M-1). One can get by with weaker mean independence assumptions (M-3) in the place of the stronger conditions (M-1). However, if (M-3) is invoked, the assumption that one can replace X by $P(X)$ does not follow from the analysis of Rosenbaum and Rubin (1983), and is an additional new assumption.

Methods for implementing matching are provided in Heckman, Ichimura, Smith, and Todd (1998) and are discussed extensively in Heckman, LaLonde, and Smith (1999). See Todd (1999, 2008, 2007) for software and extensive discussion of the mechanics of matching. I now contrast the identifying assumptions used in the method of control functions with those used in matching.

3.4 Comparing Matching and Control Functions Approaches

The method of matching eliminates the dependence between (Y_0, Y_1) and D , $(Y_0, Y_1) \not\perp D$, by assuming access to conditioning variables X such that (M-1) is satisfied: $(Y_0, Y_1) \perp D | X$. By conditioning on observables, one can identify the distributions of Y_0 and Y_1 over the support of X satisfying (M-2).

Other methods model the dependence that gives rise to the spurious relationship and in this way attempt to eliminate it. IV involves exclusion and a different type of conditional independence, $(Y_0, Y_1) \perp Z | X$, as well as a rank condition ($\Pr(D=1 | X, Z)$ depends on Z). The instrument Z plays the role of the implicit

randomization used in matching by allocating people to treatment status in a way that does not depend on (Y_0, Y_1) . I have already established that matching and IV make very different assumptions. Thus, in general, a matching assumption that $(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z$ neither implies nor is implied by $(Y_0, Y_1) \perp\!\!\!\perp Z \mid X$. One special case where they are equivalent is when treatment status is assigned by randomization with full compliance (letting $\xi = 1$ denote assignment to treatment, $\xi = 1 \Rightarrow A = 1$ and $\xi = 0 \Rightarrow A = 0$) and $Z = \xi$, so that the instrument is the assignment mechanism. $A = 1$ if the person actually receives treatment, and $A = 0$ otherwise.

The method of control functions explicitly models the dependence between (Y_0, Y_1) and D and attempts to eliminate it. Matzkin (2007) provides a comprehensive review of these methods. I previously presented a summary of some of the general principles underlying the method of control functions, the method of control variates, replacement functions, and proxy approaches as they apply to the selection problem that underlies the evaluation problem studied in this paper. All of these methods attempt to eliminate the θ in (U-1) that produces the dependence captured in (U-2).

This section relates matching to the form of the control function introduced in Heckman (1980) and Heckman and Robb (1985a, 1986a). I analyze conditional means because of their familiarity. Using the fact that $E(\mathbf{1}(Y \leq y) \mid X) = F(y \mid X)$, the analysis applies to marginal distributions as well.

Thus I work with conditional expectations of (Y_0, Y_1) given (X, Z, D) , where Z is assumed to include at least one variable not in X . Conventional applications of the control function method assume additive separability, which is not required in matching. Strictly speaking, additive separability is not required in the application of control functions either.⁴¹ What is required is a model relating the outcome unobservables to the observables and the unobservables in the choice of treatment equation. Various assumptions give operational content to (U-1).

For the additively separable case (7), the control function for mean outcomes models the conditional expectations of Y_1 and Y_0 given X , Z , and D as

$$\begin{aligned} E(Y_1 \mid Z, X, D = 1) &= \mu_1(X) + E(U_1 \mid Z, X, D = 1) \\ E(Y_0 \mid Z, X, D = 0) &= \mu_0(X) + E(U_0 \mid Z, X, D = 0). \end{aligned}$$

In the traditional method of control functions, the analyst models $E(U_1 \mid Z, X, D = 1)$ and $E(U_0 \mid Z, X, D = 0)$. If these functions can be independently varied against $\mu_1(X)$ and $\mu_0(X)$ respectively, one can identify $\mu_1(X)$ and $\mu_0(X)$ up to constant terms.⁴² It is not required that X or Z be stochastically independent of U_1 or U_0 , although conventional methods often assume this.

⁴¹Examples of nonseparable selection models are found in Cameron and Heckman (1998). See also Altonji and Matzkin (2005) and Matzkin (2007).

⁴²Heckman and Robb (1985a, 1986a) introduce this general formulation of control func-

Assume that $(U_1, U_0, V) \perp\!\!\!\perp (X, Z)$ and adopt equation (12) as the treatment choice model augmented so that X and Z are determinants of treatment choice, using V as the latent variable that generates D given X, Z : $D = \mathbf{1}(\mu_D(Z) \geq V)$. Let $U_D = F_{V|X}(V)$ and $P(Z) = F_{V|X}(\mu_D(Z))$. In this notation, the control functions are

$$E(U_1|Z, D=1) = E(U_1|\mu_D(Z) \geq V) = E(U_1 | P(Z) \geq U_D) = K_1(P(Z))$$

and

$$E(U_0|Z, D=0) = E(U_0|\mu_D(Z) < V) = E(U_0 | P(Z) < U_D) = K_0(P(Z)),$$

so the control function only depends on the propensity score $P(Z)$. The key assumption needed to represent the control function solely as a function of $P(Z)$ is

$$(U_1, U_0, V) \perp\!\!\!\perp X, Z. \quad (CF-1)^{43}$$

Under this condition

$$\begin{aligned} E(Y_1|Z, X, D=1) &= \mu_1(X) + K_1(P(Z)), \\ E(Y_0|Z, X, D=0) &= \mu_0(X) + K_0(P(Z)), \end{aligned}$$

with $\lim_{P \rightarrow 1} K_1(P) = 0$ and $\lim_{P \rightarrow 0} K_0(P) = 0$. It is assumed that Z can be independently varied for all X , and the limits are obtained by changing Z while holding X fixed.⁴⁴ These limit results state that when the values of X, Z are such that the probability of being in a sample ($D = 1$ or $D = 0$, respectively) is 1, there is no selection bias and one can separate out $\mu_1(X)$ from $K_1(P(Z))$ and $\mu_0(X)$ from $K_0(P(Z))$. This is the same identification at infinity condition that is required to identify ATE and TT in IV for models with heterogeneous responses.^{45,46}

Unlike the method of matching based on (M-1), the method of control functions allows the marginal treatment effect to be different from the average treatment effect and from the conditional effect of treatment on the treated. Although conventional practice has been to derive the functional forms of $K_0(P)$

tions. The identifiability requires that the members of the pairs $(\mu_1(X), E(U_1|X, Z, D=1))$ and $(\mu_0(X), E(U_0|X, Z, D=0))$ be variation-free so that they can be independently varied against each other.

⁴³This assumption is not strictly required. One can condition on X throughout.

⁴⁴More precisely, I assume that $\text{Supp}(Z|X) = \text{Supp}(Z)$ and that limit sets of $Z, \mathbb{Z}_0$, and $\mathbb{Z}_1$ exist so that as $Z \rightarrow \mathbb{Z}_0$, $P(Z, X) \rightarrow 0$, and as $Z \rightarrow \mathbb{Z}_1$, $P(Z, X) \rightarrow 1$.

⁴⁵One needs identification at infinity to obtain ATE and TT. This is a general feature of any evaluation model with general heterogeneity.

⁴⁶One can approximate the $K_1(P)$ and $K_0(P)$ terms by polynomials in P (see Heckman, 1980, Heckman and Robb, 1985a, 1986a, Heckman and Hotz, 1989). Ahn and Powell (1993) and Powell (1994) develop methods for eliminating $K_1(P(Z))$ and $K_0(P(Z))$ by differentiating.

and $K_1(P)$ by making distributional assumptions about (U_0, U_1, V) such as normality or other conventional distributional assumptions, this is not an intrinsic feature of the method and there are many nonnormal and semiparametric versions of this method. See Powell (1994) for a survey.

In its semiparametric implementation, the method of control functions requires an exclusion restriction (a variable in Z not in X) to achieve nonparametric identification.⁴⁷ Without any functional-form assumptions one cannot rule out a worst case analysis where, for example, if $X = Z$, then $K_1(P(X)) = \alpha\mu_1(X)$ where α is a scalar. In this situation, there is perfect collinearity between the control function and the conditional mean of the outcome equation, and it is impossible to separately identify either.⁴⁸ Even though this case is not generic, it is possible. The method of matching does not require an exclusion restriction, but achieves identification at the cost of ruling out essential heterogeneity. In the general case, the method of control functions requires that in certain limit sets of Z , $P(Z) = 1$ and $P(Z) = 0$ in order to achieve full nonparametric identification.⁴⁹ The conventional method of matching does not invoke such limit set arguments.

All methods of evaluation, including matching and control functions, require that treatment parameters be defined on a common support that is the intersection of the supports of X given $D = 1$ and X given $D = 0$: $\text{Supp}(X|D = 1) \cap \text{Supp}(X|D = 0)$. This is the requirement for any estimator that seeks to identify treatment effects by comparing samples of treated persons with samples of untreated persons. It is a great advantage of nonparametric methods that they impose this condition.

In this version of the method of control functions, $P(Z)$ is a conditioning variable used to predict U_1 conditional on D and U_0 conditional on D . In the method of matching, it is used as a conditioning variable to eliminate the stochastic independence between (U_0, U_1) and D . In the method of LATE or LIV, $P(Z)$ is used as an instrument. In the method of control functions, as conventionally applied, $(U_0, U_1) \perp\!\!\!\perp (X, Z)$, but this assumption is not intrinsic to the method.⁵⁰ This assumption plays no role in matching if the correct conditioning set is known.⁵¹ However, as noted below, exogeneity plays a key role in devising algorithms to select the conditioning variables. In addition, as developed in Heckman and Vytlačil (2005), exogeneity is helpful in making out-of-sample forecasts. The method of

⁴⁷No exclusion is required for many common functional forms for the distributions of unobservables.

⁴⁸Clearly $K_1(P(X))$ and $\mu_1(P)$ cannot be independently varied in this case. Olsen (1980) develops the worst case scenario.

⁴⁹Symmetry of the errors can be used in place of the appeal to limit sets that put $P(Z) = 0$ or $P(Z) = 1$. See Chen (1999).

⁵⁰Relaxing it, however, requires that the analyst model the dependence of the unobservables on the observables and that certain variation-free conditions are satisfied. (See Heckman and Robb, 1985a).

⁵¹That is, a conditioning set that satisfies (M-1) and (M-2).

control functions does not require that $(U_0, U_1) \perp\!\!\!\perp D | (X, Z)$, which is a central requirement of matching. Equivalently, the method of control functions does not require

$$(U_0, U_1) \perp\!\!\!\perp V | (X, Z), \quad \text{or that } (U_0, U_1) \perp\!\!\!\perp V | X,$$

whereas matching does and typically equates X and Z .

The method of control functions allows for outcome unobservables to be dependent on D even after conditioning on (X, Z) , and it models this dependence. The method of matching assumes no such D dependence. Thus in this regard, and maintaining all of the assumptions invoked for control functions in this section, matching is a special case of the method of control functions⁵² in which under assumptions (M-1) and (M-2),

$$\begin{aligned} E(U_1 | X, D = 1) &= E(U_1 | X) \\ E(U_0 | X, D = 0) &= E(U_0 | X). \end{aligned}$$

In the method of control functions, in the case where $(X, Z) \perp\!\!\!\perp (U_0, U_1, V)$, where the Z can include some or all of the elements of X , the conditional expectation of Y given X, Z, D is

$$\begin{aligned} E(Y | X, Z, D) &= E(Y_1 | X, Z, D = 1)D + E(Y_0 | X, Z, D = 0)(1 - D) \quad (20) \\ &= \mu_0(X) + [\mu_1(X) - \mu_0(X)]D \\ &\quad + E(U_1 | P(Z), D = 1)D + E(U_0 | P(Z), D = 0)(1 - D) \\ &= \mu_0(X) + K_0(P(Z)) + [\mu_1(X) - \mu_0(X) + K_1(P(Z)) \\ &\quad - K_0(P(Z))]D. \end{aligned}$$

The coefficient on D in the final equation combines $\mu_1(X) - \mu_0(X)$ with $K_1(P(Z)) - K_0(P(Z))$. It does not correspond to any treatment effect. To identify $\mu_1(X) - \mu_0(X)$, one must isolate it from $K_1(P(Z)) - K_0(P(Z))$.

Under assumptions (M-1) and (M-2) of the method of matching, the conditional expectation of Y conditional on $P(X)$ and D is

$$\begin{aligned} E(Y | P(X), D) &= \mu_0(P(X)) + E(U_0 | P(X)) \quad (21) \\ &\quad + [(\mu_1(P(X)) - \mu_0(P(X))) + E(U_1 | P(X)) \\ &\quad - E(U_0 | P(X))]D. \end{aligned}$$

⁵²See Aakvik, Heckman, and Vytlačil (2005), Carneiro, Hansen, and Heckman (2003) and Cunha, Heckman, and Navarro (2005) for a generalization of matching that allows for selection on unobservables by imposing a factor structure on the errors and estimating the distribution of the unobserved factors. These methods are discussed in Abbring and Heckman (2007).

The coefficient on D in this expression is now interpretable and is the average treatment effect. If one assumes that $(U_0, U_1) \perp\!\!\!\perp X$, which is not strictly required, one obtains a familiar representation

$$E(Y|P(X), D) = \mu_0(P(X)) + [\mu_1(P(X)) - \mu_0(P(X))]D, \quad (22)$$

since $E(U_1|P(X)) = E(U_0|P(X)) = 0$. A parallel derivation can be made conditioning on X instead of $P(X)$.

Under the assumptions that justify matching, treatment effects ATE or TT (conditional on $P(X)$) are identified from the coefficient on D in either (21) or (22). Condition (M-2) guarantees that D is not perfectly predictable by X (or $P(X)$), so the variation in D identifies the treatment parameter.

The coefficient on D in equation (20) for the more general control function model does not correspond to any treatment parameter, whereas the coefficients on D in equations (21) and (22) correspond to treatment parameters under the assumptions of the matching model. Under the additional assumption (CF-1), $\mu_1(P(X)) - \mu_0(P(X)) = \text{ATE}$ and $\text{ATE} = \text{TT} = \text{MTE}$, so the method of matching identifies all of the (conditional on $P(X)$) mean treatment parameters.⁵³

Under the assumptions justifying matching, when the means of Y_1 and Y_0 are the parameters of interest, and X satisfies (M-1) and (M-2), the bias terms vanish. They do not vanish in the more general case considered by the method of control functions. This is the mathematical counterpart of the randomization implicit in matching: conditional on X or $P(X)$, (U_1, U_0) are random with respect to D . The method of control functions allows these error terms to be nonrandom with respect to D and models the dependence. In the absence of functional form assumptions, it requires an exclusion restriction (a variable in Z not in X) to separate out $K_0(P(Z))$ from the coefficient on D . Matching produces identification without exclusion restrictions whereas identification with exclusion restrictions is a central feature of the control function method in the absence of functional form assumptions.

The fact that the control function approach allows for more general dependencies among the unobservables and the conditioning variables than the matching approach allows is implicitly recognized in the work of Rosenbaum (1995) and Robins (1997). Their “sensitivity analyses” for matching when there are unobserved conditioning variables are, in their essence, sensitivity analyses using control functions.⁵⁴ Aakvik, Heckman, and Vytlačil (2005), Carneiro, Hansen, and

⁵³This result also holds even if (CF-1) is not satisfied because $(U_0, U_1) \not\perp\!\!\!\perp X$. In this case, the treatment effects include the term

$$E(U_1 | P(X)) - E(U_0 | P(X)).$$

⁵⁴See also Vijverberg (1993) who does such a sensitivity analysis in a parametric selection model with an unidentified parameter.

Heckman (2003) and Cunha, Heckman, and Navarro (2005) explicitly model the relationship between matching and selection models using factor structure models, treating the omitted conditioning variables as unobserved factors and estimating their distribution. Abbring and Heckman (2007) discuss this work.

3.5 Comparing Matching and Classical Control Function Methods for a Generalized Roy Model

Figure 6 shows the contrast between the shape of the MTE and the OLS matching estimand as a function of p for the extended Roy model. The $MTE(p)$ shows its typical declining shape associated with diminishing returns, and the assumptions justifying matching are violated. Matching attempts to impose a flat $MTE(p)$ and therefore flattens the estimated $MTE(p)$ compared to its true value. It understates marginal returns at low levels of p (associated with unobservables that make it likely to participate in treatment) and overstates marginal returns at high levels of p .

To further illustrate the bias in matching and how the control function eliminates it, I perform sensitivity analyses under different assumptions about the parameters of the underlying selection model. In particular, I assume that the data are generated by the model of equations (10) and (11), where $\mu_D(Z) = Z\gamma$, $\mu_0(X) = \mu_0$, $\mu_1(X) = \mu_1$, and

$$\begin{aligned}(U_1, U_0, V)' &\sim N(0, \Sigma) \\ \text{corr}(U_j, V) &= \rho_{jV} \\ \text{Var}(U_j) &= \sigma_j^2; \quad j = \{0, 1\}.\end{aligned}$$

I assume in this section that $D = \mathbf{1}[\mu_D(Z) + V > 0]$, in conformity with the examples presented in Heckman and Navarro (2004), on which I build. This reformulation of choice model (12) entails a simple change in the sign of V . I assume that $Z \perp\!\!\!\perp (U_1, U_0, V)$. Using the selection formulae derived in Appendix A, one can write the biases conditional on $P(Z) = p$ as

$$\begin{aligned}\text{Bias TT}(Z = z) &= \text{Bias TT}(P(Z) = p) = \sigma_0 \rho_{0V} M(p) \\ \text{Bias ATE}(Z = z) &= \text{Bias ATE}(P(Z) = p) \\ &= M(p) [\sigma_1 \rho_{1V} (1 - p) + \sigma_0 \rho_{0V} p],\end{aligned}$$

where $M(p) = \frac{\phi(\Phi^{-1}(1-p))}{p(1-p)}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a standard normal random variable and the propensity score $P(z)$ is evaluated at $P(z) = p$. I assume that $\mu_1 = \mu_0$ so that the true average treatment effect is zero.

I simulate the mean bias for TT (Table 4) and ATE (Table 5) for different values of the ρ_{jV} and σ_j . The results in the tables show that, as one lets the

variances of the outcome equations grow, the value of the mean bias that we obtain can become substantial. With larger correlations between the outcomes and the unobservables generating choices, come larger biases. These tables demonstrate the greater generality of the control function approach, which models the bias rather than assuming it away by conditioning. Even if the correlation between the observables and the unobservables (ρ_{jV}) is small, so that one might think that selection on unobservables is relatively unimportant, we still obtain substantial biases if we do not control for relevant omitted conditioning variables. Only for special values of the parameters can one avoid bias by matching. These examples also demonstrate that sensitivity analyses can be conducted for analysis based on control function methods even when they are not fully identified. Vijverberg (1993) provides an example.

3.6 The Informational Requirements of Matching and the Bias When They Are Not Satisfied

In this Section I present some examples of when matching “works” and when it breaks down. This section extends the analysis of Heckman and Navarro (2004). In particular, I show how matching on some of the relevant information but not all can make the bias using matching worse for standard treatment parameters.

Section 1 discussed informational asymmetries between the econometrician and the agents whose behavior they are analyzing. The method of matching assumes that the econometrician has access to and uses all of the relevant information in the precise sense defined there. That means that the X that guarantees conditional independence (M-1) is available and is used. The concept of relevant information is a delicate one and it is difficult to find the true conditioning set.

Assume that the economic model generating the data is a generalized Roy model of the form

$$\begin{aligned} D^* &= Z\gamma + V && \text{where} \\ Z &\perp\!\!\!\perp V && \text{and} \\ V &= \alpha_{V1}f_1 + \alpha_{V2}f_2 + \varepsilon_V \\ D &= \begin{cases} 1 & \text{if } D^* \geq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and

$$\begin{aligned} Y_1 &= \mu_1 + U_1 && \text{where } U_1 = \alpha_{11}f_1 + \alpha_{12}f_2 + \varepsilon_1, \\ Y_0 &= \mu_0 + U_0 && \text{where } U_0 = \alpha_{01}f_1 + \alpha_{02}f_2 + \varepsilon_0. \end{aligned}$$

In this specification, $(f_1, f_2, \varepsilon_V, \varepsilon_1, \varepsilon_0)$ are assumed to be mean zero random variables that are mutually independent of each other and Z so that all the correlation among the elements of (U_0, U_1, V) is captured by $f = (f_1, f_2)$. Models that

take this form are known as factor models and have been applied in the context of selection models by Aakvik, Heckman, and Vytlačil (2005), Carneiro, Hansen, and Heckman (2001, 2003), and Hansen, Heckman, and Mullen (2004), among others. I keep implicit any dependence on X which may be general.

Generically, the minimal relevant information for this model when the factor loadings are not zero ($\alpha_{ij} \neq 0$) is, for general values of the factor loadings,

$$I_R = \{f_1, f_2\}.$$
⁵⁵

Recall that I assume independence between Z and all error terms. If the econometrician has access to I_R and uses it, (M-1) is satisfied conditional on I_R . Note that I_R plays the role of θ in (U-1). In the case where the economist knows I_R , the economist's information set $\sigma(I_E)$ contains the relevant information ($\sigma(I_E) \supseteq \sigma(I_R)$).

The agent's information set may include different variables. If one assumes that $\varepsilon_0, \varepsilon_1$ are shocks to outcomes not known to the agent at the time treatment decisions are made, but the agent knows all other aspects of the model, the agent's information is

$$I_A = \{f_1, f_2, Z, \varepsilon_V\}.$$

Under perfect certainty, the agent's information set includes ε_1 and ε_0 :

$$I_A = \{f_1, f_2, Z, \varepsilon_V, \varepsilon_1, \varepsilon_0\}.$$

In either case, all of the information available to the agent is not required to satisfy conditional independence (M-1). All three information sets guarantee conditional independence, but only the first is minimal relevant.

In the notation of Section 1, the observing economist may know some variables not in I_A, I_{R^*} or I_R but may not know all of the variables in I_R . In the following subsections, I study what happens when the matching assumption that $\sigma(I_E) \supseteq \sigma(I_R)$ does not hold. That is, I analyze what happens to the bias from matching as the amount of information used by the econometrician is changed. In order to get closed form expressions for the biases of the treatment parameters I make the additional assumption that

$$(f_1, f_2, \varepsilon_V, \varepsilon_1, \varepsilon_0) \sim N(0, \Sigma),$$

where Σ is a matrix with $(\sigma_{f_1}^2, \sigma_{f_2}^2, \sigma_{\varepsilon_V}^2, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_0}^2)$ on the diagonal and zero in all the non-diagonal elements. This assumption links matching models to conventional normal selection models. However, the examples based on this specification

⁵⁵Notice that for a fixed set of α_{ij} , the minimal information set is $(\alpha_{11} - \alpha_{01})f_1 + (\alpha_{12} - \alpha_{02})f_2$, which captures the dependence between D and $(Y_1 - Y_0)$. Notice further that one can also work with $I_{R'} = A \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$ for any nonsingular A , in place of I_R .

illustrate more general principles. I now analyze various commonly encountered cases.

3.6.1 The Economist Uses the Minimal Relevant Information: $\sigma(I_R) \subseteq \sigma(I_E)$

I begin by analyzing the case in which the information used by the economist is $I_E = \{Z, f_1, f_2\}$, so that the econometrician has access to a relevant information set and it is larger than the minimal relevant information set. In this case, it is straightforward to show that matching identifies all of the mean treatment parameters with no bias. The matching estimator has population mean

$$E(Y_1|D=1, I_E) - E(Y_0|D=0, I_E) = \mu_1 - \mu_0 + (\alpha_{11} - \alpha_{01})f_1 + (\alpha_{12} - \alpha_{02})f_2,$$

and all of the mean treatment parameters collapse to this same expression since, conditional on knowing f_1 and f_2 , there is no selection because $(\varepsilon_1, \varepsilon_0) \perp\!\!\!\perp V$. Recall that $I_R = \{f_1, f_2\}$ and the economist needs less information to achieve (M-1) than is contained in I_E for arbitrary choices of $\alpha_{11}, \alpha_{01}, \alpha_{12}, \alpha_{02}$.

In this case, the analysis of Rosenbaum and Rubin (1983) tells us that knowledge of (Z, f_1, f_2) and knowledge of $P(Z, f_1, f_2)$ are equally useful in identifying all of the treatment parameters conditional on P . If we write the propensity score as

$$\begin{aligned} P(I_E) &= \Pr\left(\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} > \frac{-Z\gamma - \alpha_{V1}f_1 - \alpha_{V2}f_2}{\sigma_{\varepsilon_V}}\right) \\ &= 1 - \Phi\left(\frac{-Z\gamma - \alpha_{V1}f_1 - \alpha_{V2}f_2}{\sigma_{\varepsilon_V}}\right) = p, \end{aligned}$$

the event $(D^* \leq 0, \text{ given } f = \tilde{f} \text{ and } Z = z)$ can be written as $\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} \leq \Phi^{-1}(1 - P(z, \tilde{f}))$, where Φ is the cdf of a standard normal random variable and $f = (f_1, f_2)$. I abuse notation slightly by using z as the realized fixed value of Z and $\tilde{f}$ as the realized value of f . The population matching condition (M-1) implies that

$$\begin{aligned}
& E\left(Y_1|D=1, P(I_E)=P(z, \tilde{f})\right) - E\left(Y_0|D=0, P(I_E)=P(z, \tilde{f})\right) \\
&= \mu_1 - \mu_0 + E\left(U_1|D=1, P(I_E)=P(z, \tilde{f})\right) \\
&\quad - E\left(U_0|D=0, P(I_E)=P(z, \tilde{f})\right) \\
&= \mu_1 - \mu_0 + E\left(U_1\left|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} > \Phi^{-1}\left(1 - P(z, \tilde{f})\right)\right.\right) \\
&\quad - E\left(U_0\left|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} \leq \Phi^{-1}\left(1 - P(z, \tilde{f})\right)\right.\right) \\
&= \mu_1 - \mu_0.
\end{aligned}$$

This expression is equal to all of the treatment parameters discussed in this paper, since

$$E\left(U_1\left|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} > \Phi^{-1}\left(1 - P(z, \tilde{f})\right)\right.\right) = \frac{Cov(U_1, \varepsilon_V)}{\sigma_{\varepsilon_V}} M_1\left(P(z, \tilde{f})\right)$$

and

$$E\left(U_0\left|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} \leq \Phi^{-1}\left(1 - P(z, \tilde{f})\right)\right.\right) = \frac{Cov(U_0, \varepsilon_V)}{\sigma_{\varepsilon_V}} M_0\left(P(z, \tilde{f})\right),$$

where

$$\begin{aligned}
M_1(P(z, \tilde{f})) &= \frac{\phi(\Phi^{-1}(1 - P(z, \tilde{f})))}{P(z, \tilde{f})} \\
M_0(P(z, \tilde{f})) &= -\frac{\phi(\Phi^{-1}(1 - P(z, \tilde{f})))}{1 - P(z, \tilde{f})},
\end{aligned}$$

where ϕ is the density of a standard normal random variable. As a consequence of the assumptions about mutual independence of the errors

$$Cov(U_i, \varepsilon_V) = Cov(\alpha_{i1}f_1 + \alpha_{i2}f_2 + \varepsilon_i, \varepsilon_V) = 0, \quad i = 0, 1.$$

In the context of the generalized Roy model, the case considered in this subsection is the one matching is designed to solve. Even though a selection model generates the data, the fact that the information used by the econometrician includes the minimal relevant information makes matching a correct solution to the selection problem. One can estimate the treatment parameters with no bias since, as a consequence of the assumptions, $(U_1, U_0) \perp\!\!\!\perp D|(f, Z)$, which is exactly what matching requires. The minimal relevant information set is even smaller. For

arbitrary factor loadings, one only needs to know (f_1, f_2) to secure conditional independence. One can define the propensity score solely in terms of f_1 and f_2 , and the Rosenbaum-Rubin result still goes through. The analysis in this section focuses on treatment parameters conditional on particular values of $P(Z, f) = P(z, \tilde{f})$, i.e., for fixed values of p , but we could condition more finely. Conditioning on $P(z, \tilde{f})$ defines the treatment parameters more coarsely. One can use either fine or coarse conditioning to construct the unconditional treatment effects.

In this example, using more information than what is in the relevant information set (i.e. using Z) is harmless. But this is not generally true. If $Z \not\perp (U_0, U_1, V)$, adding Z to the conditioning set can violate conditional independence assumption (M-1):

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid (f_1, f_2),$$

but

$$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid f_1, f_2, Z.$$

Adding extra variables can destroy the crucial conditional independence property of matching. I present an example of this point below. I first consider a case where $Z \perp (U_0, U_1, V)$ but the analyst conditions on Z and not (f_1, f_2) . In this case, there is selection on the unobservables that are not conditioned on.

3.6.2 The Economist does not Use All of the Minimal Relevant Information

Next, suppose that the information used by the econometrician is

$$I_E = \{Z\},$$

and there is selection on the unobservable (to the analyst) f_1 and f_2 , i.e. the factor loadings α_{ij} are all non zero. Recall that I assume that Z and the f are independent. In this case the event $\left(D^* \begin{smallmatrix} \leq \\ \geq \end{smallmatrix} 0, Z = z\right)$ is characterized by

$$\frac{\alpha_{V1}f_1 + \alpha_{V2}f_2 + \varepsilon_V}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} \begin{smallmatrix} \leq \\ \geq \end{smallmatrix} \Phi^{-1}(1 - P(z)).$$

Using the analysis presented in Appendix A, the bias for the different treatment parameters is given by

$$\text{Bias TT}(Z = z) = \text{Bias TT}(P(Z) = P(z)) = \eta_0 M(P(z)), \quad (23)$$

where $M(p) = M_1(P(z)) - M_0(P(z))$.

$$\text{Bias ATE}(Z = z) = \text{Bias ATE}(P(Z) = P(z)) = M(p(z))\{\eta_1[1 - P(z)] + \eta_0 P(z)\}, \quad (24)$$

where

$$\begin{aligned} \eta_1 &= \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2 + \alpha_{V2}\alpha_{12}\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} \\ \eta_0 &= \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2 + \alpha_{V2}\alpha_{02}\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}}. \end{aligned}$$

It is not surprising that matching on sets of variables that exclude the relevant conditioning variables produces bias for the conditional (on $P(z)$) treatment parameters. The advantage of working with a closed form expression for the bias is that it allows me to answer questions about the *magnitude* of this bias under different assumptions about the information available to the analyst, and to present some simple examples. I next use expressions (23) and (24) as benchmarks against which to compare the relative size of the bias when I enlarge the econometrician's information set beyond Z .

3.6.3 Adding Information to the Econometrician's Information Set I_E : Using Some but not All the Information from the Minimal Relevant Information Set I_R

Suppose that the econometrician uses more information but not all of the information in the minimal relevant information set. He still reports values of the parameters conditional on specific p values but now the model for p has different conditioning variables. For example, the data set assumed in the preceding section might be augmented or else the econometrician decides to use information previously available. In particular, assume that the econometrician's information set is

$$I'_E = \{Z, f_2\},$$

and that he uses this information set. Under conditions 1 and 2 presented below, the biases for the treatment parameters conditional on values of $P(z) = p$ are reduced in absolute value relative to their values in Section 3.6.2 by changing the conditioning set in this way. But these conditions are not generally satisfied, so that adding extra information does not necessarily reduce bias and may actually increase it. To show how this happens in the model considered here, I define expressions comparable to η_1 and η_0 for this case:

$$\begin{aligned}\eta'_1 &= \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \sigma_{\varepsilon_V}^2}} \\ \eta'_0 &= \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \sigma_{\varepsilon_V}^2}}.\end{aligned}$$

I compare the biases under the two cases using formulas (23)–(24), suitably modified, but keeping p fixed at a specific value even though this implies different conditioning sets in terms of $(z, \tilde{f})$.

Condition 1 *The bias produced by using matching to estimate TT is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if*

$$|\eta_0| > |\eta'_0|.$$

There is a similar result for ATE:

Condition 2 *The bias produced by using matching to estimate ATE is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if*

$$|\eta_1(1-p) + \eta_0p| > |\eta'_1(1-p) + \eta'_0p|.$$

Proof These conditions are a direct consequence of formulas (23) and (24), modified to allow for the different covariance structure produced by the information structure assumed in this section (replacing η_0 with η'_0 , η_1 with η'_1).

It is important to notice that I condition on the same value of p in deriving these expressions although the variables in P are different across different specifications of the model. Recall that propensity-score matching defines them conditional on $P = p$.

Conditions 1 and 2 do not always hold. In general, whether or not the bias will be reduced by adding additional conditioning variables depends on the relative importance of the additional information in both the outcome equations and on the signs of the terms inside the absolute value.

When is Condition (1) satisfied? Assume $\eta_0 > 0$ for all α_{02}, α_{V2} . Then $\eta_0 > \eta'_0$ if

$$\eta_0 = \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2 + (\alpha_{V2}^2)\left(\frac{\alpha_{02}}{\alpha_{V2}}\right)\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} > \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \sigma_{\varepsilon_V}^2}} = \eta'_0.$$

When $\frac{\alpha_{02}}{\alpha_{V2}} = 0$, clearly $\eta_0 < \eta'_0$. Adding information to the conditioning set increases bias. I can vary $\left(\frac{\alpha_{02}}{\alpha_{V2}}\right)$ holding all of the other parameters constant and hence can make the left hand side arbitrarily large.⁵⁶

As α_{02} increases, there is some critical value α_{02}^* beyond which $\eta_0 > \eta'_0$. If I assumed that $\eta_0 < 0$, however, the opposite conclusion would hold, and the conditions for reduction in bias would be harder to meet, as the relative importance of the new information is increased. Similar expressions can be derived for ATE and MTE, in which the direction of the effect depends on the signs of the terms in the absolute value.

Figures 9a and 9b illustrate the point that adding some but not all information from the minimal relevant set might increase the point-wise bias and the unconditional or average bias for ATE and TT, respectively.⁵⁷ Values of the parameters of the model are presented at the base of the figures.

The fact that the point-wise (and overall) bias might increase when adding some but not all information from I_R is a feature that is not shared by the method of control functions. Because the method of control functions models the stochastic dependence of the unobservables in the outcome equations on the observables, changing the variables observed by the econometrician to include f_2 does not generate bias. It only changes the control function used. That is, by adding f_2 one changes the control function from

$$\begin{aligned} K_1(P(Z) = P(z)) &= \eta_1 M_1(P(z)) \\ K_0(P(Z) = P(z)) &= \eta_0 M_0(P(z)) \end{aligned}$$

to

$$\begin{aligned} K'_1(P(Z, f_2) = P(z, \tilde{f}_2)) &= \eta'_1 M_1(P(z, \tilde{f}_2)) \\ K'_0(P(Z, f_2) = P(z, \tilde{f}_2)) &= \eta'_0 M_0(P(z, \tilde{f}_2)) \end{aligned}$$

but does not generate any bias by using the control function estimator. This is a major advantage of this method. It controls for the bias of the omitted conditioning variables by modeling it. Of course, if the model for the bias term is not valid, neither is the correction for the bias. Semiparametric selection estimators are

---

⁵⁶A direct computation shows that

$$\frac{\partial \eta_0}{\partial \left(\frac{\alpha_{02}}{\alpha_{V2}}\right)} = \frac{\alpha_{V2}^2 \sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} > 0.$$

⁵⁷Heckman and Navarro (2004) show comparable plots for MTE.

designed to protect the analyst against model misspecification. (See, e.g. Powell, 1994). Matching evades this problem by assuming that the analyst always knows the correct conditioning variables and that they satisfy (M-1). In actual empirical settings, agents rarely know the relevant information set. Instead they use proxies.

3.6.4 Adding Information to the Econometrician's Information Set: Using Proxies for the Relevant Information

Suppose that instead of knowing some part of the minimal relevant information set, such as f_2 , the analyst has access to a proxy for it.⁵⁸ In particular, assume that he has access to a variable $\tilde{Z}$ that is correlated with f_2 but that is not the full minimal relevant information set. That is, define the econometrician's information to be

$$\tilde{I}_{E^*} = \{Z, \tilde{Z}\},$$

and suppose that he uses it so $I_E = \tilde{I}_{E^*}$. In order to obtain closed-form expressions for the biases I assume that

$$\begin{aligned} \tilde{Z} &\sim N(0, \sigma_{\tilde{Z}}^2) \\ \text{corr}(\tilde{Z}, f_2) &= \rho, \text{ and } \tilde{Z} \perp (\varepsilon_0, \varepsilon_1, \varepsilon_V, f_1). \end{aligned}$$

Define expressions comparable to η and η' :

$$\begin{aligned} \tilde{\eta}_1 &= \frac{\alpha_{11}\alpha_{V1}\sigma_{f_1}^2 + \alpha_{12}\alpha_{V2}(1 - \rho^2)\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2(1 - \rho^2) + \sigma_{\varepsilon_V}^2}} \\ \tilde{\eta}_0 &= \frac{\alpha_{01}\alpha_{V1}\sigma_{f_1}^2 + \alpha_{02}\alpha_{V2}(1 - \rho^2)\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2(1 - \rho^2) + \sigma_{\varepsilon_V}^2}}. \end{aligned}$$

By substituting for I'_E by $\tilde{I}_E$ and η'_j by $\tilde{\eta}_j$ ($j = 0, 1$) in Conditions (1) and (2) of Section 3.6.3 I can obtain results for the bias. Whether $\tilde{I}_E$ will be bias-reducing depends on how well it spans I_R and on the signs of the terms in the absolute values in those conditions in Section 3.6.3.

In this case, however, there is another parameter to consider: the correlation ρ between $\tilde{Z}$ and f_2 , ρ . If $|\rho| = 1$ we are back to the case of $\tilde{I}_E = I'_E$ because $\tilde{Z}$

⁵⁸For example, the returns-to-schooling literature often uses different test scores, like AFQT or IQ, to proxy for missing ability variables. These proxy replacement function methods are discussed in Heckman and Vytlačil (2007b, Section 11) and in Abbring and Heckman (2007).

is a perfect proxy for f_2 . If $\rho = 0$, we are essentially back to the case analyzed in Section 3.6.3. Because we know that the bias at a particular value of p might either increase or decrease when f_2 is used as a conditioning variable but f_1 is not, we know that it is not possible to determine whether the bias increases or decreases as we change the correlation between f_2 and $\tilde{Z}$. That is, we know that going from $\rho = 0$ to $|\rho| = 1$ might change the bias in any direction. Use of a better proxy in this correlational sense may produce a *more* biased estimate.

From the analysis of Section 3.6.3, it is straightforward to derive conditions under which the bias generated when the econometrician's information is $\tilde{I}_E$ is smaller than when it is I'_E . That is, it can be the case that knowing *the proxy* variable $\tilde{Z}$ is *better* than knowing the actual variable f_2 . Take again the analysis of treatment on the treated as an example (i.e., Condition (1)). The bias in absolute value (at a fixed value of p) is reduced when $\tilde{Z}$ is used instead of f_2 if

$$\left| \frac{\alpha_{01}\alpha_{V1}\sigma_{f_1}^2 + \alpha_{02}\alpha_{V2}(1-\rho^2)\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2(1-\rho^2) + \sigma_{\varepsilon_V}^2}} \right| < \left| \frac{\alpha_{01}\alpha_{V1}\sigma_{f_1}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \sigma_{\varepsilon_V}^2}} \right|.$$

Figures 10a and 10b, use the same true model as used in the previous section to illustrate the two points being made here. Namely, *using a proxy* for an unobserved relevant variable *might increase the bias*. On the other hand, it *might be better* in terms of bias to use a *proxy* than to use the actual variable, f_2 . However, as figures 11a and 11b show, by changing α_{02} from 0.1 to 1, using a proxy might increase the bias versus using the actual variable f_2 . Notice that the bias need not be universally negative or positive but depends on p .

The general point of these examples is that matching makes very knife-edge assumptions. If the analyst gets the right conditioning set, (M-1) is satisfied and there is no bias. But determining the correct information set is not a trivial task, as I note below. Having good proxies in the standard usage of that term can create substantial bias in estimating treatment effects.

3.6.5 The Case of a Discrete Outcome Variable

The examples given in this section do not depend on all of the assumptions I have made to produce simple examples. In particular, I require neither normality nor additive separability of the outcome equations. The proposition that matching identifies the correct treatment effect if the econometrician's information set includes all the minimal relevant information is true more generally, provided that any additional extraneous information used is exogenous in a sense to be defined precisely in the next section. Heckman and Navarro (2004) present parallel analyses of discrete treatment working with odds ratios and discrete outcomes that do not rely on either normality or separability of outcome equations.

3.6.6 On the Use of Model Selection Criteria to Choose Matching Variables

I have previously shown by way of example that adding more variables from the minimal relevant information set, but not all variables in it, may increase bias. By a parallel argument, adding additional variables to the relevant conditioning set may make the bias worse. Although I have used a prototypical Roy model as the point of departure, clearly the point is more general.

There is no rigorous rule for choosing the conditioning variables that produce (M-1). Adding variables that are statistically significant in the treatment choice equation is not guaranteed to select a set of conditioning variables that satisfies condition (M-1). This is demonstrated by the analysis of subsection 3.6.3 that shows that adding f_2 when it determines D may increase bias at any selected value of p .

The existing literature (e.g. Heckman, Ichimura, Smith, and Todd, 1998) proposes criteria based on selecting a set of conditioning variables based on a goodness of fit criterion (λ), where a higher λ means a better fit in the equation predicting D . The intuition behind such criteria is that by using some measure of goodness of fit as a guiding principle one is using information relevant to the decision process. In the example of subsection 3.6.3, it is clear that knowing f_2 improves goodness of fit, so that in general such a rule is deficient if f_1 is not known or is not used.

An implicit assumption underlying such procedures is that the added conditioning variables $\mathcal{X}$ are exogenous in the following sense:

$$(Y_0, Y_1) \perp\!\!\!\perp D | I_{\text{init}}, \mathcal{X} \quad (\text{E-1})$$

where I_{init} is interpreted as the variables initially used as conditioning variables before $\mathcal{X}$ is added. Failure of exogeneity is a failure of (M-1) for the augmented conditioning set, and matching estimators based on the augmented information set $(I_{\text{init}}, \mathcal{X})$ are biased when the condition is not satisfied.

Exogeneity assumption (E-1) is not usually invoked in the matching literature, which largely focuses on problem P-1, evaluating a program in place, rather than extrapolating to new environments (P-2). Indeed, the robustness of matching to such exogeneity assumptions is often viewed as an advantage of the method. In this section, I show some examples which illustrate the general point that standard model selection criteria fail to produce correctly specified conditioning sets unless some version of exogeneity condition (E-1) is satisfied.

In the literature, the use of model selection criteria is justified in two different ways. Sometimes it is claimed that they provide a *relative* guide. Sets of variables with higher λ (better goodness of fit) are alleged to be better than sets of variables with lower λ in the sense that they generate lower biases. However, I have already shown that this is not true. Enlarging the analyst's information from $I_{\text{init}} = \{Z\}$ to $I'_{\text{init}} = \{Z, f_2\}$ will improve fit since f_2 is also in I_A and I_R . But, going from

I_{init} to I'_{init} might increase the bias. So it is not true that combinations of variables that increase some measure of fit λ necessarily reduce the bias. Table 6 illustrates this point using the normal example. Going from row 1 to row 2 (adding f_2) improves goodness of fit and increases the unconditional or overall bias for all three treatment parameters, because (E-1) is violated.

The following rule of thumb argument is sometimes invoked as an absolute standard against which to compare alternative models. In versions of the argument, the analyst asserts that there is a combination of variables I'' that satisfy (M-1) and hence produces zero bias and a value of λ larger than that of any other I . In the examples, conditioning on $\{Z, f_1, f_2\}$ generates zero bias. One can exclude Z and still obtain zero bias. Because Z is a determinant of D , this shows immediately that the best fitting model does not necessarily identify the minimal relevant information set. In this example including Z is innocuous because there is still zero bias and the added conditioning variables satisfies (E-1) where $I_{\text{init}} = (f_1, f_2)$. In general, such a rule is not innocuous if Z is not exogenous. If goodness of fit is used as a rule to choose variables on which to match, there is no guarantee it produces a desirable conditioning set. If one includes in the conditioning set variables $\mathcal{X}$ that violate (E-1), they may improve the fit of predicted probabilities but worsen the bias.

Heckman and Navarro (2004) produce a series of examples that have the following feature. Variables S (shown at the base of Table 6) are added to the information set that improve the prediction of D but are correlated with (U_0, U_1) . Their particular examples use imperfect proxies (S_1, S_2) for (f_1, f_2) . The point is more general. The S variables fail exogeneity and produce greater bias for TT and ATE but they improve the prediction of D as measured by the correct in-sample prediction rate and the pseudo- R^2 . See the bottom two rows of Table 6.

4. Summary and Conclusions

This paper exposit the key identifying assumptions of commonly used econometric evaluation estimators. The emphasis is on the economic content of these assumptions. Using the prototypical generalized Roy model as the point of departure and building on the analysis of Barros (1987), I examine the assumptions underlying the method of matching. In brief, they are: (a) the analyst knows the relevant information set, (b) a randomization can be produced by conditioning on regressors, and (c) conditioning on observables, marginal returns are the same as average returns. Other econometric estimators do not impose (c) and account for failures of (a) and (b), and in this sense are more general.

References

- Aakvik, A., Heckman, J. J., and Vytlacil, E. J. (1999). Training effects on employment when the training effects are heterogeneous: An application to Norwegian vocational rehabilitation programs, University of Bergen Working Paper 0599; and University of Chicago.
- Aakvik, A., Heckman, J. J., and Vytlacil, E. J. (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs, *Journal of Econometrics* 125, 15–51.
- Abadie, A. (2002). Bootstrap tests of distributional treatment effects in instrumental variable models, *Journal of the American Statistical Association* 97, 284–292.
- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects, *Econometrica* 74, 235–267.
- Abbring, J. H. and Heckman, J. J. (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation, in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol. 6B, Amsterdam: Elsevier, 5145–5303.
- Ahn, H. and Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism, *Journal of Econometrics* 58, 3–29.
- Aigner, D. J. (1985). The residential electricity time-of-use pricing experiments: What have we learned?, in J. A. Hausman and D. A. Wise, eds., *Social Experimentation*, Chicago: University of Chicago Press, 11–41.
- Aigner, D. J., Hsiao, C., Kapteyn, A., and Wansbeek, T. (1984). Latent variable models in econometrics, in Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*, vol. 2, chap. 23, Amsterdam: Elsevier, 1321–1393.
- Altonji, J. G. and Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors, *Econometrica* 73, 1053–1102.
- Angrist, J. D., Graddy, K., and Imbens, G. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish, *Review of Economic Studies* 67, 499–527.
- Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics, in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, vol. 3A, New York: North-Holland, 1277–1366.
- Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models, *Econometrica* 74, 431–497.

Barnow, B. S., Cain, G. G., and Goldberger, A. S. (1980). Issues in the analysis of selectivity bias, in E. Stromsdorfer and G. Farkas, eds., *Evaluation Studies*, vol. 5, Beverly Hills, California: Sage Publications, 42–59.

Barros, R. P. d. (1987). *The Impact of Social Interventions: Nonparametric Identification From Choice Based Samples*, Ph.D. thesis, University of Chicago.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?, *Quarterly Journal of Economics* 119, 249–275.

Björklund, A. and Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection, *Review of Economics and Statistics* 69, 42–49.

Blundell, R., Duncan, A., and Meghir, C. (1998). Estimating labor supply responses using tax reforms, *Econometrica* 66, 827–861.

Blundell, R. and Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models, in L. P. H. M. Dewatripont and S. J. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol. 2, Cambridge, UK: Cambridge University Press.

Blundell, R. and Powell, J. (2004). Endogeneity in semiparametric binary response models, *Review of Economic Studies* 71, 655–679.

Cameron, S. V. and Heckman, J. J. (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males, *Journal of Political Economy* 106, 262–333.

Carneiro, P. (2002). *Heterogeneity in the Returns to Schooling: Implications for Policy Evaluation*, Ph.D. thesis, University of Chicago.

Carneiro, P., Hansen, K., and Heckman, J. J. (2001). Removing the veil of ignorance in assessing the distributional impacts of social policies, *Swedish Economic Policy Review* 8, 273–301.

Carneiro, P., Hansen, K., and Heckman, J. J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice, *International Economic Review* 44, 361–422.

Chen, S. (1999). Distribution-free estimation of the random coefficient dummy endogenous variable model, *Journal of Econometrics* 91, 171–199.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review, *Sankhya* 35, 417–446.

Cunha, F. and Heckman, J. J. (2007). Identifying and estimating the distributions of *Ex Post* and *Ex Ante* returns to schooling: A survey of recent developments, *Labour Economics* 14, 870–893.

- Cunha, F. and Heckman, J. J. (2008). A new framework for the analysis of inequality, *Macroeconomic Dynamics* 12, 315–354.
- Cunha, F., Heckman, J. J., and Matzkin, R. (2003). Nonseparable factor analysis, Unpublished manuscript, University of Chicago, Department of Economics.
- Cunha, F., Heckman, J. J., and Navarro, S. (2005). Separating uncertainty from heterogeneity in life cycle earnings, The 2004 Hicks Lecture, *Oxford Economic Papers* 57, 191–261.
- Cunha, F., Heckman, J. J., and Navarro, S. (2006). Counterfactual analysis of inequality and social mobility, in S. L. Morgan, D. B. Grusky, and G. S. Fields, eds., *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, chap. 4, Stanford, CA: Stanford University Press, 290–348.
- Cunha, F., Heckman, J. J., and Schennach, S. M. (2006). Nonlinear factor analysis, Unpublished manuscript, University of Chicago, Department of Economics, revised 2008.
- Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation, *Econometrica* 78, 883–931.
- Fisher, F. M. (1966). *The Identification Problem in Econometrics*, New York: McGraw-Hill.
- Gerfin, M. and Lechner, M. (2002). A microeconomic evaluation of the active labor market policy in Switzerland, *Economic Journal* 112, 854–893.
- Gill, R. D. and Robins, J. M. (2001). Causal inference for complex longitudinal data: The continuous case, *Annals of Statistics* 29, 1785–1811.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica* 66, 315–31.
- Hansen, K. T., Heckman, J. J., and Mullen, K. J. (2004). The effect of schooling and ability on achievement test scores, *Journal of Econometrics* 121, 39–98.
- Härdle, W. (1990). *Applied Nonparametric Regression*, New York: Cambridge University Press.
- Heckman, J. J. (1980). Addendum to sample selection bias as a specification error, in E. Stromsdorfer and G. Farkas, eds., *Evaluation Studies Review Annual*, vol. 5, Beverly Hills: Sage Publications.
- Heckman, J. J. (1992). Randomization and social policy evaluation, in C. Manski and I. Garfinkel, eds., *Evaluating Welfare and Training Programs*, Cambridge, MA: Harvard University Press, 201–230.
- Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations, *Journal of Human Resources* 32, 441–462, addendum published vol. 33 no. 1 (Winter 1998).

Heckman, J. J. (1998). The effects of government policies on human capital investment, unemployment and earnings inequality, in *Third Public GAAC Symposium: Labor Markets in the USA and Germany*, vol. 5, Bonn, Germany: German-American Academic Council Foundation.

Heckman, J. J. (2008). Econometric causality, *International Statistical Review* 76, 1–27.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy, *Journal of Economic Literature* 48, 356–398.

Heckman, J. J. and Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of Manpower Training, *Journal of the American Statistical Association* 84, 862–874, rejoinder also published in Vol. 84, No. 408, (Dec. 1989).

Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. E. (1998). Characterizing selection bias using experimental data, *Econometrica* 66, 1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). How details make a difference: Semiparametric estimation of the partially linear regression model, Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1998). Matching as an econometric evaluation estimator, *Review of Economic Studies* 65, 261–294.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The economics and econometrics of active labor market programs, in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, vol. 3A, chap. 31, New York: North-Holland, 1865–2097.

Heckman, J. J. and Navarro, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models, *Review of Economics and Statistics* 86, 30–57.

Heckman, J. J. and Robb, R. (1985a). Alternative methods for evaluating the impact of interventions, in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, vol. 10, New York: Cambridge University Press, 156–245.

Heckman, J. J. and Robb, R. (1985b). Alternative methods for evaluating the impact of interventions: An overview, *Journal of Econometrics* 30, 239–267.

Heckman, J. J. and Robb, R. (1986a). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes, in H. Wainer, ed., *Drawing Inferences from Self-Selected Samples*, New York: Springer-Verlag, 63–107, reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J. and Robb, R. (1986b). Postscript: A rejoinder to Tukey, in H. Wainer, ed., *Drawing Inferences from Self-Selected Samples*, New York:

Springer-Verlag, 111–114, reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J., Smith, J. A., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts, *Review of Economic Studies* 64, 487–535.

Heckman, J. J., Urzúa, S., and Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity, *Review of Economics and Statistics* 88, 389–432.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects, *Proceedings of the National Academy of Sciences* 96, 4730–4734.

Heckman, J. J. and Vytlacil, E. J. (2001). Local instrumental variables, in C. Hsiao, K. Morimune, and J. L. Powell, eds., *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, New York: Cambridge University Press, 1–46.

Heckman, J. J. and Vytlacil, E. J. (2005). Structural equations, treatment effects and econometric policy evaluation, *Econometrica* 73, 669–738.

Heckman, J. J. and Vytlacil, E. J. (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation, in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol. 6B, Amsterdam: Elsevier, 4779–4874.

Heckman, J. J. and Vytlacil, E. J. (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments, in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol. 6B, Amsterdam: Elsevier, 4875–5144.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica* 71, 1161–1189.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models, *Econometrica* 76, 195–216.

Ichimura, H. and Todd, P. E. (2007). Implementing nonparametric and semiparametric estimators, in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol. 6B, Amsterdam: Elsevier, 5369–5468.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation, *American Economic Review* 93, 126–132.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review, *Review of Economics and Statistics* 86, 4 – 29.

- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects, *Econometrica* 62, 467–475.
- Imbens, G. W. and Newey, W. K. (2002). Identification and estimation of triangular simultaneous equations models without additivity, Technical Working Paper 285, National Bureau of Economic Research.
- Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption, in M. Lechner and F. Pfeiffer, eds., *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg: Physica/Springer.
- Matzkin, R. L. (2003). Nonparametric estimation of nonadditive random functions, *Econometrica* 71, 1339–1375.
- Matzkin, R. L. (2007). Nonparametric identification, in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*, vol. 6B, Amsterdam: Elsevier.
- Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry, *Econometrica* 64, 1263–1297.
- Olsen, R. J. (1980). A least squares correction for selectivity bias, *Econometrica* 48, 1815–1820.
- Pearl, J. (2000). *Causality*, Cambridge, England: Cambridge University Press.
- Powell, J. L. (1994). Estimation of semiparametric models, in R. Engle and D. McFadden, eds., *Handbook of Econometrics, Volume 4*, Amsterdam: Elsevier, 2443–2521.
- Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes, *Journal of the American Statistical Association* 53, 873–880.
- Quandt, R. E. (1972). A new approach to estimating switching regressions, *Journal of the American Statistical Association* 67, 306–310.
- Robins, J. M. (1997). Causal inference from complex longitudinal data. , in M. Berkane, ed., *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics*, New York: Springer-Verlag, 69–117.
- Rosenbaum, P. R. (1995). *Observational Studies*, New York: Springer-Verlag.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* 70, 41–55.
- Roy, A. (1951). Some thoughts on the distribution of earnings, *Oxford Economic Papers* 3, 135–146.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* 66, 688–701.

- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization, *Annals of Statistics* 6, 34–58.
- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association* 74, 318–328.
- Schennach, S. M. (2004). Estimation of nonlinear models with measurement error, *Econometrica* 72, 33–75.
- Telser, L. G. (1964). Iterative estimation of a set of linear regression equations, *Journal of the American Statistical Association* 59, 845–862.
- Todd, P. E. (1999). A practical guide to implementing matching estimators, Unpublished manuscript, University of Pennsylvania, Department of Economics. Prepared for the IADB meeting in Santiago, Chile.
- Todd, P. E. (2007). Evaluating social programs with endogenous program placement and selection of the treated, in *Handbook of Development Economics*, Amsterdam: Elsevier, forthcoming.
- Todd, P. E. (2008). Matching estimators, in S. Durlauf and L. E. Blume, eds., *The New Palgrave Dictionary of Economics*, New York: Palgrave Macmillan, forthcoming.
- Vijverberg, W. P. M. (1993). Measuring the unidentified parameter of the extended roy model of selectivity, *Journal of Econometrics* 57, 69–89.
- Vytlacil, E. J. (2002). Independence, monotonicity, and latent index models: An equivalence result, *Econometrica* 70, 331–341.
- Willis, R. J. and Rosen, S. (1979). Education and self-selection, *Journal of Political Economy* 87, S7–S36.

Appendix A: Selection Formula for the Matching Examples

Consider a generalized Roy model of the form $Y_1 = \mu_1 + U_1$; $Y_0 = \mu_0 + U_0$; $D^* = \mu_D(Z) + V$; $D = 1$ if $D^* \geq 0$, $D = 0$ otherwise; and $Y = DY_1 + (1 - D)Y_0$, where

$$\begin{aligned}(U_1, U_0, V)' &\sim N(0, \Sigma); \text{Var}(U_i) = \sigma_i^2 & i = 0, 1 \\ \text{Var}(V) &= \sigma_V^2; \text{Cov}(U_1, U_0) = \sigma_{10} \\ \text{Cov}(U_1, V) &= \sigma_{1V}; \text{Cov}(U_0, V) = \sigma_{0V}.\end{aligned}$$

Assume $Z \perp\!\!\!\perp (U_0, U_1, V)$. Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the pdf and the cdf of a standard normal random variable. Then, the propensity score for this model for $Z = z$ is given by:

$$\Pr(D^* > 0 | Z = z) = \Pr(V > -\mu_D(z)) = P(z) = \Phi\left(\frac{\mu_D(z)}{\sigma_V}\right).$$

Thus $\frac{\mu_D(z)}{\sigma_V} = \Phi^{-1}(P(z))$, and

$$\frac{-\mu_D(z)}{\sigma_V} = \Phi^{-1}(1 - P(z)).$$

The event $\left(V \begin{smallmatrix} \leq \\ \geq \end{smallmatrix} 0, Z = z\right)$ can be written as $\frac{V}{\sigma_V} \begin{smallmatrix} \leq \\ \geq \end{smallmatrix} -\frac{\mu_D(z)}{\sigma_V} \Leftrightarrow \frac{V}{\sigma_V} \begin{smallmatrix} \leq \\ \geq \end{smallmatrix} \Phi^{-1}(1 - P(z))$. We can write the conditional expectations required to get the biases for the treatment parameters as a function of $P(z) = p$. For U_1 :

$$\begin{aligned}E(U_1 | D^* > 0, Z = z) &= \frac{\sigma_{1V}}{\sigma_V} E\left(\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > \frac{-\mu_D(z)}{\sigma_V}\right) \\ &= \frac{\sigma_{1V}}{\sigma_V} E\left(\frac{V}{\sigma_V} \mid \frac{V}{\sigma_V} > \Phi^{-1}(1 - P(z))\right) \\ &= \eta_1 M_1(P(z))\end{aligned}$$

where

$$\eta_1 = \frac{\sigma_{1V}}{\sigma_V}.$$

Similarly for U_0 :

$$\begin{aligned}E(U_0 | D^* > 0, Z = z) &= \eta_0 M_1(P(z)) \\ E(U_0 | D^* < 0, Z = z) &= \eta_0 M_0(P(z)),\end{aligned}$$

where $\eta_0 = \frac{\sigma_{0V}}{\sigma_V}$ and $M_1(P(z)) = \frac{\phi(\Phi^{-1}(1-P(z)))}{P(z)}$ and $M_0(P(z)) = \frac{-\phi(\Phi^{-1}(1-P(z)))}{(1-P(z))}$ are inverse Mills ratio terms.

Substituting these into the expressions for the biases for the treatment parameters conditional on z we obtain

$$\begin{aligned}\text{Bias } TT(P(z)) &= \eta_0 M_1(P(z)) - \eta_0 M_0(P(z)) \\ &= \eta_0 M(P(z)),\end{aligned}$$

$$\begin{aligned}\text{Bias } ATE(P(z)) &= \eta_1 M_1(P(z)) - \eta_0 M_0(P(z)) \\ &= M(P(z)) (\eta_1 (1 - P(z)) + \eta_0 P(z)).\end{aligned}$$

List of Tables

1. Identifying assumptions under commonly used methods.
2. A. Treatment effects and estimands as weighted averages of the marginal treatment effect; B. Weights. Source: Heckman and Vytlačil (2005)
3. Treatment parameters and estimands in the generalized Roy example. Source: Heckman and Vytlačil (2005)
4. Mean bias for treatment on the treated. Source: Heckman and Navarro (2004)
5. Mean bias for average treatment effect. Source: Heckman and Navarro (2004)
6. Goodness-of-fit statistics and average bias. Source: Heckman and Navarro (2004)

List of Figures

1. Distribution of gains in the Roy economy. Source: Heckman, Urzúa, and Vytlačil (2006)
2. A. Weights for the marginal treatment effect for different parameters; B. Marginal treatment effect vs. linear instrumental variables and ordinary least squares weights. Source: Heckman and Vytlačil (2005)
3. Conditional expectation of Y on $P(Z)$ and the MTE, the extended Roy economy. Source: Heckman, Urzúa, and Vytlačil (2006)
4. A. Plot of the $E(Y | P(Z) = p)$; B. Plot of the identified marginal treatment effect from figure 4A (the derivative). Source: Heckman and Vytlačil (2005)
5. The local average treatment effect, the extended Roy economy. Source: Heckman, Urzúa, and Vytlačil (2006)
6. Treatment parameters and OLS matching as a function of $P(Z) = p$. Source: Heckman, Urzúa, and Vytlačil (2006)
7. The least squares extrapolation problem avoided by using nonparametric regression or matching
8. A. MTE vs. linear instrumental variables, ordinary least squares, and policy relevant treatment effect weights: when $P(Z)$ is the instrument; B. MTE vs. linear IV with $P(Z(1 + t(1[Z > 0]))) = P(Z, t)$ as an instrument, and policy relevant treatment effect weights; C. MTE vs. IV policy and policy relevant treatment effect weights. Source: Heckman and Vytlačil (2005)

9. Bias for treatment on the treated ($P(Z)$, $P(Z, f_2)$). Source: Heckman and Navarro (2004)
10. Bias for treatment on the treated ($P(Z)$, $P(Z, \tilde{Z})$) Source: Heckman and Navarro (2004)
11. Bias for treatment on the treated ($P(Z, f_2)$, $P(Z, \tilde{Z})$) Source: Heckman and Navarro (2004)

Table 1
Identifying Assumptions Under Commonly Used Methods

(Y_0, Y_1) are potential outcomes that depend on X .
 $D = \begin{cases} 1 & \text{if assigned (or chose) status 1} \\ 0 & \text{otherwise.} \end{cases}$

Z are determinants of D , θ is a vector of unobservables.
 For random assignments, A is a vector of actual treatment status. $A = 1$ if treated; $A = 0$ if not.
 $\xi = 1$ if a person is randomized to treatment status; $\xi = 0$ otherwise.

	Identifying Assumptions	Identifies marginal distributions?	Exclusion condition needed?
Random Assignment	$(Y_0, Y_1) \perp\!\!\!\perp \xi$, $\xi = 1 \implies A = 1, \xi = 0 \implies A = 0$ (full compliance) Alternatively, if self-selection is random with respect to outcomes, $(Y_0, Y_1) \perp\!\!\!\perp D$. Assignment can be conditional on X .	Yes	No
Matching	$(Y_0, Y_1) \not\perp\!\!\!\perp D$, but $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$, $0 < \Pr(D = 1 \mid X) < 1$ for all X . Thus D conditional on X is a nondegenerate random variable	Yes	No
Control Functions and Extensions	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp D \mid X, Z, \theta$. The method models dependence induced by θ or else proxies θ (replacement function) Version (i) Replacement functions (substitute out θ or components of θ by observables) (Blundell and Powell 2003, Heckman and Robb 1985b; Olley and Pakes 1996). Factor models (Carneiro, Hansen, and Heckman, 2003) allow for measurement error in the proxies. Version (ii) Integrate out θ assuming $\theta \perp\!\!\!\perp (X, Z)$ (Aakvik, Heckman, and Vytlacil 2005; Carneiro, Hansen, and Heckman 2003). Version (iii) For separable models for mean response expect θ conditional on X, Z, D as in standard selection models (control functions in the same sense of Heckman and Robb).	Yes	Yes (for semiparametric models)
IV	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp Z \mid X, Z$, $\Pr(D = 1 \mid Z)$ is a nondegenerate function of Z .	Yes	Yes

Source: Heckman and Vytlacil (2007b).

Table 2A
Treatment Effects and Estimands as Weighted Averages
of the Marginal Treatment Effect

$\text{ATE}(x) = E(Y_1 - Y_0 \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) du_D$
$\text{TT}(x) = E(Y_1 - Y_0 \mid X = x, D = 1) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) \omega_{\text{TT}}(x, u_D) du_D$
$\text{TUT}(x) = E(Y_1 - Y_0 \mid X = x, D = 0) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) \omega_{\text{TUT}}(x, u_D) du_D$
Policy Relevant Treatment Effect $(x) = E(Y_{a'} \mid X = x) - E(Y_a \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) \omega_{\text{PRTE}}(x, u_D) du_D$
for two policies a and a' that affect the Z but not the X
$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) \omega_{\text{IV}}^J(x, u_D) du_D$, given instrument J
$\text{OLS}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D) \omega_{\text{OLS}}(x, u_D) du_D$

Table 2B
Weights

$\omega_{\text{ATE}}(x, u_D) = 1$
$\omega_{\text{TT}}(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x) dp \right] \frac{1}{E(P \mid X = x)}$
$\omega_{\text{TUT}}(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x) dp \right] \frac{1}{E((1 - P) \mid X = x)}$
$\omega_{\text{PRTE}}(x, u_D) = \left[\frac{FP_{a', X}(u_D) - FP_{a, X}(u_D)}{\Delta \bar{P}} \right]$
$\omega_{\text{IV}}^J(x, u_D) = \left[\int_{u_D}^1 (J(Z) - E(J(Z) \mid X = x)) f_{J, P \mid X}(j, t \mid X = x) dt dj \right] \frac{1}{\text{Cov}(J(Z), D \mid X = x)}$
$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1 \mid X = x, U_D = u_D) \omega_1(x, u_D) - E(U_0 \mid X = x, U_D = u_D) \omega_0(x, u_D)}{\Delta^{\text{MTE}}(x, u_D)}$
$\omega_1(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x) dp \right] \left[\frac{1}{E(P \mid X = x)} \right]$
$\omega_0(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x) dp \right] \frac{1}{E((1 - P) \mid X = x)}$

Source: Heckman and Vytlacil (2005).

Table 3
Treatment Parameters and Estimands
in the Generalized Roy Example

Treatment on the Treated	0.2353
Treatment on the Untreated	0.1574
Average Treatment Effect	0.2000
Sorting Gain ^a	0.0353
Selection Bias ^b	-0.0628
Linear Instrumental Variables ^c	0.2013
Ordinary Least Squares	0.1725

$$^a TT - ATE = E(Y_1 - Y_0 \mid D = 1) - E(Y_1 - Y_0)$$

$$^b OLS - TT = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$$

^cUsing Propensity Score $P(Z)$ as the instrument.

Note: The model used to create Table 2 is the same as those used to create Figures 1A and 1B.

Source: Heckman and Vytlačil (2005).

Table 4
Mean Bias for Treatment on the Treated

ρ_{0V}	Average Bias ($\sigma_0 = 1$)	Average Bias ($\sigma_0 = 2$)
-1.00	-1.7920	-3.5839
-0.75	-1.3440	-2.6879
-0.50	-0.8960	-1.7920
-0.25	-0.4480	-0.8960
0.00	0.0000	0.0000
0.25	0.4480	0.8960
0.50	0.8960	1.7920
0.75	1.3440	2.6879
1.00	1.7920	3.5839
BIAS TT = $\rho_{0V} * \sigma_0 * M(p)$		
$M(p) = \frac{\phi(\Phi^{-1}(p))}{[p*(1-p)]}$		

Source: Heckman and Navarro (2004).

Table 5
Mean Bias for Average Treatment Effect
($\sigma_0 = 1$)

ρ_{0V}	-1.00	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1.00
$\rho_{1V}(\sigma_1 = 1)$									
-1.00	-1.7920	-1.5680	-1.3440	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0
-0.75	-1.5680	-1.3440	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240
-0.50	-1.3440	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480
-0.25	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480	0.6720
0	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480	0.6720	0.8960
0.25	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200
0.50	-0.4480	-0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440
0.75	-0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440	1.5680
1.00	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440	1.5680	1.7920
$\rho_{1V}(\sigma_1 = 2)$									
-1.00	-2.6879	-2.2399	-1.7920	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960
-0.75	-2.4639	-2.0159	-1.5680	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200
-0.50	-2.2399	-1.7920	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960	1.3440
-0.25	-2.0159	-1.5680	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200	1.5680
0	-1.7920	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960	1.3440	1.7920
0.25	-1.5680	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200	1.5680	2.0159
0.50	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960	1.3440	1.7920	2.2399
0.75	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200	1.5680	2.0159	2.4639
1.00	-0.8960	-0.4480	0	0.4480	0.8960	1.3440	1.7920	2.2399	2.6879
BIAS ATE = $\rho_{1V} * \sigma_1 * M_1(p) - \rho_{0V} * \sigma_0 * M_0(p)$									
$M_1(p) = \frac{\phi(\Phi^{-1}(p))}{p}$									
$M_0(p) = \frac{-\phi(\Phi^{-1}(p))}{1-p}$									

Source: Heckman and Navarro (2004).

Table 6

Variables in Probit	Goodness of fit statistics λ		Average Bias	
	Correct in-sample prediction rate	Pseudo R^2	TT	ATE
Z	66.88%	0.1284	1.1380	1.6553
Z, f_2	75.02%	0.2791	1.2671	1.9007
Z, f_1, f_2	83.45%	0.4844	0.0000	0.0000
Z, S_1	77.38%	0.3282	0.9612	1.3981
Z, S_2	92.25%	0.7498	0.9997	1.4541

Source: Heckman and Navarro (2004).

Model: $V = Z + f_1 + f_2 + \varepsilon_V$; $\varepsilon_V \sim N(0, 1)$; $Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$; $\varepsilon_1 \sim N(0, 1)$; $Y_0 = f_1 + 0.1f_2 + \varepsilon_0$; $\varepsilon_0 \sim N(0, 1)$; $S_1 = V + u_1$; $u_1 \sim N(0, 4)$; $S_2 = V + u_2$; $u_2 \sim N(0, 0.25)$; $f_1 \sim N(0, 1)$; $f_2 \sim N(0, 1)$.

Figure 1. Distribution of Gains
The Roy Economy
 $U_1 - U_0 \not\propto D$

TT= 2.666, TUT= -0.632
Return to Marginal Agent = $C = 1.5$
ATE = $\mu_1 - \mu_0 = \bar{\beta} = 0.2$

The Model	
Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 & \text{if } D^* > 0 \\ 0 & \text{if } D^* \leq 0 \end{cases}$
General Case	
$(U_1 - U_0) \not\propto D$ $ATE \neq TT \neq TUT$	

The Researcher Observes (Y, D, C)

$Y = \alpha + \beta D + U_0$ where $\beta = Y_1 - Y_0$

Parameterization

$\alpha = 0.67$ $(U_1, U_0) \sim N(\mathbf{0}, \Sigma)$ $D^* = Y_1 - Y_0 - C$
 $\bar{\beta} = 0.2$ $\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$ $C = 1.5$

Source: Heckman, Urzúa, and Vytlačil (2006).

Figure 2A
Weights for the Marginal Treatment Effect for Different Parameters

Figure 2B
Marginal Treatment Effect vs Linear Instrumental Variables and Ordinary Least Squares
Weights

$$\begin{array}{llll}
 Y_1 = \alpha + \bar{\beta} + U_1 & U_1 = \sigma_1 \varepsilon & \alpha = 0.67 & \sigma_1 = 0.012 \\
 Y_0 = \alpha + U_0 & U_0 = \sigma_0 \varepsilon & \bar{\beta} = 0.2 & \sigma_0 = -0.050 \\
 D = 1 \text{ if } Z - V > 0 & V = \sigma_V \varepsilon & \varepsilon \sim N(0, 1) & \sigma_V = -1.000 \\
 U_D = \Phi\left(\frac{V}{\sigma_V \sigma_\varepsilon}\right) & & & Z \sim N(-0.0026, 0.2700)
 \end{array}$$

Source: Heckman and Vytlacil (2005).

Figure 3. Conditional Expectation of Y on $P(Z)$ and the Marginal Treatment Effect (MTE)
The Extended Roy Economy
A. $E(Y|P(Z) = p)$

Outcomes		Choice Model
$Y_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \alpha + U_0$		$D = \begin{cases} 1 & \text{if } D^* > 0 \\ 0 & \text{if } D^* \leq 0 \end{cases}$
Case IA	Case IB	Case II
$U_1 = U_0$ $\bar{\beta} = \text{ATE} = \text{TT} = \text{TUT} = \text{IV}$	$U_1 - U_0 \perp\!\!\!\perp D$ $\bar{\beta} = \text{ATE} = \text{TT} = \text{TUT} = \text{IV}$	$U_1 - U_0 \not\perp\!\!\!\perp D$ $\bar{\beta} = \text{ATE} \neq \text{TT} \neq \text{TUT} \neq \text{IV}$

Parameterization		
Cases IA, IB and II	Cases IB and II	Case II
$\alpha = 0.67$ $\bar{\beta} = 0.2$	$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$	$D^* = Y_1 - Y_0 - \gamma Z$ $Z \sim N(\mu_Z, \mathbf{\Sigma}_Z)$ $\mu_Z = (2, -2)$ and $\mathbf{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$ $\gamma = (0.5, 0.5)$

Source: Heckman, Urzúa, and Vytlačil (2006).

Figure 4A
Plot of the $E(Y|P(Z) = p)$

Figure 4B
Plot of the Identified Marginal Treatment Effect from Figure 4A (the Derivative).

Note : Parameters for the general heterogeneous case are the same as those used in Figures 2A and 2B. For the homogeneous case we impose $U_1 = U_0$ ($\sigma_1 = \sigma_0 = 0.012$).

Source: Heckman and Vytlacil (2005).

Figure 5. The Local Average Treatment Effect:
The Extended Roy Economy
A. $E(Y|P(Z) = p)$ and $\Delta^{\text{LATE}}(p_\ell, p_{\ell+1})$

B. $\Delta^{\text{MTE}}(u_D)$ and $\Delta^{\text{LATE}}(p_\ell, p_{\ell+1})$

$$\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) = \frac{E(Y|P(Z) = p_{\ell+1}) - E(Y|P(Z) = p_\ell)}{p_{\ell+1} - p_\ell} = \frac{\int_{p_\ell}^{p_{\ell+1}} \Delta^{\text{MTE}}(u_D) du_D}{p_{\ell+1} - p_\ell}$$

$$\begin{aligned}\Delta^{\text{LATE}}(0.6, 0.9) &= -1.17 \\ \Delta^{\text{LATE}}(0.1, 0.35) &= 1.719\end{aligned}$$

Outcomes

Choice Model

$$\begin{aligned}Y_1 &= \alpha + \bar{\beta} + U_1 \\ Y_0 &= \alpha + U_0\end{aligned}\quad \begin{aligned}D &= \begin{cases} 1 & \text{if } D^* > 0 \\ 0 & \text{if } D^* \leq 0 \end{cases} \\ \text{with } D^* &= Y_1 - Y_0 - \gamma Z\end{aligned}$$

Parameterization

$$\begin{aligned}(U_1, U_0) &\sim N(\mathbf{0}, \mathbf{\Sigma}) \text{ and } Z \sim N(\mu_Z, \mathbf{\Sigma}_Z) \\ \mathbf{\Sigma} &= \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \mu_Z = (2, -2) \text{ and } \mathbf{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix} \\ \alpha &= 0.67, \bar{\beta} = 0.2, \gamma = (0.5, 0.5)\end{aligned}$$

Source: Heckman, Urzúa, and Vytlačil (2006).

Figure 6. Treatment Parameters and OLS/Matching as a Function of $P(Z) = p$

Parameter	Definition	Under Assumptions (*)
Marginal Treatment Effect	$E[Y_1 - Y_0 D^* = 0, P(Z) = p]$	$\bar{\beta} + \sigma_{U_1 - U_0} \Phi^{-1}(1 - p)$
Average Treatment Effect	$E[Y_1 - Y_0 P(Z) = p]$	$\bar{\beta}$
Treatment on the Treated	$E[Y_1 - Y_0 D^* > 0, P(Z) = p]$	$\bar{\beta} + \sigma_{U_1 - U_0} \frac{\phi(\Phi^{-1}(1 - p))}{p}$
Treatment on the Untreated	$E[Y_1 - Y_0 D^* \leq 0, P(Z) = p]$	$\bar{\beta} + \sigma_{U_1 - U_0} \frac{\phi(\Phi^{-1}(1 - p))}{1 - p}$
OLS/Matching on $P(Z)$	$E[Y_1 D^* > 0, P(Z) = p] - E[Y_0 D^* \leq 0, P(Z) = p]$	$\bar{\beta} + \left(\frac{\sigma_{U_1 - U_0}^2}{\sqrt{\sigma_{U_1 - U_0}^2}} \right) \left(\frac{1 - 2p}{p(1 - p)} \right) \phi(\Phi^{-1}(1 - p))$

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.
 (*): The model in this case is the same as the one presented below Figure 5.
 Source: Heckman, Urzúa, and Vytlačil (2006).

Figure 7. The Least Squares Extrapolation Problem Avoided by Using Nonparametric Regression or Matching

Source: Heckman and Vytlačil (2007b).

Figure 8

Marginal Treatment Effect vs Linear Instrumental Variables and Ordinary Least Squares
 Effect Weights: When $P(Z)$ is the Instrument
 The model parameters are given at the base of Figure 2

Source: Heckman and Vytlačil (2005).

Figure 9A. Bias for Treatment on the Treated

Figure 9B. Bias for Average Treatment Effect

Note: Using proxy $\tilde{Z}$ for f_2 increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.
Model:

$$\begin{aligned} V &= Z + f_1 + f_2 + \varepsilon_\nu; & Y_1 &= 2f_1 + 0.1f_2 + \varepsilon_1; & Y_0 &= f_1 + 0.1f_2 + \varepsilon_0 \\ \varepsilon_\nu &\sim N(0, 1); & \varepsilon_1 &\sim N(0, 1); & \varepsilon_0 &\sim N(0, 1) \\ f_1 &\sim N(0, 1); & f_2 &\sim N(0, 1) \end{aligned}$$

Source: Heckman and Navarro (2004).

Figure 10A. Bias for Treatment on the Treated

Figure 10B. Bias for Average Treatment Effect

Note: Using proxy $\tilde{Z}$ for f_2 increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.
Model:

$$\begin{aligned}
 V &= Z + f_1 + f_2 + \varepsilon_\nu; & Y_1 &= 2f_1 + 0.1f_2 + \varepsilon_1; & Y_0 &= f_1 + 0.1f_2 + \varepsilon_0 \\
 \varepsilon_\nu &\sim N(0, 1); & \varepsilon_1 &\sim N(0, 1); & \varepsilon_0 &\sim N(0, 1) \\
 f_1 &\sim N(0, 1); & f_2 &\sim N(0, 1)
 \end{aligned}$$

Source: Heckman and Navarro (2004).

Figure 11A. Bias for Treatment on the Treated

Figure 11B. Bias for Average Treatment Effect

Note: Using proxy $\tilde{Z}$ for f_2 increases the bias. Correlation $(\tilde{Z}, f_2) = 0.5$.
Model:

$$\begin{aligned}
 V &= Z + f_1 + f_2 + \varepsilon_\nu; & Y_1 &= 2f_1 + 0.1f_2 + \varepsilon_1; & Y_0 &= f_1 + 0.1f_2 + \varepsilon_0 \\
 \varepsilon_\nu &\sim N(0, 1); & \varepsilon_1 &\sim N(0, 1); & \varepsilon_0 &\sim N(0, 1) \\
 f_1 &\sim N(0, 1); & f_2 &\sim N(0, 1)
 \end{aligned}$$

Source: Heckman and Navarro (2004).