Chapter 71

ECONOMETRIC EVALUATION OF SOCIAL PROGRAMS, PART II: USING THE MARGINAL TREATMENT EFFECT TO ORGANIZE ALTERNATIVE ECONOMETRIC ESTIMATORS TO EVALUATE SOCIAL PROGRAMS, AND TO FORECAST THEIR EFFECTS IN NEW ENVIRONMENTS[*]

JAMES J. HECKMAN

The University of Chicago, USA

American Bar Foundation, USA

University College Dublin, Ireland

EDWARD J. VYTLACIL

Columbia University, USA

Contents

[*] This research was supported by NSF: 97-09-873, 00-99195, and SES-0241858 and NICHD: R01-HD32058-03. We have benefited from comments received from Thierry Magnac and Costas Meghir at the Handbook of Econometrics Conference, December 1998; general comments at the 2001 Chicago Conference; and specific and very helpful comments from Jaap Abbring, Thomas Amorde, Hugo Garduño, Seong Moon, Rodrigo Pinto, Heleno Pioner, Jean-Marc Robin, Peter Saveleyev, G. Adam Savvas, Daniel Schmierer, John Trujillo, Semih Tumen, Sergio Urzua and Jordan Weil. Parts of this document draw on joint work with Sergio Urzua.

Handbook of Econometrics, Volume 6B
Copyright © 2007 Elsevier B.V. All rights reserved
DOI: 10.1016/S1573-4412(07)06071-0

Abstract

This chapter uses the marginal treatment effect (MTE) to unify and organize the econometric literature on the evaluation of social programs. The marginal treatment effect is a choice-theoretic parameter that can be interpreted as a willingness to pay parameter for persons at a margin of indifference between participating in an activity or not. All of the conventional treatment parameters as well as the more economically motivated treatment effects can be generated from a baseline marginal treatment effect. All of the estimation methods used in the applied evaluation literature, such as matching, instrumental variables, regression discontinuity methods, selection and control function methods, make assumptions about the marginal treatment effect which we exposit. Models for multiple outcomes are developed. Empirical examples of the leading methods are presented. Methods are presented for bounding treatment effects in partially identified models, when the marginal treatment effect is known only over a limited support. We show how to use the marginal treatment in econometric cost benefit analysis, in defining limits of policy experiments, in constructing the average marginal treatment effect, and in forecasting the effects of programs in new environments.

Keywords

marginal treatment effect, policy evaluation, instrumental variables, forecasting new policies, econometric cost benefit analysis, regression discontinuity, matching, bounds

JEL classification: C10, C13, C50

1. Introduction

This part of our contribution to this Handbook reviews and extends the econometric literature on the evaluation of social policy. We organize our discussion around choice-theoretic models for objective and subjective outcomes of the sort discussed in Chapter 70. Specifically, we organize our discussion of the literature around the concept of the marginal treatment effect (MTE) that was introduced in Chapter 70. Using the marginal treatment effect, we define a variety of treatment effects and show how they can be generated by a single economic functional, the MTE. We then show what various econometric methods assume about the MTE.

In this part, we focus exclusively on microeconomic partial equilibrium evaluation methods, deferring analysis of general equilibrium issues to Abbring and Heckman (Chapter 72). Thus throughout this chapter, except when we discuss randomized evaluation of social programs, we assume that potential outcomes are not affected by interventions but choices among the potential outcomes are affected. Thus, we invoke policy invariance assumptions (PI-3) and (PI-4) of Chapter 70. We also focus primarily on mean responses, leaving analysis of distributions of responses for Abbring and Heckman, Chapter 72.

The plan of this chapter is as follows. In Section 2, we present some basic principles that underlie conventional econometric evaluation estimators. In Section 3, we define the marginal treatment effect in a two potential outcome model that is a semiparametric version of the generalized Roy model. We then show how treatment parameters can be generated as weighted averages of the MTE. We carefully distinguish the definition of parameters from issues of identification. Section 4 considers how instrumental variable methods that supplement the classical instrumental variable assumptions of econometrics can be used to identify treatment parameters. We discuss the crucial role of monotonicity assumptions in the recent IV literature.

They impart an asymmetry to the admissible forms of agent heterogeneity. Outcomes are permitted to be heterogeneous in a general way but responses of choices to external inputs are not. When heterogeneity in choices and outcomes is allowed, the IV enterprise breaks down. Treatment parameters can still be defined but IV does not identify them.

Section 5 extends our analysis to consider regression discontinuity estimators introduced in Campbell (1969) and adapted to modern econometrics in Hahn, Todd and Van der Klaauw (2001). We interpret the regression discontinuity estimator within the MTE framework, as a special type of IV estimator. In Section 6, we show how the output of the IV analysis of Section 4 can be used to extend parameters identified in one population to other populations and to forecast the effects of new programs. These are questions P-2 and P-3 introduced in Chapter 70. Sections 2–5 focus solely on the problem of internal validity, which is the problem defined as P-1. We also develop a cost benefit analysis based on the MTE and we analyze marginal policy changes. In Section 7, we generalize the analysis of instrumental variables to consider models with multiple outcomes. We develop both unordered and ordered choice models linking them to an explicit choice-theoretic literature.

In Section 8, we consider matching as a special case of our framework. Matching applied to estimating conditional means is a version of nonparametric least squares. It assumes that marginal and average returns are the same whereas our general framework allows us to distinguish marginal from average returns and to identify both. Matching is more robust than IV to violations of conventional monotonicity assumptions but the price for this robustness is steep in terms of its economic content. In Section 9, we develop randomization as an instrumental variable. We consider problems with compliance induced by agent self-selection decisions. In Section 10, we consider how to bound the various treatment parameters when models are not identified. Section 11 develops alternative methods for controlling for selection: control functions, replacement functions and proxy variables. Section 12 concludes.

2. The basic principles underlying the identification of the major econometric evaluation estimators

In this section, we review the main principles underlying the major evaluation estimators used in the econometric literature. We assume two potential outcomes (Y_0, Y_1). Models for multiple outcomes are developed in later sections of this chapter. As in Chapter 70, $D = 1$ if Y_1 is observed, and $D = 0$ corresponds to Y_0 being observed. The observed objective outcome is

$$Y = DY_1 + (1 - D)Y_0. \tag{2.1}$$

To briefly recapitulate the lessons of Chapter 70, we distinguish two distinct econometric problems. For simplicity, we focus our discussion on identification of objective outcomes. A parallel analysis can be made for subjective outcomes.

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 or quantile outcomes or various distributions of outcomes as described in Chapter 70. For example, a common approach is to focus attention on average treatment effects, such as $\text{ATE} = E(Y_1 - Y_0)$.

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

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

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

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

The *selection problem* is a key aspect of the problem of evaluating social programs. Many methods have been proposed to solve both problems. This chapter unifies these methods using the concept of the marginal treatment effect (MTE) introduced in Chapter 70 of this Handbook.

The method with the greatest intuitive appeal, which is sometimes called the "gold standard" in evaluation analysis, is the method of random assignment. Nonexperimental 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) $(Y_0, Y_1) \perp\!\!\!\perp D$.

It is useful to distinguish several cases where (R-1) will be satisfied. The first is that agents (decision makers whose choices are being investigated) 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 $D = 1$. This assumes that persons randomized out of the program cannot participate in it. 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 and the parameters treatment on the treated (TT) and treatment on the untreated (TUT) as defined in Chapter 70:

$$\text{TT} = \text{MTE} = \text{TUT} = \text{ATE} = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

Observe that even with random assignment of treatment status and full compliance, we cannot, in general, identify the distribution of the treatment effects $(Y_1 - Y_0)$, although we can identify the marginal distributions $F_1(Y_1 \mid A = 1, X = x) = F_1(Y_1 \mid X = x)$ and $F_0(Y_0 \mid A = 0, X = x) = F_0(Y_0 \mid X = x)$. One special assumption, common in the conventional econometrics literature, is that $Y_1 - Y_0 = \Delta(x)$, a constant given x. Since $\Delta(x)$ can be identified from $E(Y_1 \mid A = 1, X = x) - E(Y_0 \mid A = 0, X = x)$ because A is allocated by randomization, the analyst can identify the joint distribution of (Y_0, Y_1).[1] However, this approach assumes that (Y_0, Y_1) have the same distribution up to a parameter Δ (Y_0 and Y_1 are perfectly dependent). One can make other assumptions about the dependence across ranks from perfect positive or negative ranking to independence.[2] In general, 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 ($\Pr(Y_1 \geqslant Y_0)$). This problem plagues all evaluation methods. Abbring and Heckman discuss methods for identifying joint distributions of outcomes in Chapter 72.

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 equations model, assumption (M-1) that D is independent from (Y_0, Y_1) given X justifies application of least squares on D to eliminate selection bias in mean outcome

[1] Heckman (1992), Heckman, Smith and Clements (1997).

[2] Heckman, Smith and Clements (1997).

parameters. For means, matching is just nonparametric regression.[3] In order to be able to compare X-comparable people, we must assume

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

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 $X = x$ and there is full compliance.

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

$$\begin{aligned} F_1(Y_1 \mid D = 1, X = x) &= F_1(Y_1 \mid X = x) \\ &= F_1(Y_1 \mid D = 0, X = x). \end{aligned}$$

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

$$\begin{aligned} F_0(Y_0 \mid D = 0, X = x) &= F_0(Y_0 \mid X = x) \\ &= F_0(Y_0 \mid D = 1, X = x), \end{aligned}$$

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.

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

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

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. We can define the residual variation in D not accounted for by X as

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

If the variance of $\mathcal{E}(x)$ is zero, it is not possible to construct contrasts in 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

[3] See the discussion in Section 8. Barnow, Cain and Goldberger (1980) present one application of matching in a regression setting.

expectations, under (M-1), to obtain

$$E(Y \mid X, D) = E(Y_0 \mid X) + D\big[E(Y_1 - Y_0 \mid X)\big].[4]$$

If $\mathrm{Var}(\mathcal{E}(x)) > 0$ for all x in the support of X, one can use nonparametric least squares to identify $E(Y_1 - Y_0 \mid X = x) = \mathrm{ATE}(x)$ by regressing Y on D and X. The function identified from the coefficient on D is the average treatment effect.[5] If $\mathrm{Var}(\mathcal{E}(x)) = 0$, $\mathrm{ATE}(x)$ 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 we write

$$Y_0 = X\alpha + U, \qquad Y_1 = X\alpha + \beta + U,$$

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

$$E(Y \mid X, D) = X\alpha + D\beta + E(U \mid X).$$

If D is perfectly predictable by X, we cannot identify β because of a multicollinearity problem. (M-2) rules out perfect collinearity.[6] 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). However, matching does not assume exogeneity of X.

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.[7] 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). We 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

[4] This follows because $E(Y \mid X, D) = E(Y_0 \mid X, D) + DE(Y_1 - Y_0 \mid X, D)$, but from (M-1), $E(Y_0 \mid X, D) = E(Y_0 \mid X)$ and $E(Y_1 - Y_0 \mid X, D) = E(Y_1 - Y_0 \mid X)$.

[5] Under the conditional independence assumption (M-1), it is also the effect of treatment on the treated $E(Y_1 - Y_0 \mid X, D = 1)$.

[6] Clearly (M-1) and (M-2) are sufficient but not necessary conditions. For the special case of OLS, as a consequence of the assumed linearity in the functional form of the estimating equation, we achieve identification of β if $\mathrm{Cov}(X, U) = 0$, $\mathrm{Cov}(D, U) = 0$ and (D, X) are not perfectly collinear. Observe that (M-1) does not imply that $E(U \mid X) = 0$. Thus, we can identify β but not necessarily α.

[7] Heckman et al. (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.

versions of (M-1) and (M-2). Without invoking further assumptions, there is no objective principle for determining what conditioning variables produce (M-1).

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 we exposit them in this chapter.

To analyze the informational assumptions invoked in matching, and other econometric evaluation strategies, it is helpful to introduce five distinct information sets and establish some relationships among them.[8] (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. We will denote the random variables generated by these sets as I_{R^*}, I_R, I_A, I_{E^*}, and I_E, respectively.[9]

DEFINITION 1. We say that $\sigma(I_{R^*})$ is 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 that

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

DEFINITION 2. We say that $\sigma(I_R)$ is 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.[10]

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

[9] We start with a primitive probability space (Ω, σ, P) with associated random variables I. We 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.

[10] Observe 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

If we define a 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 possible, it is desirable to use the minimal relevant information set.

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, we call I_A the *agent's information*.

By the agent we 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 $\sigma(I_E) \subseteq \sigma(I_{E^*})$.[11] We 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.

The third restriction requires that the information used by the econometrician must be part of the information that the econometrician observes. 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 does not 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. Assuming a minimal relevant information set exists, matching assumption (M-1) implies that

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 transformations and affine transformations of I_R also preserve the property). In the more general case of nonnested 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.

[11] This 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. The requirement $\sigma(I_R) \subseteq \sigma(I_{R^*})$ is satisfied by nested sets.

$\sigma(I_R) \subseteq \sigma(I_E)$, so that the econometrician uses at least the minimal relevant 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. We present examples of the consequences of the asymmetry in agent and analyst information sets in Section 8.

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 matching. Those advocates assume that they know the X that produces a relevant information set. Heckman and Navarro (2004) 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) and we summarize their analysis in Section 8. Conditional independence condition (M-1) cannot be tested without maintaining other assumptions.[12] As noted in Chapter 70, choosing the appropriate conditioning variables is a problem that plagues *all* econometric estimators.

The methods of control functions, replacement functions, proxy variables and instrumental variables recognize the possibility of asymmetry in information between the agent being studied and the econometrician and further 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).[13] These methods postulate the existence of some unobservables θ, which may be vector-valued, with the property that

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

but allow for the possibility that

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

In the event (U-2) holds, these approaches model the relationship of the unobservable θ with (Y_0, Y_1) and D in various ways. The content in the 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. We present examples of models that satisfy (U-1) but not (U-2) in Section 8.

[12] We discuss the required "exogeneity" conditions in our discussion of matching in Section 8. 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). Abbring and Heckman (Chapter 72) discuss this case.

[13] The term and concept of control function is due to Heckman and Robb (1985a, 1985b, 1986a, 1986b). See Blundell and Powell (2003) who call the Heckman–Robb replacement functions control functions. A more recent nomenclature is "control variate". Matzkin (2007) (Chapter 73 in this Handbook) provides a comprehensive discussion of identification principles for these, and other, econometric estimators.

The early literature focused on mean outcomes conditional on covariates [Heckman and Robb (1985a, 1985b, 1986a, 1986b)] 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 (2003)]. Abbring and Heckman review this work in Chapter 72.

The normal Roy model discussed in Chapter 70 makes distributional assumptions and identifies the joint distribution of outcomes. (Recall the discussion in Section 6.1 of Chapter 70.) A large literature surveyed in Chapter 73 (Matzkin) of this Handbook 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 using observables.[14] Aakvik, Heckman and Vytlacil (1999, 2005), Carneiro, Hansen and Heckman (2001, 2003), Cunha, Heckman and Navarro (2005), and Cunha, Heckman and Schennach (2006b, 2007) 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 methods are discussed in Chapter 72.

The normal selection model discussed in Section 6.1 of Chapter 70 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, and 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 Chapter 73 (Matzkin) of this Handbook survey methods for identifying semiparametric versions of these models. Appendix B of Chapter 70 presents 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 heterogenous responses is the probability of selection. Let Z denote variables in the choice equation. Fixing Z at different values (denoted z), we 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 the separable index model introduced in Chapter 70, for a fixed value of z,

$$D(z) = \mathbf{1}\big(\mu_D(z) \geqslant V\big),$$

where $Z \perp\!\!\!\perp V \mid X$. Thus fixing $Z = z$, values of z do not affect the realizations of V 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 $D(z) \perp\!\!\!\perp Z \mid X$ for all $z \in \mathcal{Z}$,

[14] This is the "control variate" of Blundell and Powell (2003). Heckman and Robb (1985a) and Olley and Pakes (1996) use a similar idea. Chapter 73 (Matzkin) of this Handbook discusses replacement functions.

where $\mathcal{Z}$ is the support of Z. The Imbens–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 independent of Z.

The method of instrumental variables (IV) postulates that

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

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

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

Alternatively, we can write that $\mathrm{Var}(E(D \mid X, Z)) \neq \mathrm{Var}(E(D \mid X))$.

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 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. We exposit the implicit model of the relationship between D and (Y_0, Y_1) used in instrumental variables in this chapter.

(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 we allow for $\mathrm{Cov}(D, U) \neq 0$, (IV-1) and (IV-2) identify β.[15] When β varies in the population and is correlated with D, additional assumptions must be invoked for IV to identify interpretable parameters. We discuss these conditions in Section 4 of this chapter, drawing on and extending the analysis of Heckman and Vytlacil (1999, 2001b, 2005) and Heckman, Urzua and Vytlacil (2006).

Assumptions (IV-1) and (IV-2), with additional assumptions in the case where β varies in the population which we discuss in this chapter, can be used to identify mean treatment parameters. Replacing Y_1 with $\mathbf{1}(Y_1 \leqslant t)$ and Y_0 with $\mathbf{1}(Y_0 \leqslant t)$, where t is a constant, the IV approach allows us to identify marginal distributions $F_1(y_1 \mid X)$ or $F_0(y_0 \mid X)$.

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 pro-

[15] $\beta = \frac{\mathrm{Cov}(Z,Y)}{\mathrm{Cov}(Z,D)}$.

vide 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 $\mathcal{E}(X) = D - E(D \mid X)$ that is nondegenerate and is one manifestation of the randomness that causes persons to switch status.[16]

In the IV method, it is the choice probability $E(D \mid X, Z) = P(X, Z)$ 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:

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

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 in (U-1) and (U-2) 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)].

Table 1 summarizes some of the main lessons of this section. We stress that 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 more fully and we exposit them in this Handbook. We start with the method of instrumental variables and analyze the general case where responses to treatment are heterogeneous and persons select into treatment status in response to the heterogeneity in treatment response.

Our strategy in this chapter is to anchor all of our analysis around the economic theory of choice as embodied in discrete choice theory and versions of the generalized Roy model developed in Chapter 70. We next show how recent developments allow analysts to define treatment parameters within a well-posed economic framework but without the strong assumptions maintained in the early literature on selection models. To focus our discussion, we first consider the analysis of a prototypical policy evaluation program.

2.1. *A prototypical policy evaluation problem*

To motivate our discussion in this chapter, 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

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

Table 1
Identifying assumptions under commonly used methods

	Identifying assumptions	Identifies marginal distributions?	Exclusion condition needed?
Random assignment	$(Y_0, Y_1) \perp\!\!\!\perp \xi$, $\xi = 1 \Rightarrow A = 1, \xi = 0 \Rightarrow A = 0$ (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. So 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 $(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta$. The method models dependence induced by θ or else proxies θ (replacement function). Version (i). Replacement functions (substitute out θ by observables) [Blundell and Powell (2003), Heckman and Robb (1985a), 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 out 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) No (under some parametric assumptions)
IV	$(Y_0, Y_1) \not\!\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp Z \mid X$, $\Pr(D = 1 \mid Z)$ is a nondegenerate function of Z.	Yes	Yes

Notes: (Y_0, Y_1) are potential outcomes that depend on X;

$$D = \begin{cases} 1 & \text{if assigned (or choose) status 1,} \\ 0 & \text{otherwise;} \end{cases}$$

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.

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. We can make a similar decomposition for Y_0:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1, \\ Y_0 &= \mu_0(X) + U_0. \end{aligned} \tag{2.2}$$

We do not need to assume additive separability but it is convenient and we 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. We develop more general nonseparable models in later sections of this chapter.

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 treatment parameters introduced in Chapter 70. We use a special form of the generalized Roy model with constant cost C of adopting the policy. This is called the "extended Roy model". We 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 we could estimate the distribution of $Y_1 - Y_0$ given X and there is research that does this. Abbring and Heckman survey methods for doing so in Chapter 72.

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 (2.2) into a regression model, it is conventional to use the switching regression framework.[17] Define $D = 1$ if a country adopts a policy; $D = 0$ if it does not. The observed outcome Y is the switching regression model (2.1). Substituting (2.2) into this expression, and keeping all X implicit, we obtain

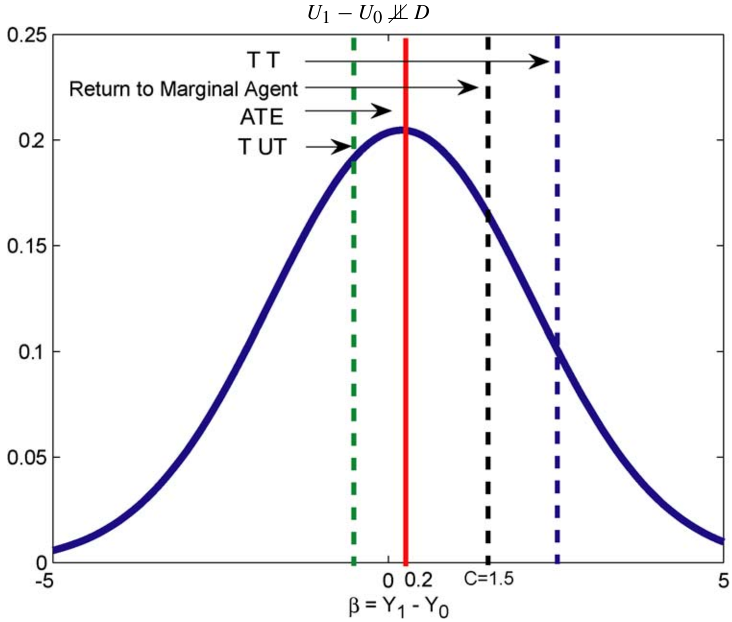
$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned} \tag{2.3}$$

Using conventional regression notation,

$$Y = \alpha + \beta D + \varepsilon, \tag{2.4}$$

where $\alpha = \mu_0$, $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$ and $\varepsilon = U_0$. We will also use the notation that $\eta = U_1 - U_0$, letting $\bar{\beta} = \mu_1 - \mu_0$ and $\beta = \bar{\beta} + \eta$. Throughout this section we 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 ($\beta(X)$)

[17] Statisticians 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 our discussion of this basic model of counterfactuals in Chapter 70.



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
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{if } D^* < 0 \end{cases}$

General case

$(U_1 - U_0) \not\perp D$
ATE $\neq$ TT $\neq$ TUT

The researcher observes (Y, D, C).
$Y = \alpha + \beta D + U_0$ where $\beta = Y_1 - Y_0$.

Parameterization

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

Figure 1. Distribution of gains in the Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

and estimates interactions of D with X. The case where β varies even after accounting for X is called the "random coefficient" or "heterogenous treatment effect" case. The case where $\eta = U_1 - U_0$ depends on D is the case of essential heterogeneity analyzed by Heckman, Urzua and Vytlacil (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 we develop the implications of this model in this chapter.

3. An index model of choice and treatment effects: Definitions and unifying principles

We now present the model of treatment effects developed in Heckman and Vytlacil (1999, 2001b, 2005) and Heckman, Urzua and Vytlacil (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 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. We follow Heckman and Vytlacil (1999, 2005) and Heckman, Urzua and Vytlacil (2006) in considering binary treatments. We analyze multiple treatments in Section 7. Florens et al. (2002) develop a model with a continuum of treatments and we briefly survey that work at the end of Section 7.

Y is the measured outcome variable. It is produced from the switching regression model (2.1). Outcomes are general nonlinear, nonseparable functions of observables and unobservables:

$$Y_1 = \mu_1(X, U_1), \tag{3.1}$$

$$Y_0 = \mu_0(X, U_0). \tag{3.2}$$

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: $Y_i = \mathbf{1}(Y_i^* \geqslant 0)$, where $Y_i^* = \mu_i(X) + U_i$, $i = 0, 1$. Notice that in the general case, $\mu_i(X, U_i) - E(Y_i \mid X) \neq U_i$, $i = 0, 1$.

As defined in Chapter 70, 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, we characterize the decision rule for program participation by an index model:

$$D^* = \mu_D(Z) - V, \qquad D = 1 \quad \text{if } D^* \geqslant 0, \qquad D = 0 \quad \text{otherwise}, \tag{3.3}$$

where, from the point of view of the econometrician, (Z, X) is observed and (U_0, U_1, V) is unobserved. The random variable V may be a function of (U_0, U_1). 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 (recall the discussion in Section 3.3 of Chapter 70). Without loss of generality, we define Z so that it includes all of the elements of X as well as any additional variables unique to the choice equation.

We invoke the following assumptions that are weaker than those used in the conventional literature on structural econometrics or the recent literature on semiparametric selection models and at the same time can be used both to define and to identify different treatment parameters.[18] The assumptions are:

(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*[19];
(A-4) *the values of* $E(|Y_1|)$ *and* $E(|Y_0|)$ *are finite* (*Finite means*);
(A-5) $0 < \Pr(D = 1 \mid X) < 1$.

(A-1) assumes that V is independent of Z given X, and is used below to generate counterfactuals. For the definition of treatment effects, we do not need either (A-1) or (A-2). Our definitions of treatment effects and their unification through MTE do not require any elements of Z that are not elements of X or independence assumptions. However, our analysis of instrumental variables requires that Z contain 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 so we 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.*)

[18] A much weaker set of conditions is required to define the parameters than is required to identify them. See the discussion in Appendix B. As noted in Section 6, stronger conditions are required for policy forecasting.

[19] Absolutely continuous with respect to Lebesgue measure.

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. Until we discuss the problems of external validity and policy forecasting in Section 6, we analyze treatment effects conditional on X, and maintain assumption (A-6).

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.[20] We sometimes denote $P(Z)$ by P, suppressing the Z argument. We also work with U_D, a uniform random variable ($U_D \sim \mathrm{Unif}[0, 1]$) defined by $U_D = F_{V|X}(V)$.[21] 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 chapter.

Vytlacil (2002) establishes that assumptions (A-1)–(A-5) for selection model (2.1) and (3.1)–(3.3) are equivalent to the assumptions used to generate the LATE model of Imbens and Angrist (1994) which are developed below in Section 4. Thus the nonparametric selection model for treatment effects developed by Heckman and Vytlacil is implied by the assumptions of the Imbens–Angrist instrumental variable model for treatment effects. Our approach links the IV literature to the literature on economic choice models exposited in Chapter 70. Our latent variable model is a version of the standard sample selection bias model. We weave together two strands of the literature often thought to be distinct [see, e.g., Angrist and Krueger (1999)].

The model of Equations (3.1)–(3.3) 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 $j = 0, 1$,

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

Z (given X) enters the model only through the propensity score $P(Z)$.[22] 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 $E(Y(1 - D) \mid$

[20] Throughout this chapter, we 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 $F_{A|B}(\cdot)$. We will write the cumulative distribution function of A conditional on $B = b$ by $F_{A|B}(\cdot \mid b)$.

[21] This 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.

[22] The set $\mathcal{A}$ is assumed to be measurable.

$X = x, P = p$). Heckman and Vytlacil (2005, Appendix A) develop this condition further, and show that it is testable.

Even though the model of treatment effects we exposit 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. We compare the assumptions used to identify IV with the assumptions used in matching in Section 8.

3.1. Definitions of treatment effects in the two outcome model

As developed in Chapter 70, 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.[23] The most commonly 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.[24] 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 $P(Z)$: $\Delta^{\text{TT}}(x, p) \equiv E(\Delta \mid X = x, P(Z) = p, D = 1)$.[25]

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 this chapter:

$$\Delta^{\text{MTE}}(x, u_D) \equiv E(\Delta \mid X = x, U_D = u_D). \tag{3.4}$$

This parameter is defined independently of any instrument. We 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_0 and Y_1 are value outcomes, MTE is a mean willingness-to-pay measure. MTE is a

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

[24] See, e.g., Imbens (2004).

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

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 Chapter 73 (Matzkin) of this Handbook, or "control function" interpretation in the sense of Blundell and Powell (2003). These three interpretations are equivalent under separability in D^*, i.e., when (3.3) characterizes the choice equation, but lead to three different definitions of MTE when a more general nonseparable model is developed. This point is developed in Section 4.10 where we discuss a general nonseparable model. The additive separability of Equation (3.3) 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. We present their full conditions for identification in Section 4. Here we define it in the notation used in this chapter. LATE is defined by an instrument in their analysis. As in Chapter 70, we define LATE independently of any instrument after first presenting the Imbens–Angrist definition. Define $D(z)$ as a counterfactual choice variable, with $D(z) = 1$ if state 1 ($D = 1$) 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, developed in Chapter 70 and defined in the introduction to this section, Heckman and Vytlacil (1999, 2005) show that LATE can be written as

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

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 estimator. Their analysis conflates issues of definition of parameters with issues of identification. Our representation of LATE allows us 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 assump-

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: $\text{PRTE}(x) = E(Y_{a'} \mid X = x) - E(Y_a \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\omega_{\text{PRTE}}(x, u_D)\, du_D$ 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

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

Source: Heckman and Vytlacil (2005).

tions (A-1), (A-3), and (A-4), $\Delta^{\text{LATE}}(x, u_D, u'_D)$ is continuous in u_D and u'_D and $\lim_{u'_D \uparrow u_D} \Delta^{\text{LATE}}(x, u_D, u'_D) = \Delta^{\text{MTE}}(x, u_D)$.[26]

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. Appendix A presents the formal derivation of the parameters and associated weights and graphically illustrates the relationship between ATE and TT. There we establish that all treatment parameters may be expressed as weighted averages of the MTE:

$$\text{Treatment parameter } (j) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_j(x, u_D)\, du_D,$$

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. We analyze how negative weights for IV might arise in Section 4.

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

$$\omega_{\text{TT}}(x, u_D) = \frac{1 - F_{P|X}(u_D \mid x)}{\int_0^1 (1 - F_{P|X}(t \mid x))\, dt} = \frac{S_{P|X}(u_D \mid x)}{E(P(Z) \mid X = x)}, \tag{3.5}$$

[26] This 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.

Table 2B
Weights

$$\omega_{\mathrm{ATE}}(x, u_D) = 1$$

$$\omega_{\mathrm{TT}}(x, u_D) = \Big[\int_{u_D}^{1} f_{P|X}(p \mid X = x)\, dp\Big] \frac{1}{E(P|X=x)}$$

$$\omega_{\mathrm{TUT}}(x, u_D) = \Big[\int_{0}^{u_D} f_{P|X}(p \mid X = x)\, dp\Big] \frac{1}{E((1-P)|X=x)}$$

$$\omega_{\mathrm{PRTE}}(x, u_D) = \left[\frac{F_{P_{a'}|X}(u_D|x) - F_{P_a|X}(u_D|x)}{\Delta \overline{P}(x)}\right], \text{ where}$$

$$\Delta \overline{P}(x) = E(P_a \mid X = x) - E(P_{a'} \mid X = x)$$

$$\omega_{\mathrm{IV}}^{J}(x, u_D) = \Big[\int_{u_D}^{1} (J(Z) - E(J(Z) \mid X = x)) f_{J,P|X}(j, t \mid X = x)\, dt\, dj\Big] \frac{1}{\mathrm{Cov}(J(Z), D|X=x)}$$

$$\omega_{\mathrm{OLS}}(x, u_D) = 1 + \frac{E(U_1|X=x, U_D=u_D)\omega_1(x, u_D) - E(U_0|X=x, U_D=u_D)\omega_0(x, u_D)}{\Delta^{\mathrm{MTE}}(x, u_D)}$$

$$\omega_1(x, u_D) = \Big[\int_{u_D}^{1} f_{P|X}(p \mid X = x)\, dp\Big] \frac{1}{E(P|X=x)}$$

$$\omega_0(x, u_D) = \Big[\int_{0}^{u_D} f_{P|X}(p \mid X = x)\, dp\Big] \frac{1}{E((1-P)|X=x)}$$

Source: Heckman and Vytlacil (2005).

and $S_{P|X}(u_D \mid x)$ is $\Pr(P(Z) > u_D \mid X = x)$ and $\omega_{\mathrm{TT}}(x, u_D)$ is a weighted distribution. The parameter $\Delta^{\mathrm{TT}}(x)$ oversamples $\Delta^{\mathrm{MTE}}(x, u_D)$ 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 least likely to participate. The various weights are displayed in Table 2B. The other weights, treatment effects and estimands shown in this table are discussed later. A central theme of this chapter is that under our assumptions all estimators and estimands can be written as weighted averages of MTE. This allows us to unify the treatment effect literature using a common functional $\Delta^{\mathrm{MTE}}(x, u_D)$.

Observe that if $E(Y_1 - Y_0 \mid X = x, U_D = u_D) = E(Y_1 - Y_0 \mid X = x)$, so $\Delta = Y_1 - Y_0$ is mean independent of U_D given $X = x$, then $\Delta^{\mathrm{MTE}} = \Delta^{\mathrm{ATE}} = \Delta^{\mathrm{TT}} = \Delta^{\mathrm{LATE}}$. 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. This is an instance of the general model developed in Chapter 70, Section 5. The model allows for costs to vary in the population and is more general than the extended Roy model. We discuss the weights for IV depicted in Figure 2B in Section 4 and the weights for OLS in Section 8. 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

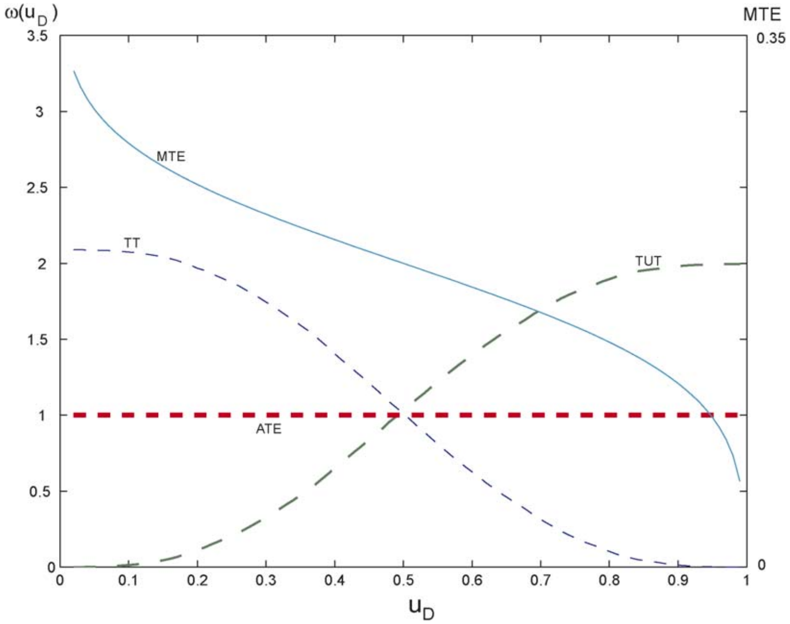
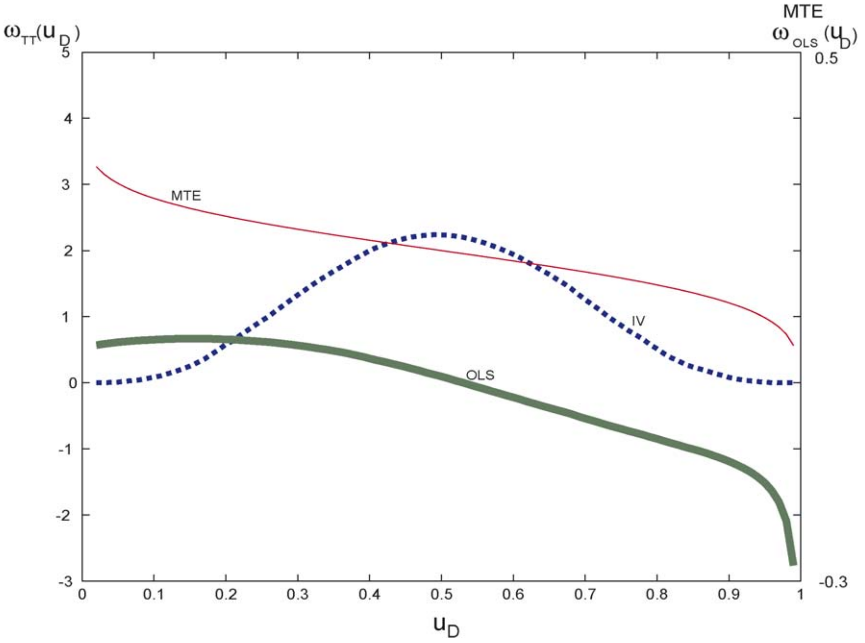


Figure 2A. Weights for the marginal treatment effect for different parameters. *Source*: Heckman and Vytlacil (2005).

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 ($E(Y_1 - Y_0 \mid X = x, D = 0)$), 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 TT $>$ ATE $>$ 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 $\geqslant$ ATE $\geqslant$ TT. Later in this chapter, we return to this table to discuss the other numbers in it.

Table 4 reproduced from Heckman (2001) presents evidence on the nonconstancy of the MTE in U_D drawn from a variety of studies of schooling, job training, migration and unionism. Most of the evidence is obtained using parametric normal selection models or variants of such models. With the exception of studies of unionism, a common finding



$Y_1 = \alpha + \bar{\beta} + U_1$	$U_1 = \sigma_1 \tau$	$\alpha = 0.67$	$\sigma_1 = 0.012$
$Y_0 = \alpha + U_0$	$U_0 = \sigma_0 \tau$	$\bar{\beta} = 0.2$	$\sigma_0 = -0.050$
$D = 1$ if $Z - V \geqslant 0$	$V = \sigma_V \tau$	$\tau \sim N(0, 1)$	$\sigma_V = -1.000$
	$U_D = \Phi(\frac{V}{\sigma_V \sigma_\tau})$		$Z \sim N(-0.0026, 0.2700)$

Figure 2B. Marginal treatment effect vs. linear instrumental variables and ordinary least squares weights. *Source*: Heckman and Vytlacil (2005).

in the empirical literature is the nonconstancy of MTE given X.[27] The evidence from the literature suggests that different treatment parameters measure different effects, and persons participate in programs based on heterogeneity in responses to the program being studied. The phenomenon of nonconstancy of the MTE that we analyze in this chapter is of substantial empirical interest.

The additively separable latent index model for D [Equation (3.3)] and assumptions (A-1)–(A-5) are far stronger than what is required to define the parameters in terms of the MTE. The representations of treatment effects defined in Table 2A remain valid even if Z is not independent of U_D, if there are no variables in Z that are not also contained in X, or if a more general nonseparable choice model generates D [so $D^* = \mu_D(Z, U_D)$]. An important advantage of our approach over other approaches to the analysis of instrumental variables in the recent literature is that no instrument Z is

[27] However, most of the empirical evidence is based on parametric selection models.

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
Policy relevant treatment effect (PRTE)	0.1549
Selection bias[b]	−0.0628
Linear instrumental variables[c]	0.2013
Ordinary least squares	0.1725

Source: Heckman and Vytlacil (2005).
Note: The model used to create Table 3 is the same as those used to create Figures 2A and 2B. The PRTE is computed using a policy t characterized as follows:

- If $Z > 0$ then $D = 1$ if $Z(1 + t) - V \geqslant 0$.
- If $Z \leqslant t$ then $D = 1$ if $Z - V \geqslant 0$.

For this example t is set equal to 0.2.

[a] $\text{TT} - \text{ATE} = E(Y_1 - Y_0 \mid D = 1) - E(Y_1 - Y_0)$.

[b] $\text{OLS} - \text{TT} = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$.

[c] Using propensity score $P(Z)$ as the instrument.

needed to define the parameters. We separate the tasks of definition and identification of parameters as discussed in Table 1 of Chapter 70, and present an analysis more closely rooted in economics. Appendices A and B define the treatment parameters for both separable (Appendix A) and nonseparable choice equations (Appendix B). We show that the treatment parameters can be defined even if there is no instrument or if instrumental variables methods break down as they do in nonseparable models.

As noted in Chapter 70, the literature on structural econometrics is clear about the basic parameters of interest although it is not always clear about the exact combinations of parameters needed to answer specific policy problems. The literature on treatment effects offers a variety of evaluation parameters. Missing from that literature is an algorithm for defining treatment effects that answer precisely formulated economic policy questions. The MTE provides a framework for developing such an algorithm. In the next section, we present one well defined policy parameter that can be used to generate Benthamite policy evaluations as discussed in Section 5 of Chapter 70.

3.2. *Policy relevant treatment parameters*

The conventional treatment parameters do not always answer economically interesting questions. Their link to cost-benefit analysis and interpretable economic frameworks is sometimes obscure. Each answers a different question. Many investigators estimate a treatment effect and hope that it answers an interesting question. A more promising approach for defining parameters is to postulate a policy question or decision problem

Table 4
Evidence on selection on unobservables and constancy of the MTE for separable models

Study	Method	Finding on the hypothesis of constancy of the MTE
	Unionism	
Lee (1978)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} = \sigma_{0V}$ Do not reject
Farber (1983)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} = \sigma_{0V}$ Do not reject
Duncan and Leigh (1985)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} = \sigma_{0V}$ Do not reject
Robinson (1989)	Normal selection model $(\mu_1 - \mu_0)_{\text{IV}} = (\mu_1 - \mu_0)_{\text{normal}}$	$\sigma_{1V} \neq \sigma_{0V}$ Do not reject
	Schooling (college vs. high school)	
Willis and Rosen (1979)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} \neq \sigma_{0V}$ Reject
Heckman, Tobias and Vytlacil (2003)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} \neq \sigma_{0V}$ Reject
	Job training	
Björklund and Moffitt (1987)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} \neq \sigma_{0V}$ Reject
Heckman et al. (1998; Suppl.)	$E(U_1 - U_0 \mid D = 1, Z, X)$ $= E(U_1 - U_0 \mid D = 1, X)$	Reject selection on unobservables
	Sectoral choice	
Heckman and Sedlacek (1990)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} \neq \sigma_{0V}$ Reject
	Migration	
Pessino (1991)	Normal selection model $(H_0\colon \sigma_{1V} = \sigma_{0V})$	$\sigma_{1V} \neq \sigma_{0V}$ Reject
Tunali (2000)	$H_0\colon\ E(U_1 - U_0 \mid D = 1) = 0$ (estimated using robust selection)	Cannot reject

Source: Heckman (2001).
Notes: $Y = DY_1 + (1 - D)Y_0$
$Y_1 = \mu_1(X) + U_1$
$Y_0 = \mu_0(X) + U_0$
$Z \perp\!\!\!\perp (U_0, U_1)$, $Z \not\perp\!\!\!\perp D$
$D = \mathbf{1}(\mu_D(Z) - V \geqslant 0)$, where $\mu_D(Z) - V$ is the index determining selection into “1” or “0”
Hypothesis: No selection on unobservables (constancy of the MTE)
$H_0\colon\ E(U_1 - U_0 \mid D = 1, Z, X)$ does not depend on D where $\operatorname{Cov}(U_1, U_V) = \sigma_{1V}$, $\operatorname{Cov}(U_0, U_V) = \sigma_{0V}$ (in normal model, the null hypothesis is $\sigma_{1V} = \sigma_{0V}$).

of interest and to derive the treatment parameter that answers it. Taking this approach does not in general produce the conventional treatment parameters or the estimands produced from instrumental variables.

Consider a class of policies that affect P, the probability of participation in a program, but do not affect Δ^{MTE}. The policies analyzed in the treatment effect literature that change the Z not in X are more restrictive than the general policies that shift X and Z analyzed in the structural literature. An example from the schooling literature would be policies that change tuition or distance to school but do not directly affect the gross returns to schooling [Card (2001)]. Since we ignore general equilibrium effects in this chapter, the effects on (Y_0, Y_1) from changes in the overall level of education are assumed to be negligible.

Let p and p' denote two potential policies and let D_p and $D_{p'}$ denote the choices that would be made under policies p and p'. When we discuss the policy relevant treatment effect, we use "p" to denote the policy and distinguish it from the realized value of $P(Z)$. Under our assumptions, the policies affect the Z given X, but not the potential outcomes. Let the corresponding decision rules be $D_p = \mathbf{1}[P_p(Z_p) \geqslant U_D]$, $D_{p'} = \mathbf{1}[P_{p'}(Z_{p'}) \geqslant U_D]$, where $P_p(Z_p) = \Pr(D_p = 1 \mid Z_p)$ and $P_{p'}(Z_{p'}) = \Pr(D_{p'} = 1 \mid Z_{p'})$. To simplify the exposition, we will suppress the arguments of these functions and write P_p and $P_{p'}$ for $P_p(Z_p)$ and $P_{p'}(Z_{p'})$. Define $(Y_{0,p}, Y_{1,p}, U_{D,p})$ as (Y_0, Y_1, U_D) under policy p, and define $(Y_{0,p'}, Y_{1,p'}, U_{D,p'})$ correspondingly under policy p'. We assume that Z_p and $Z_{p'}$ are independent of $(Y_{0,p}, Y_{1,p}, U_{D,p})$ and $(Y_{0,p'}, Y_{1,p'}, U_{D,p'})$, respectively, conditional on X_p and $X_{p'}$. Let $Y_p = D_p Y_{1,p} + (1 - D_p) Y_{0,p}$ and $Y_{p'} = D_{p'} Y_{1,p'} + (1 - D_{p'}) Y_{0,p'}$ denote the outcomes that would be observed under policies p and p', respectively.

Δ^{MTE} is policy invariant in the sense of Hurwicz as defined in Chapter 70 if

$E(Y_{1,p} \mid U_{D,p} = u_D, X_p = x)$ and $E(Y_{0,p} \mid U_{D,p} = u_D, X_p = x)$ are invariant to the choice of policy p (Policy invariance for the marginal treatment effect).

Policy invariance can be justified by the strong assumption that the policy being investigated does not change the counterfactual outcomes, covariates, or unobservables, i.e., $(Y_{0,p}, Y_{1,p}, X_p, U_{D,p}) = (Y_{0,p'}, Y_{1,p'}, X_{p'}, U_{D,p'})$. However, Δ^{MTE} is policy invariant if this assumption is relaxed to the weaker assumption that the policy change does not affect the distribution of these variables conditional on X:

(A-7) *The distribution of $(Y_{0,p}, Y_{1,p}, U_{D,p})$ conditional on $X_p = x$ is the same as the distribution of $(Y_{0,p'}, Y_{1,p'}, U_{D,p'})$ conditional on $X_{p'} = x$ (policy invariance for distribution).*

Assumption (A-7) guarantees that manipulations of the distribution of Z do not affect anything in the model except the choice of outcomes. These are specialized versions of (PI-3) and (PI-4) invoked in Chapter 70.

For the widely used Benthamite social welfare criterion $\Upsilon(Y)$, where Υ is a utility function, comparing policies using mean utilities of outcomes and considering the effect for individuals with a given level of $X = x$ we obtain the *policy relevant treatment effect*, PRTE, denoted $\Delta^{\text{PRTE}}(x)$:

$$\begin{aligned} &E\big(\Upsilon(Y_p) \mid X = x\big) - E\big(\Upsilon(Y_{p'}) \mid X = x\big) \\ &\quad = \int_0^1 \Delta_\Upsilon^{\text{MTE}}(x, u_D)\big\{F_{P_{p'}|X}(u_D \mid x) - F_{P_p|X}(u_D \mid x)\big\}\, du_D, \end{aligned} \tag{3.6}$$

where $F_{P_p|X}(\cdot \mid x)$ and $F_{P_{p'}|X}(\cdot \mid x)$ are the distributions of P_p and $P_{p'}$ conditional on $X = x$, respectively, defined for the different policy regimes and $\Delta_\Upsilon^{\text{MTE}}(x, u_D) = E(\Upsilon(Y_{1,p}) - \Upsilon(Y_{0,p}) \mid U_{D,p} = u_D, X_p = x)$.[28,29] The weights in expression (3.6) are derived in Appendix C under the assumption that the policy does not change the joint distribution of outcomes. To simplify the notation, throughout the rest of this chapter when we discuss PRTE, we assume that $\Upsilon(Y) = Y$. Modifications of our analysis for the more general case are straightforward. We also discuss the implications of noninvariance for the definition and interpretation of the PRTE in Appendix C.

Define $\Delta\bar{P}(x) = E(P_p \mid X = x) - E(P_{p'} \mid X = x)$, the change in the proportion of people induced into the program due to the intervention. Assuming $\Delta\bar{P}(x)$ is positive, we may define per person affected weights as

$$\omega_{\text{PRTE}}(x, u_D) = \frac{F_{P_{p'}|X}(u_D \mid x) - F_{P_p|X}(u_D \mid x)}{\Delta\bar{P}(x)}.$$

These weights are displayed in Table 2B. As demonstrated in the next section, in general, conventional IV weights the MTE differently than either the conventional treatment parameters (Δ^{ATE} or Δ^{TT}) or the policy relevant parameter, and so does not recover these parameters.

Instead of hoping that conventional treatment parameters or favorite estimators answer interesting economic questions, the approach developed by Heckman and Vytlacil (1999, 2001a, 2001b, 2005) is to estimate the MTE and weight it by the appropriate weight determined by how the policy changes the distribution of P to construct Δ^{PRTE}. In Heckman and Vytlacil (2005), we also develop an alternative approach that produces a policy weighted instrument to identify Δ^{PRTE} by standard instrumental variables. We elaborate our discussion of policy analysis based in the MTE and develop other policy

[28] We could define policy invariance for Δ^{MTE} in terms of expectations of $\Upsilon(Y_{1,p})$ and $\Upsilon(Y_{0,p})$.

[29] If we assume that the marginal distribution of X_p and $X_{p'}$ are the same as the marginal distribution of a benchmark X, the weights can be integrated against the distribution of X to obtain the total effect of the policy in the population:

$$\begin{aligned} &E\big(\Upsilon(Y_p)\big) - E\big(\Upsilon(Y_{p'})\big) \\ &\quad = E_X\big[E\big(\Upsilon(Y_p) \mid X\big) - E\big(\Upsilon(Y_{p'}) \mid X\big)\big] \\ &\quad = \int\bigg[\int_0^1 \Delta_\Upsilon^{\text{MTE}}(x, u_D)\big\{F_{P_{p'}|X}(u_D \mid x) - F_{P_p|X}(u_D \mid x)\big\}\, du_D\bigg]\, dF_X(x). \end{aligned}$$

parameters for local and global perturbations of policy in Section 6 after developing the instrumental variable estimator and the related regression discontinuity estimator. The analyses of Sections 4 and 5 give us tools to make specific the discussion of alternative approaches to policy evaluation.

4. Instrumental variables

The method of instrumental variables (IV) is currently the most widely used method in economics for estimating economic models when unobservables are present that violate the matching assumption (M-1).[30] We first present an intuitive exposition of the method and then present a more formal development. We analyze a model with two outcomes. We generalize the analysis to multiple outcomes in Section 7.

Return to the policy adoption example presented at the end of Section 2. The distribution of returns to adoption is depicted in Figure 1. First, consider the method of IV, where β (given X), which is the same as $Y_1 - Y_0$ given X, is the same for every country. This is the familiar case and we develop it first. The model is

$$Y = \alpha + \beta D + \varepsilon, \tag{4.1}$$

where conditioning on X is implicit. A simple least squares regression of Y on D (equivalently a mean difference in outcomes between countries with $D = 1$ and countries with $D = 0$) is possibly subject to a selection bias on Y_0. Countries that adopt the policy may be atypical in terms of their Y_0 ($= \alpha + \varepsilon$). Thus if countries that would have done well in terms of unobservable ε ($= U_0$) even in the absence of the policy are the ones that adopt the policy, β estimated from OLS (or its semiparametric version – matching) is upward biased because $\operatorname{Cov}(D, \varepsilon) > 0$.

If there is an instrument Z, with the properties that

$$\operatorname{Cov}(Z, D) \neq 0, \tag{4.2}$$

$$\operatorname{Cov}(Z, \varepsilon) = 0, \tag{4.3}$$

then standard IV identifies β, at least in large samples,

$$\operatorname{plim} \hat{\beta}_{\mathrm{IV}} = \frac{\operatorname{Cov}(Z, Y)}{\operatorname{Cov}(Z, D)} = \beta.$$[31]

If other instruments exist, each identifies β. Z produces a controlled variation in D relative to ε. Randomization of assignment with full compliance to experimental protocols

[30] More precisely, IV is the most widely used alternative to OLS. OLS is a version of matching that imposes linearity of the functional form of outcome equations and assumes exogeneity of the regressors. See our discussion of matching in Section 8.

[31] The proof is straightforward. Under general conditions [see, e.g., White (1984)],

$$\operatorname{plim} \hat{\beta}_{\mathrm{IV}} = \beta + \frac{\operatorname{Cov}(Z, \varepsilon)}{\operatorname{Cov}(Z, D)} \quad \text{and} \quad \operatorname{Cov}(Z, \varepsilon) = 0,$$

so the second term on the right-hand side vanishes.

is an example of an instrument. From the instrumental variable estimators, we can identify the effect of adopting the policy in any country since all countries respond to the policy in the same way controlling for their X.

If β $(= Y_1 - Y_0)$ varies in the population even after controlling for X, there is a distribution of responses that cannot in general be summarized by a single number. Even if we are interested in the mean of the distribution, a new phenomenon distinct from selection bias might arise. This is a problem of sorting on the gain, which is distinct from sorting on levels. If β varies, even after controlling for X, there may be sorting on the gain $(\mathrm{Cov}(\beta, D) \neq 0)$. This is the model of *essential heterogeneity* as defined by Heckman, Urzua and Vytlacil (2006). It is also called a correlated random coefficient model [Heckman and Vytlacil (1998)].

The application of instrumental variables to this case is more problematic. Suppose that we augment the standard instrumental variable assumptions (4.2) and (4.3) by the following assumption:

$$\mathrm{Cov}(Z, \beta) = 0. \tag{4.4}$$

Can we identify the mean of $(Y_1 - Y_0)$ using IV? In general we cannot.[32]

To see why, let $\bar{\beta} = (\mu_1 - \mu_0)$ be the mean treatment effect (the mean of the distribution in Figure 1). $\beta = \bar{\beta} + \eta$, where $U_1 - U_0 = \eta$ and $\bar{\beta} = \mu_1 - \mu_0$ and we keep the conditioning on X implicit. Write Equation (4.1) in terms of these parameters:

$$Y = \alpha + \bar{\beta} D + [\varepsilon + \eta D].$$

The error term of this equation $(\varepsilon + \eta D)$ contains two components. By assumption, Z is uncorrelated with ε and η. But to identify $\bar{\beta}$, we need IV to be uncorrelated with $[\varepsilon + \eta D]$. That requires Z to be uncorrelated with ηD.

If policy adoption is made without knowledge of η $(= U_1 - U_0)$, the idiosyncratic gain to policy adoption after controlling for the observables, then η and D are statistically independent and hence uncorrelated, and IV identifies $\bar{\beta}$.[33] If, however, policy adoption is made with partial or full knowledge of η, IV does not identify $\bar{\beta}$ because $E(\eta D \mid Z) = E(\eta \mid D = 1, Z) \Pr(D = 1 \mid Z)$ and if there is sorting on the unobserved gain η, the first term is not zero. Similar calculations show that IV does not identify the mean gain to the countries that adopt the policy $(E(\beta \mid D = 1))$ and many other summary treatment parameters.[34] Whether η $(= U_1 - U_0)$ is correlated with D depends on the quality of the data available to the empirical economist and cannot be settled

[32] This point was made by Heckman and Robb (1985a, 1986a). See also Heckman (1997).

[33] The proof is straightforward:

$$\operatorname{plim} \hat{\beta}_{\mathrm{IV}} = \bar{\beta} + \frac{\mathrm{Cov}(Z, \varepsilon + \eta D)}{\mathrm{Var}(D, Z)}.$$

But $\mathrm{Cov}(Z, \varepsilon + \eta D) = \mathrm{Cov}(Z, \varepsilon) + \mathrm{Cov}(Z, \eta D)$ and $\mathrm{Cov}(Z, \eta D) = E(Z\eta D) - E(Z)E(\eta D)$, $E(\eta D) = 0$ by the assumed independence. $E(Z\eta D) = E[E(\eta DZ \mid Z)] = E[E(\eta D \mid Z)Z] = 0$ since $E(\eta D \mid Z) = 0$.

[34] See Heckman and Robb (1985a, 1986a), Heckman (1997) or Heckman and Vytlacil (1999).

a priori. The conservative position is to allow for such a correlation. However, this rules out IV as an interesting econometric strategy for identifying any of the familiar mean treatment parameters.

In light of the negative conclusions about IV in the literature preceding their paper, it is remarkable that Imbens and Angrist (1994) establish that under certain conditions, in the model with essential heterogeneity, IV can identify an interpretable parameter. The parameter they identify is a discrete approximation to the marginal gain parameter introduced by Björklund and Moffitt (1987). The Björklund–Moffitt parameter is a version of MTE for a parametric normal selection model. We derive their parameter from a selection model in Section 4.8. Björklund and Moffitt (1987) demonstrate how to use a selection model to identify the marginal gain to persons induced into a treatment status by a marginal change in the cost of treatment. Imbens and Angrist (1994) show how to estimate a discrete approximation to the Björklund–Moffitt parameter using instrumental variables.

Imbens and Angrist (1994) assume the existence of an instrument Z that takes two or more distinct values. This is implicit in (4.2). If Z assumes only one value, the covariance in (4.2) would be zero. Strengthening the covariance conditions of Equations (4.3) and (4.4), they assume (IV-1) and (IV-2) (independence and rank, respectively) and that Z is independent of $\beta = (Y_1 - Y_0)$ and Y_0. Recall that we denote by $D(z)$ the random variable indicating receipt of treatment when Z is set to z. ($D(z) = 1$ if treatment is received; $D(z) = 0$ otherwise.) The Imbens–Angrist independence and rank assumptions are (IV-1) and (IV-2).

They supplement the standard IV assumptions with what they call a "monotonicity" assumption. It is a condition across persons. The assumption maintains that if Z is fixed first at one and then at the other of two distinct values, say $Z = z$ and $Z = z'$, then all persons respond in their choice of D to the change in Z in the same way. In our policy adoption example, this condition states that a movement from z to z', causes all countries to move toward (or against) adoption of the public policy being studied. If some adopt, others do not drop the policy in response to the same change.

More formally, letting $D_i(z)$ be the indicator ($= 1$ if adopted; $= 0$ if not) for adoption of a policy if $Z = z$ for country i, then for any distinct values z and z' Imbens and Angrist (1994) assume:

(IV-3) $D_i(z) \geqslant D_i(z')$ *for all* i, *or* $D_i(z) \leqslant D_i(z')$ *for all* $i = 1, \ldots, I$ (*Monotonicity or uniformity*).

The content in this assumption is not in the order for any person. Rather, the responses have to be uniform across people for a given choice of z and z'. One possibility allowed under (IV-3) is the existence of three values of $z < z' < z''$ such that for all i, $D_i(z) \geqslant D_i(z')$ but $D_i(z') \leqslant D_i(z'')$. The standard usage of the term monotonicity rules out this possibility by requiring that one of the following hold for all i: (a) $z < z'$ componentwise implies $D_i(z) \geqslant D_i(z')$ or (b) $z < z'$ componentwise implies $D_i(z) \leqslant D_i(z')$. Of course, if the $D_i(z)$ are monotonic in Z in the same direction for all i, they are monotonic in the sense of Imbens and Angrist.

For any value of z' in the domain of definition of Z, from (IV-1) and (IV-2) and the definition of $D(z)$, $(Y_0, Y_1, D(z'))$ is independent of Z. For any two values of the instrument $Z = z$ and $Z = z'$, we may write

$$\begin{aligned} &E(Y \mid Z = z) - E(Y \mid Z = z') \\ &\quad = E\big(Y_1 D + Y_0(1 - D) \bigm| Z = z\big) - E\big(Y_1 D + Y_0(1 - D) \bigm| Z = z'\big) \\ &\quad = E\big(Y_0 + D(Y_1 - Y_0) \bigm| Z = z\big) - E\big(Y_0 + D(Y_1 - Y_0) \bigm| Z = z'\big). \end{aligned}$$

From the independence condition (IV-1) and the definition of $D(z)$ and $D(z')$, we may write this expression as $E[(Y_1 - Y_0)(D(z) - D(z'))]$. Using the law of iterated expectations,

$$\begin{aligned} &E(Y \mid Z = z) - E(Y \mid Z = z') \\ &\quad = E\big(Y_1 - Y_0 \bigm| D(z) - D(z') = 1\big) \Pr\big(D(z) - D(z') = 1\big) \\ &\qquad - E\big(Y_1 - Y_0 \bigm| D(z) - D(z') = -1\big) \Pr\big(D(z) - D(z') = -1\big). \end{aligned} \tag{4.5}$$

By the monotonicity condition (IV-3), we eliminate one or the other term in the final expression. Suppose that $\Pr(D(z) - D(z') = -1) = 0$, then

$$\begin{aligned} &E(Y \mid Z = z) - E(Y \mid Z = z') \\ &\quad = E\big(Y_1 - Y_0 \bigm| D(z) - D(z') = 1\big) \Pr\big(D(z) - D(z') = 1\big). \end{aligned}$$

Observe that, by monotonicity, $\Pr(D(z) - D(z') = 1) = \Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')$. For values of z and z' that produce distinct propensity scores $\Pr(D = 1 \mid Z = z)$, using monotonicity once more, we obtain LATE:

$$\begin{aligned} \text{LATE} &= \frac{E(Y \mid Z = z) - E(Y \mid Z = z')}{\Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')} \\ &= E\big(Y_1 - Y_0 \bigm| D(z) - D(z') = 1\big).^{35} \end{aligned} \tag{4.6}$$

This is the mean gain to those induced to switch from "0" to "1" by a change in Z from z' to z.

This is not the mean of $Y_1 - Y_0$ (average treatment effect) unless the Z assume values (z, z') such that $\Pr(D(z) = 1) = 1$ and $\Pr(D(z') = 1) = 0$.[36] It is also not the effect of treatment on the treated $(E(Y_1 - Y_0 \mid D = 1) = E(\beta \mid D = 1))$ unless the analyst has access to one or more values of Z such that $\Pr(D(z) = 1) = 1$.

The LATE parameter is defined by a hypothetical manipulation of instruments. It depends on the particular instrument used.[37] If monotonicity (uniformity) is violated,

[35] $\Pr(D(z) - D(z') = 1) = \Pr(D(z = 1) \wedge D(z' = 0)) = \Pr(D(z) = 1) - \Pr(D(z') = 1)$ from monotonicity.

[36] Such values produce "identification at infinity" or more accurately limit points where $P(z) = 1$ and $P(z') = 0$.

[37] Dependence of the estimands on the choices of IV used to estimate models with essential heterogeneity was first noted in Heckman and Robb (1985a, 1986a).

IV estimates an average response of those induced to switch into the program and those induced to switch out of the program by the change in the instrument because both terms in (4.5) are present.[38]

In an application to wage equations, Card (1999, 2001) interprets the LATE estimator as identifying returns to marginal persons. Heckman (1996) notes that the actual margin of choice selected by the IV estimator is not identified by the instrument. It is unclear as to which segment of the population the return estimated by LATE applies.

If the analyst is interested in knowing the average response $(\bar{\beta})$, the effect of the policy on the outcomes of countries that adopt it $(E(\beta \mid D = 1))$ or the effect of the policy if a particular country adopts it, there is no guarantee that the IV estimator comes any closer to the desired target than the OLS estimator and indeed it may be more biased than OLS. Because different instruments define different parameters, having a wealth of different strong instruments does not improve the precision of the estimate of any particular parameter. This is in stark contrast with the traditional model with $\beta \perp\!\!\!\perp D$. In that case, all valid instruments identify $\bar{\beta}$. The Durbin (1954) – Wu (1973) – Hausman (1978) test for the validity of extra instruments applies to the traditional model. In the more general case with essential heterogeneity, because different instruments estimate different parameters, no clear inference emerges from such specification tests.

When there are more than two distinct values of Z, Imbens and Angrist draw on the analysis of Yitzhaki (1989), which was refined in Yitzhaki (1996) and Yitzhaki and Schechtman (2004), to produce a weighted average of pairwise LATE parameters where the scalars Z are ordered to define the LATE parameter. In this case, IV is a weighted average of LATE parameters with nonnegative weights.[39] Imbens and Angrist generalize this result to the case of vector Z assuming that instruments are monotonic functions of the probability of selection.

Heckman and Vytlacil (1999, 2001b, 2005), Heckman, Urzua and Vytlacil (2006) and Carneiro, Heckman and Vytlacil (2006) generalize the analysis of Imbens and Angrist (1994) in several ways and we report their results in this chapter. Using a choice-theoretic parameter (the marginal treatment effect or MTE) introduced into the literature on selection models by Björklund and Moffitt (1987), they relate the parameters estimated by IV to well formulated choice models. This allows treatment parameters to be defined independent of any values assumed by instruments. It is possible to generate all treatment effects as different weighted averages of the MTE. IV can also be interpreted

[38] Angrist, Imbens and Rubin (1996) consider the case of two way flows for the special case of a scalar instrument when the monotonicity assumption is violated. Their analysis is a version of Yitzhaki's (1989, 1996) analysis, which we summarize in Appendix D. He analyzes the net effect whereas they break the net effect into two components corresponding to the two gross flows that produce the two way flows.

[39] Yitzhaki (1989) shows for a scalar instrument that two stage least squares estimators of Y on $P(Z) = E(D \mid Z)$ identify weighted averages of terms like the second terms in (4.6) with positive weights. See also Yitzhaki (1996) and Yitzhaki and Schechtman (2004). We discuss this work in greater detail in Section 4.3.1, and we derive his weights in Appendix D. The original Yitzhaki (1989) paper is posted at the website of Heckman, Urzua and Vytlacil (2006).

as a weighted average of MTE. Different instruments weight different segments of the MTE differently. Using the nonparametric generalized Roy model, MTE is a limit form of LATE. Using MTE, we overcome a problem that plagues the LATE literature. LATE estimates marginal returns at an unidentified margin (or intervals of margins). We show how to use the MTE to unify diverse instrumental variables estimates and to determine what margins (or intervals of margins) they identify. Instead of reporting a marginal return for unidentified persons, we show how to report marginal returns for all persons identified by their location on the scale of a latent variable that arises from a well defined choice model and is related to the propensity of persons to make the choice being studied. We can interpret the margins of choice identified by various instruments and place diverse instruments on a common interpretive footing.

Heckman and Vytlacil (1999, 2005) establish the central role of the propensity score ($\Pr(D = 1 \mid Z = z) = P(z)$) in both selection and IV models.[40] They show that with vector Z and a scalar instrument $J(Z)$ constructed from vector Z, the weights on LATE and MTE that are implicit in standard IV are not guaranteed to be nonnegative. Thus IV can be negative even though all pairwise LATEs and pointwise MTEs are positive. Thus the treatment effects for any pair of (z, z') can be positive but the IV can be negative. We present examples below. Certain instruments produce positive weights and avoid this particular interpretive problem. Our analysis generalizes the analyses of weights on treatment effects by Yitzhaki and Imbens–Angrist, who analyze a special case where all weights are positive.

We establish the special status of $P(z)$ as an instrument. It always produces nonnegative weights for MTE and LATE. It enables analysts to identify MTE or LATE. With knowledge of $P(z)$, and the MTE or LATE, we can decompose any IV estimate into identifiable MTEs (at points) or LATEs (over intervals) and identifiable weights on MTE (or LATE) where the weights can be constructed from data. The ability to decompose IV into interpretable components allows analysts to determine the response to treatment of persons at different levels of unobserved factors that determine treatment status.

We present a simple test for essential heterogeneity (β dependent on D) that allows analysts to determine whether or not they can avoid the complexities of the more general model with heterogeneity in response to treatments. In Section 7, we generalize the analysis of IV in the two-outcome model to a multiple outcome model, analyzing both ordered and unordered choice cases.[41] We also demonstrate the fundamental asymmetry in the recent IV literature for models with heterogeneous outcomes. Responses to treatment are permitted to be heterogeneous in a general way. Responses of choices to instruments are not. When heterogeneity in choice is allowed for in a general way, IV and local IV do not estimate parameters that can be interpreted as weighted averages of MTEs or LATEs. We now turn to an analysis of the two-outcome model.

[40] Rosenbaum and Rubin (1983) establish the control role of the propensity score in matching models.

[41] Angrist and Imbens (1995) consider an ordered choice case with instruments common across all choices. Heckman, Urzua and Vytlacil (2006) consider both common and choice-specific instruments for both ordered and unordered cases.

4.1. IV in choice models

A key contribution of the analysis of Heckman and Vytlacil is to adjoin choice equation (3.3) to the outcome equations (2.1), (3.1) and (3.2). A standard binary threshold cross model for D is $D = \mathbf{1}(D^* \geqslant 0)$, where $\mathbf{1}(\cdot)$ is an indicator ($\mathbf{1}(A) = 1$ if A is true, 0 otherwise). A familiar version of (3.3) sets $\mu_D(Z) = Z\gamma$ and writes

$$D^* = Z\gamma - V, \tag{4.7}$$

where $(V \perp\!\!\!\perp Z) \mid X$. ($V$ is independent of Z given X.) In this notation, the propensity score or choice probability is

$$P(z) = \Pr(D = 1 \mid Z = z) = \Pr(Z\gamma \geqslant V) = F_V(Z\gamma),$$

where F_V is the distribution of V which is assumed to be continuous. In terms of the generalized Roy model where C is the cost of participation in sector 1, $D = \mathbf{1}[Y_1 - Y_0 - C > 0]$. For a separable model in outcomes and in costs,

$$C = \mu_D(W) + U_C,$$

we have $Z = (X, W)$, $\mu_D(Z) = \mu_1(X) - \mu_0(X) - \mu_D(W)$, and $V = -(U_1 - U_0 - U_C)$. In constructing many of our examples, we work with a special version where $U_C = 0$. We call this version the extended Roy model.[42] It is the model used to produce Figure 1. Our analysis, however, applies to more general models, and we also offer examples of generalized Roy models, as we have in Figure 2 and Table 3.

In the case where β (given X) is a constant, under (IV-1) and (IV-2) it is not necessary to specify the choice model to identify β. In a general model with heterogenous responses, the specification of $P(z)$ and its relationship with the instrument play crucial roles. To see this, study the covariance between Z and ηD discussed in the introduction to this section.[43] By the law of iterated expectations, letting $\bar{Z}$ denote the mean of Z,

$$\begin{aligned}\operatorname{Cov}(Z, \eta D) &= E\big((Z - \bar{Z})D\eta\big)\\ &= E\big((Z - \bar{Z})\eta \bigm| D = 1\big)\Pr(D = 1)\\ &= E\big((Z - \bar{Z})\eta \bigm| Z\gamma > V\big)\Pr(Z\gamma \geqslant V).\end{aligned}$$

Thus, even if Z and η are independent, they are not independent conditional on $D = \mathbf{1}[Z\gamma \geqslant V]$ if $\eta = (U_1 - U_0)$ is dependent on V (i.e., if the decision maker has partial knowledge of η and acts on it). Selection models allow for this dependence [see Heckman and Robb (1985a, 1986a), Ahn and Powell (1993), and Powell (1994)]. Keeping X implicit, assuming that

$$(U_1, U_0, V) \perp\!\!\!\perp Z \tag{4.8}$$

[42] Recall that the generalized Roy model has $U_C \not\equiv 0$, whereas the extended Roy model sets $U_C = 0$.

[43] Recall that $\eta = U_1 - U_0$.

(alternatively, assuming that $(\varepsilon, \eta) \perp\!\!\!\perp Z$), we obtain

$$\begin{aligned} E(Y \mid D = 0, Z = z) &= E(Y_0 \mid D = 0, Z = z) \\ &= \alpha + E(U_0 \mid z\gamma < V), \end{aligned}$$

where α and possibly $E(U_0 \mid z\gamma < V)$ depend on X, which can be written as

$$E(Y \mid D = 0, Z = z) = \alpha + K_0\big(P(z)\big),$$

where the functional form of K_0 is produced from the distribution of (U_0, V).[44] Focusing on means, the conventional selection approach models the conditional mean dependence between (U_0, U_1) and V.

Similarly,

$$\begin{aligned} E(Y \mid D = 1, Z = z) &= E(Y_1 \mid D = 1, Z = z) \\ &= \alpha + \bar{\beta} + E(U_1 \mid z\gamma \geqslant V) \\ &= \alpha + \bar{\beta} + K_1\big(P(z)\big), \end{aligned}$$

where α, $\bar{\beta}$ and $K_1(P(z))$ may depend on X. $K_0(P(z))$ and $K_1(P(z))$ are control functions in the sense of Heckman and Robb (1985a, 1986a). The control functions expect out the unobservables θ that give rise to selection bias (see (U-1)). Under standard conditions developed in the literature, analysts can identify $\bar{\beta}$. Powell (1994) discusses semiparametric identification. Because we condition on $Z = z$ (or $P(z)$), correct specification of the Z plays an important role in econometric selection methods. This sensitivity to the full set of instruments in Z appears to be absent from the IV method.

If β is a constant (given X), or if η $(= \beta - \bar{\beta})$ is independent of V, only one instrument from vector Z needs to be used to identify the parameter. Missing or unused instruments play no role in identifying mean responses but may affect the efficiency of the IV estimators. In a model where β is variable and not independent of V, misspecification of Z plays an important role in interpreting what IV estimates analogous to its role in selection models. Misspecification of Z affects both approaches to identification. This is a new phenomenon in models with heterogenous β. We now review results from the recent literature on instrumental variables in the model with essential heterogeneity.

4.2. Instrumental variables and local instrumental variables

In this section, we use Δ^{MTE} defined in Section 3 for a general nonseparable model (3.1)–(3.3) to organize the literature on econometric evaluation estimators. In terms of our simple regression model,

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

[44] This representation is derived in Heckman (1980), Heckman and Robb (1985a, 1986a), Ahn and Powell (1993) and Powell (1994).

$$= E\big(\beta \bigm| X = x, V = F_V^{-1}(u_D)\big)$$
$$= \bar{\beta}(x) + E(\eta \mid X = x, V = v),$$

where $v = F_V^{-1}(u_D)$. We assume policy invariance in the sense of Hurwicz for mean parameters (assumption (A-7)). For simplicity, we suppress the a and a' subscripts that indicate specific policies. We focus primarily on instrumental variable estimators and review the method of local instrumental variables. Section 4.1 demonstrated in a simple but familiar case that well established intuitions about instrumental variable identification strategies break down when Δ^{MTE} is nonconstant in u_D given X ($\beta \not\perp\!\!\!\perp D \mid X$). We acquire the probability of selection $P(z)$ as a determinant of the IV covariance relationships.

Two sets of instrumental variable conditions are presented in the current literature for this more general case: those associated with conventional instrumental variable assumptions, which are implied by the assumption of "no selection on heterogenous gains", ($\beta \perp\!\!\!\perp D \mid X$) and those which permit selection on heterogeneous gains. Neither set of assumptions implies the other, nor does either identify the policy relevant treatment effect or other economically interpretable parameters in the general case. Each set of conditions identifies different treatment parameters.

In place of standard instrumental variables methods, Heckman and Vytlacil (1999, 2001b, 2005) advocate a new approach to estimating policy impacts by estimating Δ^{MTE} using local instrumental variables (LIV) to identify all of the treatment parameters from a generator Δ^{MTE} that can be weighted in different ways to answer different policy questions. For certain classes of policy interventions covered by assumption (A-7) and analyzed in Section 6, Δ^{MTE} possesses an invariance property analogous to the invariant parameters of traditional structural econometrics.

4.2.1. Conditions on the MTE that justify the application of conventional instrumental variables

In the general case where $\Delta^{\text{MTE}}(x, u_D)$ is nonconstant in u_D ($E(\beta \mid X = x, V = v)$ depends on v), IV does not in general estimate any of the treatment effects defined in Section 3. We consider a scalar instrument $J(Z)$ constructed from Z which may be vector-valued. We sometimes denote $J(Z)$ by J, leaving implicit that J is a function of Z. If Z is a vector, $J(Z)$ can be one coordinate of Z, say Z_1. We develop this particular case in presenting our examples.

The notation is sufficiently general to make $J(Z)$ a general function of Z. The standard conditions $J(Z) \perp\!\!\!\perp (U_0, U_1) \mid X$ and $\text{Cov}(J(Z), D \mid X) \neq 0$ corresponding to (IV-1) and (IV-2), respectively, do not, by themselves, imply that instrumental variables using $J(Z)$ as the instrument will identify conventional or policy relevant treatment effects. When responses to treatment are heterogenous, we must supplement the standard conditions to identify interpretable parameters. To link our analysis to conventional analyses of IV, we continue to invoke familiar-looking representations of additive separability of outcomes in terms of (U_0, U_1) so we invoke (2.2). This is not required. All derivations and results in this subsection hold without assuming additive separability if $\mu_1(x)$ and $\mu_0(x)$ are replaced by $E(Y_1 \mid X = x)$ and $E(Y_0 \mid X = x)$, respectively, and U_1 and U_0 are replaced by $Y_1 - E(Y_1 \mid X)$ and $Y_0 - E(Y_0 \mid X)$, respectively. This highlights the point that all of our analysis of IV is conditional on X and X need not be exogenous with respect to (U_0, U_1) to identify the MTE conditional on X. To simplify the notation, we keep the conditioning on X implicit unless it is useful to break it out separately.

Two distinct sets of instrumental variable conditions in the literature are those due to Heckman and Robb (1985a, 1986a) and Heckman (1997), and those due to Imbens and Angrist (1994) which we previously discussed. We review the conditions of Heckman and Robb (1985a, 1986a) and Heckman (1997) in Appendix L, which is presented in the context of our discussion of matching in Section 8, where we compare IV and matching. In the case where Δ^{MTE} is nonconstant in u_D, standard IV estimates different parameters depending on which assumptions are maintained. We have already shown that when responses to treatment are heterogeneous, and choices are made on the basis of this heterogeneity, standard IV does not identify $\mu_1 - \mu_0 = \bar{\beta}$.

There are two important cases of the variable response model. The first case arises when responses are heterogeneous, but conditional on X, people do not base their participation on these responses. In this case, keeping the conditioning on X implicit,

(C-1) $D \perp\!\!\!\perp \Delta \Rightarrow E(\Delta \mid U_D) = E(\Delta)$, $\Delta^{\text{MTE}}(u_D)$ *is constant in* u_D and $\Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}} = \Delta^{\text{LATE}}$, *i.e.*, $E(\beta \mid D = 1) = E(\beta)$, *because* $\beta \perp\!\!\!\perp D$.

In this case, all mean treatment parameters are the same. The second case arises when selection into treatment depends on β:

(C-2) $D \not\!\perp\!\!\!\perp \Delta$ *and* $E(\Delta \mid U_D) \neq E(\Delta)$ (*i.e.*, $\beta \not\!\perp\!\!\!\perp D$).

In this case, Δ^{MTE} is nonconstant, and in general, the treatment parameters differ among each other. In this case (IV-1) and (IV-2) for general instruments do not identify $\bar{\beta}$ (as shown in Section 4.1) or $E(\beta \mid D = 1)$.

A sufficient condition that generates (C-1) is the information condition that decisions to participate in the program are not made on the basis of $U_1 - U_0$ $(= \eta)$ (in the notation of Section 4.1):

(I-1) $\Pr(D = 1 \mid Z, U_1 - U_0) = \Pr(D = 1 \mid Z)$
(*i.e.*, $\Pr(D = 1 \mid Z, \beta) = \Pr(D = 1 \mid Z)$).[45]

[45] Given the assumption that $U_1 - U_0$ is independent of Z (given X), (I-1) implies $E(U_1 - U_0 \mid Z, X, D = 1) = E(U_1 - U_0 \mid X)$ so that the weaker mean independence condition is certainly satisfied:

(I-2) $E(U_1 - U_0 \mid Z, X, D = 1) = E(U_1 - U_0 \mid X, D = 1)$,

which is generically necessary and sufficient for linear IV to identify Δ^{TT} and Δ^{ATE}.

Before we investigate what standard instrumental variables estimators identify, we first present the local instrumental variables estimator which directly estimates the MTE. It is a limit form of LATE.

4.2.2. Estimating the MTE using local instrumental variables

Heckman and Vytlacil (1999, 2001b, 2005) develop the local instrumental variable (LIV) estimator to recover Δ^{MTE} pointwise. LIV is the derivative of the conditional expectation of Y with respect to $P(Z) = p$. This is defined as

$$\Delta^{\mathrm{LIV}}(p) \equiv \frac{\partial E(Y \mid P(Z) = p)}{\partial p}. \tag{4.9}$$

It is the population mean response to a policy change embodied in changes in $P(Z)$ analyzed by Björklund and Moffitt (1987). $E(Y \mid P(Z))$ is well defined as a consequence of assumption (A-4), and $E(Y \mid P(Z))$ can be recovered over the support of $P(Z)$.[46] Under our assumptions, LIV identifies MTE at all points of continuity in $P(Z)$ (conditional on X). This expression does not require additive separability of $\mu_1(X, U_1)$ or $\mu_0(X, U_0)$.[47]

Under standard regularity conditions, a variety of nonparametric methods can be used to estimate the derivative of $E(Y \mid P(Z))$ and thus to estimate Δ^{MTE}. With Δ^{MTE} in hand, if the support of the distribution of $P(Z)$ conditional on X is the full unit interval, one can generate all the treatment parameters defined in Section 3 as well as the policy relevant treatment parameter presented in Section 3.2 as weighted versions of Δ^{MTE}. When the support of the distribution of $P(Z)$ conditional on X is not full, it is still possible to identify some parameters. Heckman and Vytlacil (2001b) show that to identify ATE under assumptions (A-1)–(A-5), it is necessary and sufficient that the support of the distribution of $P(Z)$ include 0 and 1. Thus, identification of ATE does not require that the distribution of $P(Z)$ be the full unit interval or that the distribution of $P(Z)$ be continuous. But the support must include $\{0, 1\}$. Sharp bounds on the treatment parameters can be constructed under the same assumptions imposed in this chapter without imposing full support conditions. The resulting bounds are simple and easy to apply

[46] Assumptions (A-1), (A-3) and (A-4) jointly allow one to use Lebesgue's theorem for the derivative of an integral to show that $E(Y \mid P(Z) = p)$ is differentiable in p. Thus we can recover $\frac{\partial}{\partial p} E(Y \mid P(Z) = p)$ for almost all p that are limit points of the support of the distribution of $P(Z)$ (conditional on $X = x$). For example, if the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure, then all points in the support of the distribution of $P(Z)$ are limit points of that support and we can identify $\Delta^{\mathrm{LIV}}(p) = \frac{\partial E(Y|P(Z)=p)}{\partial p}$ for p (almost everywhere).

[47] Note, however, that it does require the assumption of additive separability between U_D and Z in the latent index for selection into treatment. Specifically, for LIV to identify MTE, we require additive separability in the choice equation. See our discussion in Section 4.10.

compared with those presented in the previous literature. We discuss these and other bounds in Section 10.

To establish the relationship between LIV and ordinary IV based on $P(Z)$ and to motivate how LIV identifies Δ^{MTE}, notice that from the definition of Y, the conditional expectation of Y given $P(Z)$ is, recalling that $\Delta = Y_1 - Y_0$,

$$E\big(Y \bigm| P(Z) = p\big) = E\big(Y_0 \bigm| P(Z) = p\big) + E\big(\Delta \bigm| P(Z) = p, D = 1\big)p,$$

where we keep the conditioning on X implicit. Our model and conditional independence assumption (A-1) imply

$$E\big(Y \bigm| P(Z) = p\big) = E(Y_0) + E(\Delta \mid p \geqslant U_D)p.$$

Applying the IV (Wald) estimator for two different values of $P(Z)$, p and p', for $p \neq p'$, we obtain:

$$\begin{aligned}&\frac{E(Y \mid P(Z) = p) - E(Y \mid P(Z) = p')}{p - p'}\\&\quad = \Delta^{\text{ATE}} + \frac{E(U_1 - U_0 \mid p \geqslant U_D)p - E(U_1 - U_0 \mid p' \geqslant U_D)p'}{p - p'},\end{aligned} \tag{4.10}$$

where this particular expression is obtained under the assumption of additive separability in the outcomes.[48,49] Exactly the same equation holds without additive separability if one replaces U_1 and U_0 with $Y_1 - E(Y_1 \mid X)$ and $Y_0 - E(Y_0 \mid X)$.

When $U_1 \equiv U_0$ or $(U_1 - U_0) \perp\!\!\!\perp U_D$ (case (C-1)), IV based on $P(Z)$ estimates Δ^{ATE} because the second term on the right-hand side of the expression (4.10) vanishes. Otherwise, IV estimates a combination of MTE parameters which we analyze further below.

Assuming additive separability of the outcome equations, another representation of $E(Y \mid P(Z) = p)$ reveals the index structure. It writes (keeping the conditioning on X implicit) that

$$\begin{aligned}&E\big(Y \bigm| P(Z) = p\big)\\&\quad = E(Y_0) + \Delta^{\text{ATE}}p + \int_0^p E(U_1 - U_0 \mid U_D = u_D)\,du_D.\end{aligned} \tag{4.11}$$

[48] The Wald estimator is IV for two values of the instrument.

[49] Observe that

$$\begin{aligned}E\big(Y \bigm| P(z) = p\big) &= E\big(Y_0 + D(Y_1 - Y_0) \bigm| P(z) = p\big)\\&= \mu_0 + E(Y_1 - Y_0 \mid P(z) = p, D = 1)\Pr(D = 1 \mid Z)\\&= \mu_0 + (\mu_1 - \mu_0)p + E(U_1 - U_0 \mid p \geqslant U_D)p.\end{aligned}$$

We can differentiate with respect to p and use LIV to identify Δ^{MTE}:

$$\Delta^{\mathrm{MTE}}(p) = \frac{\partial E(Y \mid P(Z) = p)}{\partial p} = \Delta^{\mathrm{ATE}} + E(U_1 - U_0 \mid U_D = p).$$[50]

Notice that IV estimates Δ^{ATE} when $E(Y \mid P(Z) = p)$ is a linear function of p so the third term on the right-hand side of (4.11) vanishes. Thus a test of the linearity of $E(Y \mid P(Z) = p)$ in p is a test of the validity of linear IV for Δ^{ATE}, i.e., it is a test of whether or not the data are consistent with a correlated random coefficient model ($\beta \not\perp\!\!\!\perp D$). The nonlinearity of $E(Y \mid P(Z) = p)$ in p provides a way to distinguish whether case (C-1) or case (C-2) describes the data. It is also a test of whether or not agents can at least partially anticipate future unobserved (by the econometrician) gains (the $Y_1 - Y_0$ given X) at the time they make their participation decisions. The levels and derivatives of $E(Y \mid P(Z) = p)$ and standard errors can be estimated using a variety of semiparametric methods. Heckman, Urzua and Vytlacil (2006) present an algorithm for estimating Δ^{MTE} using local linear regression.[51]

This analysis generalizes to the nonseparable outcomes case. We use separability in outcomes only to simplify the exposition and link to more traditional models. In particular, exactly the same expression holds with exactly the same derivation for the nonseparable case if we replace U_1 and U_0 with $Y_1 - E(Y_1 \mid X)$ and $Y_0 - E(Y_0 \mid X)$, respectively. This simple test for the absence of general heterogeneity based on linearity of $E(Y \mid Z)$ in $P(Z)$ applies to the case of LATE for any pair of instruments. An equivalent way is to check that all pairwise LATEs are the same over the sample support of Z.[52]

Figure 3A plots two cases of $E(Y \mid P(Z) = p)$ based on the generalized Roy model used to generate the example in Figures 2A and 2B. Recall that in this model, there are unobserved components of cost. When Δ^{MTE} $(= E(\beta \mid X = x, V = v))$ does not depend on u_D (or v) the expectation is a straight line. This is case (C-1). Figure 3B plots the derivatives of the two curves in Figure 3A. When Δ^{MTE} depends on u_D (or v) (case (C-2)), people sort into the program being studied positively on the basis of gains from the program, and one obtains the curved line depicted in Figure 3A.

[50] Making the conditioning on X explicit, we obtain that $E(Y \mid X = x, P(Z) = p) = E(Y_0 \mid X = x) + \Delta^{\mathrm{ATE}}(x)p + \int_0^p E(U_1 - U_0 \mid X = x, U_D = u_D)\,du_D$, with derivative with respect to p given by $\Delta^{\mathrm{MTE}}(x, p)$.

[51] Thus, one can apply any one of the large number of available tests for a parametric null versus a nonparametric alternative [see, e.g., Ellison and Ellison (1999), Zheng (1996)]. With regressors, the null is nonparametric leaving $E(Y \mid X = x, P(Z) = p)$ unspecified except for restrictions on the partial derivatives with respect to p. In this case, the formal test is that of a nonparametric null versus a nonparametric alternative, and a formal test of the null hypothesis can be implemented using the methodology of Chen and Fan (1999).

[52] Note that it is possible that $E(Y \mid Z)$ is linear in $P(Z)$ only over certain intervals of U_D, so there can be local dependence and local independence of (U_0, U_1, U_D).

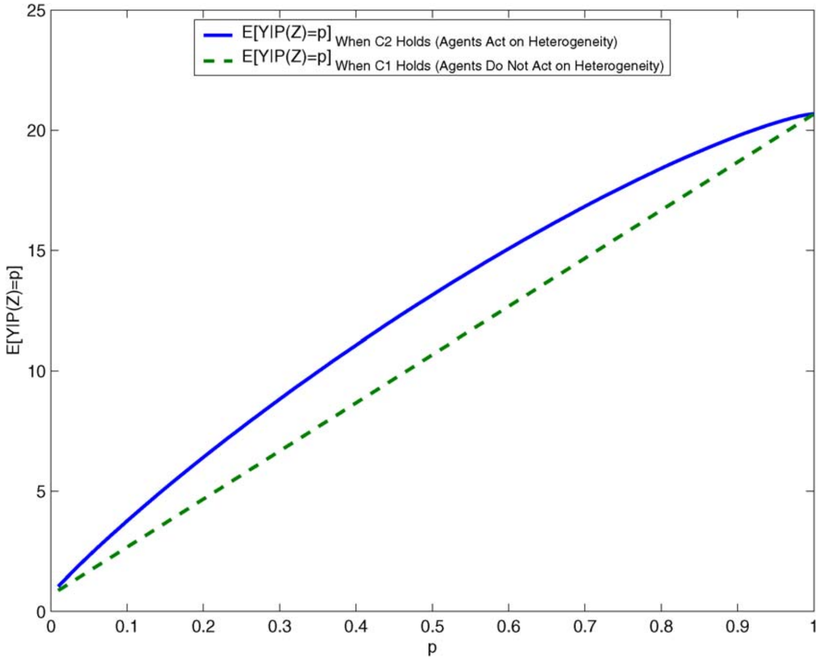


Figure 3A. Plot of the $E(Y \mid P(Z) = p)$. *Source*: Heckman and Vytlacil (2005).

4.3. What does linear IV estimate?

It is instructive to determine what linear IV estimates when Δ^{MTE} is nonconstant and conditions (A-1)–(A-5) hold. We analyze the general nonseparable case. We consider instrumental variables conditional on $X = x$ using a general function of Z as an instrument. We then specialize our result using $P(Z)$ as the instrument. As before, let $J(Z)$ be any function of Z such that $\text{Cov}(J(Z), D) \neq 0$. Define the IV estimator:

$$\beta_{\text{IV}}(J) \equiv \frac{\text{Cov}(J(Z), Y)}{\text{Cov}(J(Z), D)},$$

where to simplify the notation we keep the conditioning on X implicit. Appendix D derives a representation of this expression in terms of weighted averages of the MTE displayed in Table 2B. We exposit this expression in this section.

In Appendix D, we establish that

$$\begin{aligned}&\text{Cov}\bigl(J(Z), Y\bigr)\\&\quad = \int_0^1 \Delta^{\text{MTE}}(u_D) E\bigl(J(Z) - E\bigl(J(Z)\bigr) \bigm| P(Z) \geqslant u_D\bigr) \Pr\bigl(P(Z) \geqslant u_D\bigr)\, du_D.\end{aligned} \tag{4.12}$$

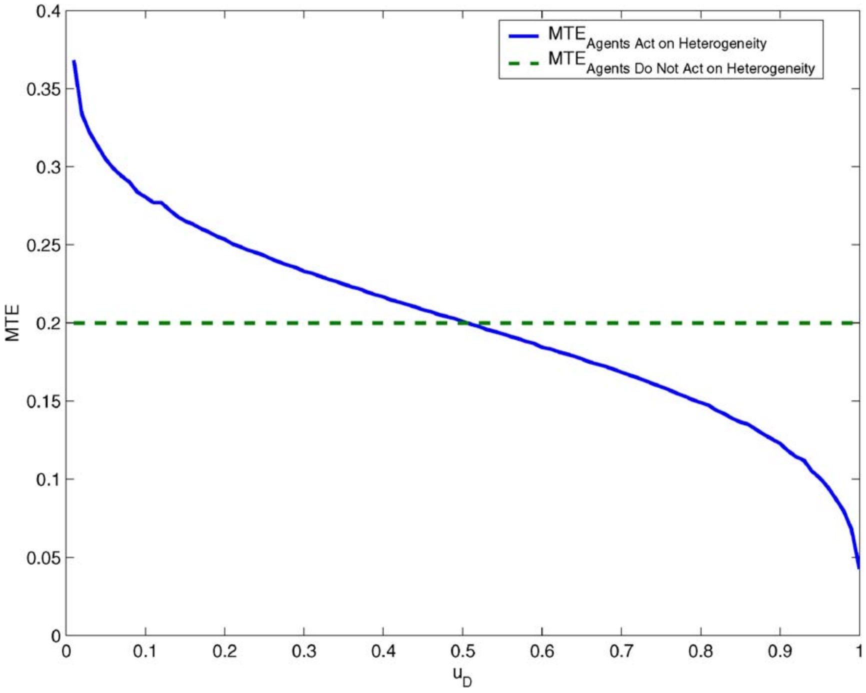


Figure 3B. Plot of the identified marginal treatment effect from Figure 3A (the derivative). *Source*: Heckman and Vytlacil (2005). *Note*: Parameters for the general heterogeneous case are the same as those used in Figures 2A and 2B. For the homogeneous case we impose $U_1 = U_0$ ($\sigma_1 = \sigma_2 = 0.012$).

By the law of iterated expectations, $\mathrm{Cov}(J(Z), D) = \mathrm{Cov}(J(Z), P(Z))$. Thus

$$\beta_{\mathrm{IV}}(J) = \int_0^1 \Delta^{\mathrm{MTE}}(u_D)\omega_{\mathrm{IV}}(u_D \mid J)\, du_D,$$

where

$$\omega_{\mathrm{IV}}(u_D \mid J) = \frac{E(J(Z) - E(J(Z)) \mid P(Z) \geqslant u_D)\Pr(P(Z) \geqslant u_D)}{\mathrm{Cov}(J(Z), P(Z))}, \tag{4.13}$$

assuming the standard rank condition (IV-2) holds: $\mathrm{Cov}(J(Z), P(Z)) \neq 0$. The weights integrate to one,

$$\int_0^1 \omega_{\mathrm{IV}}(u_D \mid J)\, du_D = 1,$$

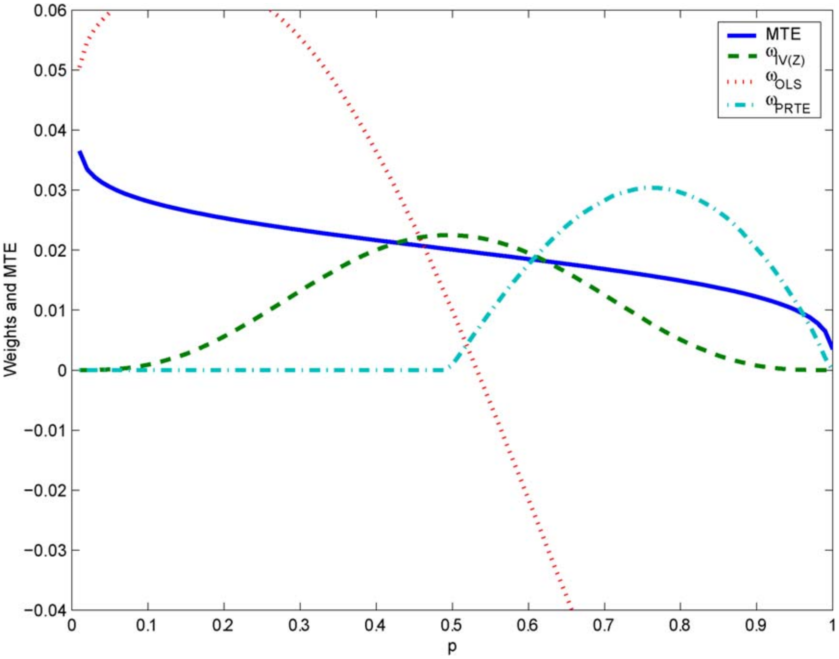


Figure 4A. MTE vs. linear instrumental variables, ordinary least squares, and policy relevant treatment effect weights: when $P(Z)$ is the instrument. The policy is given at the base of Table 3. The model parameters are given at the base of Figure 2. *Source*: Heckman and Vytlacil (2005).

and can be constructed from the data on $P(Z)$, $J(Z)$ and D. Assumptions about the properties of the weights are testable.[53]

We discuss additional properties of the weights for the special case where the propensity score is the instrument $J(Z) = P(Z)$. We then analyze the properties of the weights for a general instrument $J(Z)$. When $J(Z) = P(Z)$, Equation (4.13) specializes to

$$\omega_{\mathrm{IV}}\big(u_D \mid P(Z)\big) = \frac{[E(P(Z) \mid P(Z) \geqslant u_D) - E(P(Z))]\,\Pr(P(Z) \geqslant u_D)}{\operatorname{Var}(P(Z))}.$$

Figure 4A plots the IV weight for $J(Z) = P(Z)$ and the MTE for our generalized Roy model example developed in Figures 2 and 3 and Table 3. The weights are positive and peak at the mean of P. Figure 4A also plots the OLS weight given in Table 2 and the weight for a policy exercise described below Table 3 and discussed further below.

[53] Expressions for IV and OLS as weighted averages of marginal response functions, and the properties and construction of the weights, were first derived by Yitzhaki in 1989 in a paper that was eventually published in 1996 [see Yitzhaki (1996)]. Under monotonicity (IV-3), his expression is a weighted average of MTEs or LATEs. We present Yitzhaki's derivation in Appendix D.

Let p^{Min} and p^{Max} denote the minimum and maximum points in the support of the distribution of $P(Z)$ (conditional on $X = x$). The weights on MTE when $P(Z)$ is the instrument are nonnegative for all evaluation points, are strictly positive for $u_D \in (p^{\text{Min}}, p^{\text{Max}})$ and are zero for $u_D < p^{\text{Min}}$ and for $u_D > p^{\text{Max}}$.[54]

The properties of the weights for general $J(Z)$ depend on the conditional relationship between $J(Z)$ and $P(Z)$. From the general expression for (4.13), it is clear that the IV estimator with $J(Z)$ as an instrument satisfies the following properties:

(i) Two instruments J and J^* weight MTE equally at all values of u_D if and only if they have the same (centered) conditional expectation of J given P, i.e., $E(J \mid P(Z) = p) - E(J) = E(J^* \mid P(Z) = p) - E(J^*)$ for all p in the support of the distribution of $P(Z)$.

(ii) The support of $\omega_{\text{IV}}(u_D \mid J)$ is contained in $[p^{\text{Min}}, p^{\text{Max}}]$ the minimum and maximum value of p in the population (given x). Therefore $\omega_{\text{IV}}(t \mid J) = 0$ for $t < p^{\text{Min}}$ and for $t > p^{\text{Max}}$. Using any instrument other than $P(Z)$ leads to nonzero weights only on a subset of $[p^{\text{Min}}, p^{\text{Max}}]$, and using the propensity score as an instrument leads to nonnegative weights on a larger range of evaluation points than using any other instrument.

(iii) $\omega_{\text{IV}}(u_D \mid J)$ is nonnegative for all u_D if $E(J \mid P(Z) \geqslant p)$ is weakly monotonic in p. Using J as an instrument yields nonnegative weights on Δ^{MTE} if $E(J \mid P(Z) \geqslant p)$ is weakly monotonic in p. This condition is satisfied when $J(Z) = P(Z)$. More generally, if J is a monotonic function of $P(Z)$, then using J as the instrument will lead to nonnegative weights on Δ^{MTE}. There is no guarantee that the weights for a general $J(Z)$ will be nonnegative for all u_D, although the weights integrate to unity and thus must be positive over some range of evaluation points. We produce examples below where the instrument leads to negative weights for some evaluation points. Imbens and Angrist (1994) assume that $J(Z)$ is monotonic in $P(Z)$ and thus produce positive weights. Our analysis is more general.

[54] For u_D evaluation points between p^{Min} and p^{Max}, $u_D \in (p^{\text{Min}}, p^{\text{Max}})$, we have that

$$E\bigl(P(Z) \bigm| P(Z) \geqslant u_D\bigr) > E\bigl(P(Z)\bigr) \quad \text{and} \quad \Pr\bigl(P(Z) \geqslant u_D\bigr) > 0,$$

so that $\omega_{\text{IV}}(u_D \mid P(Z)) > 0$ for any $u_D \in (p^{\text{Min}}, p^{\text{Max}})$. For $u_D < p^{\text{Min}}$,

$$E\bigl(P(Z) \bigm| P(Z) \geqslant u_D\bigr) = E\bigl(P(Z)\bigr).$$

For any $u_D > p^{\text{Max}}$, $\Pr(P(Z) \geqslant u_D) = 0$. Thus, $\omega_{\text{IV}}(u_D \mid P(Z)) = 0$ for any $u_D < p^{\text{Min}}$ and for any $u_D > p^{\text{Max}}$. $\omega_{\text{IV}}(u_D \mid P(Z))$ is strictly positive for $u_D \in (p^{\text{Min}}, p^{\text{Max}})$, and is zero for all $u_D < p^{\text{Min}}$ and all $u_D > p^{\text{Max}}$. Whether the weights are nonzero at the endpoints depends on the distribution of $P(Z)$. However, since the weights are defined for integration with respect to Lebesgue measure, the value taken by the weights at p^{Min} and p^{Max} does not affect the value of the integral.

The propensity score plays a central role in determining the properties of the weights. The IV weighting formula critically depends on the conditional mean dependence between instrument $J(Z)$ and the propensity score.

The interpretation placed on the IV estimand depends on the specification of $P(Z)$ even if only Z_1 (e.g., the first coordinate of Z) is used as the instrument. This drives home the point about the difference between IV in the traditional model and IV in the more general model with heterogeneous responses analyzed in this chapter. In the traditional model, the choice of any valid instrument and the specification of instruments in $P(Z)$ not used to construct a particular IV estimator does not affect the IV estimand. In the more general model, these choices matter. Two economists, using the same $J(Z) = Z_1$, will obtain the same IV point estimate, but the interpretation placed on that estimate will depend on the specification of the Z in $P(Z)$ even if $P(Z)$ is not used as an instrument. The weights can be positive for one instrument and negative for another. We show some examples after developing the properties of the IV weights.

4.3.1. Further properties of the IV weights

Expression (4.13) for the weights does not impose any support conditions on the distribution of $P(Z)$, and thus does not require either that $P(Z)$ be continuous or discrete. To demonstrate this, consider two extreme special cases: (i) when $P(Z)$ is a continuous random variable, and (ii) when $P(Z)$ is a discrete random variable.

To simplify the exposition, initially assume that $J(Z)$ and $P(Z)$ are jointly continuous random variables. This assumption plays no essential role in any of the results of this chapter and we develop the discrete case after developing the continuous case. The weights defined in Equation (4.13) can be written as

$$\omega_{\mathrm{IV}}(u_D) = \frac{\int (j - E(J(Z))) \int_{u_D}^{1} f_{J,P}(j,t)\,dt\,dj}{\operatorname{Cov}(J(Z), D)}, \tag{4.14}$$

where $f_{J,P}$ is the joint density of $J(Z)$ and $P(Z)$ and we implicitly condition on X. The weights can be negative or positive. Observe that $\omega(0) = 0$ and $\omega(1) = 0$. The weights integrate to 1 because as shown in Appendix D,

$$\iint \bigl(j - E\bigl(J(Z)\bigr)\bigr) \int_{u_D}^{1} f_{J,P}(j,t)\,dt\,dj\,du_D = \operatorname{Cov}\bigl(J(Z), D\bigr),$$

so even if the weight is negative over some intervals, it must be positive over other intervals. Observe that when there is one instrument (Z is a scalar), and assumptions (A-1)–(A-5) are satisfied, the weights are always positive provided $J(Z)$ is a monotonic function of the scalar Z. In this case, which is covered by (4.13) but excluded in deriving (4.14), $J(Z)$ and $P(Z)$ have the same distribution and $f_{J,P}(j,t)$ collapses to a univariate distribution. The possibility of negative weights arises when $J(Z)$ is not a monotonic function of $P(Z)$. It also arises when there are two or more instruments, and the analyst computes estimates with only one instrument or a combination of the Z instruments that is not a monotonic function of $P(Z)$ so that $J(Z)$ and $P(Z)$ are not perfectly dependent. If the instrument is $P(Z)$ (so $J(Z) = P(Z)$) then the weights are everywhere nonnegative because from (4.14), $E(P(Z) \mid P(Z) > u_D) - E(P(Z)) \geqslant 0$. In this case, the density of $(P(Z), J(Z))$ collapses to the density of $P(Z)$. For any scalar Z, we can define $J(Z)$ and $P(Z)$ so that they are perfectly dependent, provided that $J(Z)$ and $P(Z)$ are monotonic in Z. Generally, the weight (4.13) is positive if $E(J(Z) \mid P(Z) > u_D)$ is weakly monotonic in u_D. Nonmonotonicity of this expression can produce negative weights.[55]

4.3.2. Constructing the weights from data

Observe that the weights can be constructed from data on (J, P, D). Data on $(J(Z), P(Z))$ pairs and $(J(Z), D)$ pairs (for each X value) are all that is required. We can use a smoothed sample frequency to estimate the joint density $f_{J,P}$. Thus, given our maintained assumptions, any property of the weight, including its positivity at any point (x, u_D), can be examined with data. We present examples of this approach below.

As is evident from Tables 2A and 2B and Figures 2A and 2B, the weights on $\Delta^{\mathrm{MTE}}(u_D)$ generating Δ^{IV} are different from the weights on $\Delta^{\mathrm{MTE}}(u_D)$ that generate the average treatment effect which is widely regarded as an important policy parameter [see, e.g., Imbens (2004)] or from the weights associated with the policy relevant treatment parameter which answers well-posed policy questions [Heckman and Vytlacil (2001b, 2005)]. It is not obvious why the weighted average of $\Delta^{\mathrm{MTE}}(u_D)$ produced by IV is of any economic interest. Since the weights can be negative for some values of u_D, $\Delta^{\mathrm{MTE}}(u_D)$ can be positive everywhere in u_D but IV can be negative. Thus, IV may not estimate a treatment effect for any person. We present some examples of IV models with negative weights below. A basic question is why estimate the model with IV at all given the lack of any clear economic interpretation of the IV estimator in the general case.

4.3.3. Discrete instruments

The representation (4.13) can be specialized to cover discrete instruments, $J(Z)$. Consider the case where the distribution of $P(Z)$ (conditional on X) is discrete. The support of the distribution of $P(Z)$ contains a finite number of values $p_1 < p_2 < \cdots < p_K$ and the support of the instrument $J(Z)$ is also discrete taking I distinct values where I and K may be distinct. $E(J(Z) \mid P(Z) \geqslant u_D)$ is constant in u_D, for u_D within any $(p_\ell, p_{\ell+1})$ interval, and $\Pr(P(Z) \geqslant u_D)$ is constant in u_D, for u_D within any $(p_\ell, p_{\ell+1})$ interval, and thus $\omega_{\mathrm{IV}}^{J}(u_D)$ is constant in u_D over any $(p_\ell, p_{\ell+1})$ interval. Let λ_ℓ denote

[55] If it is weakly monotonically increasing, the claim is evident from (4.13). If it is decreasing, the sign of the numerator and the denominator are both negative so the weight is nonnegative.

the weight on LATE for the interval $(\ell, \ell+1)$. In this notation,

$$\begin{aligned}\Delta_J^{\mathrm{IV}} &= \int E(Y_1 - Y_0 \mid U_D = u_D)\omega_{\mathrm{IV}}^J(u_D)\,du_D \\ &= \sum_{\ell=1}^{K-1} \lambda_\ell \int_{p_\ell}^{p_{\ell+1}} E(Y_1 - Y_0 \mid U_D = u_D)\frac{1}{(p_{\ell+1} - p_\ell)}\,du_D \\ &= \sum_{\ell=1}^{K-1} \Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1})\lambda_\ell. \end{aligned} \tag{4.15}$$

Let j_i be the ith smallest value of the support of $J(Z)$. The discrete version of Equation (4.13) is

$$\lambda_\ell = \frac{\sum_{i=1}^{I}(j_i - E(J(Z)))\sum_{t>\ell}^{K}(f(j_i, p_t))}{\operatorname{Cov}(J(Z), D)}(p_{\ell+1} - p_\ell), \tag{4.16}$$

where f is the probability frequency of (j_i, p_t): the probability that $J(Z) = j_i$ and $P(Z) = p_t$. There is no presumption that high values of $J(Z)$ are associated with high values of $P(Z)$. $J(Z)$ can be one coordinate of Z that may be positively or negatively dependent on $P(Z)$, which depends on the full vector. In the case of scalar Z, as long as $J(Z)$ and $P(Z)$ are monotonic in Z there is perfect dependence between $J(Z)$ and $P(Z)$. In this case, the joint probability density collapses to a univariate density and the weights have to be positive, exactly as in the case for continuous instruments previously discussed. Our expression for the weight on LATE generalizes the expression presented by Imbens and Angrist (1994) who in their analysis of the case of vector Z only consider the case where $J(Z)$ and $P(Z)$ are perfectly dependent because $J(Z)$ is a monotonic function of $P(Z)$.[56] More generally, the weights can be positive or negative for any ℓ but they must sum to 1 over all ℓ.

Monotonicity or uniformity is a property needed with just two values of Z, $Z = z_1$ and $Z = z_2$, to guarantee that IV estimates a treatment effect. With more than two values of Z, we need to weight the LATEs and MTEs. If the instrument $J(Z)$ shifts $P(Z)$ in the same way for everyone, it shifts D in the same way for everyone since $D = \mathbf{1}[P(Z) \geqslant U_D]$ and Z is independent of U_D. If $J(Z)$ is not monotonic in $P(Z)$, it may shift $P(Z)$ in different ways for different people. Negative weights are a tip-off of two-way flows. We present examples below.

4.3.4. *Identifying margins of choice associated with each instrument and unifying diverse instruments within a common framework*

We have just established that different instruments weight the MTE differently. Using $P(Z)$ in the local IV estimator, we can identify the MTE. We can construct the weights

[56] In their case, $I = K$ and $f(j_i, p_t) = 0, \forall i \neq t$.

associated with each instrument from the joint distribution of $(J(Z), P(Z))$ given X. By plotting the weights for each instrument, we can determine the margins identified by the different instruments. Using $P(Z)$ as the instrument enables us to extend the support associated with any single instrument, and to determine which segment of the MTE is identified by any particular instrument. As before, we keep conditioning on X implicit.

4.3.5. *Yitzhaki's derivation of the weights*

An alternative and in some ways more illuminating way to derive the weights used in IV is to follow Yitzhaki (1989, 1996) and Yitzhaki and Schechtman (2004) who prove for a general regression function $E(Y \mid P(Z) = p)$ that a linear regression of Y on P estimates

$$\beta_{Y,P} = \int_0^1 \left[\frac{\partial E(Y \mid P(Z) = p)}{\partial p}\right] \omega(p)\, dp,$$

where

$$\omega(p) = \frac{\int_p^1 (t - E(P))\, dF_P(t)}{\operatorname{Var}(P)},$$

which is exactly the weight (4.13) when P is the instrument. Thus we can interpret (4.13) as the weight on $\frac{\partial E(Y|P(Z)=p)}{\partial p}$ when two-stage least squares (2SLS) based on $P(Z)$ is used to estimate the "causal effect" of D on Y. Under uniformity,

$$\left.\frac{\partial E(Y \mid P(Z) = p)}{\partial p}\right|_{p=u_D} = E(Y_1 - Y_0 \mid U_D = u_D) = \Delta^{\text{MTE}}(u_D).[57]$$

Our analysis is more general than that of Yitzhaki (1989) or Imbens and Angrist (1994) because we allow for instruments that are not monotonic functions of $P(Z)$, whereas the Yitzhaki weighting formula only applies to instruments that are monotonic functions of $P(Z)$.[58] The analysis of Yitzhaki (1989) is more general than that of Imbens and Angrist (1994), because he does not impose uniformity (monotonicity). We present some further examples of these weights after discussing the role of $P(Z)$ and the role of monotonicity and uniformity. We present Yitzhaki's Theorem and the relationship of our analysis to Yitzhaki's analysis in Appendices D.1 and D.2.

[57] Yitzhaki's weights are used by Angrist and Imbens (1995) to interpret what 2SLS estimates in the model of Equation (4.1) with heterogeneous β. Yitzhaki (1989) derives the finite sample weights used by Imbens and Angrist. See the refinement in Yitzhaki and Schechtman (2004).

[58] Heckman and Vytlacil (2001b) generalize the Yitzhaki analysis of the IV weights by relaxing separability (monotonicity).

4.4. The central role of the propensity score

Observe that both (4.13) and (4.14) (and their counterparts for LATE (4.15) and (4.16)) contain expressions involving the propensity score $P(Z)$, the probability of selection into treatment. Under our assumptions, it is a monotonic function of the mean utility of treatment, $\mu_D(Z)$. The propensity score plays a central role in selection models as a determinant of control functions in selection models [Heckman and Robb (1985a, 1986a)] as noted in Section 4.1. In matching models, it provides a computationally convenient way to condition on Z [see, e.g., Rosenbaum and Rubin (1983), Heckman and Navarro (2004), and the discussion in Section 8]. For the IV weight to be correctly constructed and interpreted, we need to know the correct model for $P(Z)$, i.e., we need to know exactly which Z determine $P(Z)$. As previously noted, this feature is not required in the traditional model for instrumental variables based on response heterogeneity. In that simpler framework, any instrument will identify $\mu_1(X)-\mu_0(X)$ and the choice of a particular instrument affects efficiency but not identifiability. One can be casual about the choice model in the traditional setup, but not in the model of choice of treatment with essential heterogeneity. Thus, unlike the application of IV to traditional models under condition (C-1), IV applied in the model of essential heterogeneity depends on (a) the choice of the instrument $J(Z)$, (b) its dependence with $P(Z)$, the true propensity score or choice probability, and (c) the specification of the propensity score (i.e., what variables go into Z). Using the propensity score one can identify LIV and LATE and the marginal returns at values of the unobserved U_D. From the MTE identified by $P(Z)$ and the weights that can be constructed from the joint distribution of $(J(Z), P(Z))$ given X, we can identify the segment of the MTE identified by any IV.

4.5. Monotonicity, uniformity and conditional instruments

Monotonicity, or uniformity condition (IV-3), is a condition on a collection of counterfactuals for each person and hence is not testable, since we know only one element of the collection for any person. It rules out general heterogeneous responses to treatment choices in response to changes in vector Z. The recent literature on instrumental variables with heterogeneous responses is thus asymmetric. Outcome equations can be heterogeneous in a general way while choice equations cannot be. If $\mu_D(Z) = Z\gamma$, where γ is a common coefficient shared by everyone, the choice model satisfies the uniformity property. On the other hand, if γ is a random coefficient (i.e., has a nondegenerate distribution) that can take both negative and positive values, and there are two or more variables in Z with nondegenerate γ coefficients, uniformity can be violated. Different people can respond to changes in Z differently, so there can be nonuniformity. The uniformity condition can be violated even when all components of γ are of the same sign if Z is a vector and γ is a nondegenerate random variable.[59]

[59] Thus if $\gamma > 0$ for each component and some components of Z are positive and others are negative, changes from z' to z can increase γZ for some and decrease γZ for others since the γ are different among persons.

Changing one coordinate of Z, holding the other coordinates at different values across people is *not* the experiment that defines monotonicity or uniformity. Changing one component of Z, allowing the other coordinates of Z to vary across people, does not necessarily produce uniform flows toward or against participation in the treatment status. For example, let $\mu_D(z) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_1 z_2$, where γ_0, γ_1, γ_2 and γ_3 are constants, and consider changing z_1 from a common base state while holding z_2 fixed at different values across people. If $\gamma_3 < 0$, then $\mu_D(z)$ does not necessarily satisfy the uniformity condition. If we move (z_1, z_2) as a pair from the same base values to the same destination values z', uniformity is satisfied even if $\gamma_3 < 0$, although $\mu_D(z)$ is not a monotonic function of z.[60]

Positive weights and uniformity are distinct issues.[61] Under uniformity, and assumptions (A-1)–(A-5), the weights on MTE for any particular instrument may be positive or negative. The weights for MTE using $P(Z)$ must be positive as we have shown so the propensity score has a special status as an instrument. Negative weights associated with the use of $J(Z)$ as an instrument do not necessarily imply failure of uniformity in Z. Even if uniformity is satisfied for Z, it is not necessarily satisfied for $J(Z)$. Condition (IV-3) is an assumption about a vector. Fixing one combination of Z (when J is a function of Z) or one coordinate of Z does not guarantee uniformity in J even if there is uniformity in Z. The flow created by changing one coordinate of Z can be reversed by the flow created by the other components of Z if there is negative dependence among components even if *ceteris paribus* all components of Z affect D in the same direction. We present some examples below.

The issues of positive weights and the existence of one way flows in response to an intervention are conceptually distinct. Even with two values for a scalar Z, flows may be two way [see Equation (4.5)]. If we satisfy (IV-3) for a vector, so uniformity applies, weights for a particular instrument may be negative for certain intervals of U_D (i.e., for some of the LATE parameters).

[60] Associated with $Z = z$ is the counterfactual random variable $D(z)$. Associated with the scalar random variable $J(Z)$ constructed from Z is a counterfactual random variable $D(j(z))$ which is in general different from $D(z)$. The random variable $D(z)$ is constructed from (3.3) using $\mathbf{1}[\mu_D(z) \geqslant V]$. In this expression, V assumes individual specific values which remain fixed as we set different z values. From (A-1), $\Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z)$. The random variable $D(j)$ is defined by the following thought experiment. For each possible realization j of $J(Z)$, define $D(j)$ by setting $D(j) = D(Z(j))$ where $Z(j)$ is a random draw from the distribution of Z conditional on $J(Z) = j$. Set $D(j)$ equal to the choice that would be made given that draw of $Z(j)$. Thus $D(j)$ is a function of $(Z(j), u_D)$. As long as we draw $Z(j)$ randomly (so independent of Z), we have that $(Z(j), U_D) \perp\!\!\!\perp Z$ so $D(j) \perp\!\!\!\perp Z$. There are other possible constructions of the counterfactual $D(j)$ since there are different possible distributions from which Z can be drawn, apart from the actual distribution of Z. The advantage of this construction is that it equates the counterfactual probability that $D(j) = 1$ given $J(Z) = j$ with the population probability. If the Z were uncertain to the agent, this would be a rational expectations assumption. At their website, Heckman, Urzua and Vytlacil (2006) discuss this assumption further.

[61] When they analyze the vector case, Imbens and Angrist (1994) analyze instruments that are monotonic functions of $P(Z)$. Our analysis is more general and recognizes that, in the vector case, IV weights may be negative or positive.

If we condition on $Z_2 = z_2, \ldots, Z_K = z_K$ using Z_1 as an instrument, then a uniform flow condition is satisfied. We call this *conditional uniformity*. By conditioning, we effectively convert the problem back to that of a scalar instrument where the weights must be positive. If uniformity holds for Z_1, fixing the other Z at common values, then one-dimensional LATE/MTE analysis applies. Clearly, the weights have to be defined conditionally.

The concept of conditioning on other instruments to produce positive weights for the selected instrument is a new idea, not yet appreciated in the empirical IV literature and has no counterpart in the traditional IV model. In the conventional model, the choice of a valid instrument affects efficiency but not the definition of the parameters as it does in the more general case.[62]

In summary, nothing in the economics of choice guarantees that if Z is changed from z to z', that people respond in the same direction to the change. See the general expression (4.5). The condition that people respond to choices in the same direction for the same change in Z does not imply that $D(z)$ is monotonic in z for any person in the usual mathematical usage of the term monotonicity. If $D(z)$ is monotonic in the usual usage of this term and responses are in the same direction for all people, then "monotonicity" or better "uniformity" condition (IV-3) would be satisfied.

If responses to a common change of Z are heterogenous in a general way, we obtain (4.5) as the general case. Vytlacil's 2002 Theorem breaks down and IV cannot be expressed in terms of a weighted average of MTE terms. Nonetheless, Yitzhaki's characterization of IV, derived in Appendix D, remains valid and the weights on $\frac{\partial E(Y|P=p)}{\partial p}$ are positive and of the same form as the weights obtained for MTE (or LATE) when the monotonicity condition holds. IV can still be written as a weighted average of LIV terms, even though LIV does not identify the MTE.

4.6. Treatment effects vs. policy effects

Even if uniformity condition (IV-3) fails, IV may answer relevant policy questions. By Yitzhaki's analysis, summarized in Section 4.3.5, IV or 2SLS estimates a weighted average of marginal responses which may be pointwise positive or negative. Policies may induce some people to switch into and others to switch out of choices, as is evident from Equation (4.5). These net effects are of interest in many policy analyses. Thus, subsidized housing in a region supported by higher taxes may attract some to migrate to the region and cause others to leave. The net effect from the policy is all that is required to perform cost benefit calculations of the policy on outcomes. If the housing subsidy is the instrument, and the net effect of the subsidy is the parameter of interest, the issue of monotonicity is a red herring. If the subsidy is exogenously imposed, IV estimates the

[62] In the conventional model, with homogenous responses, a linear probability approximation to $P(Z)$ used as an instrument would identify the same parameter as $P(Z)$. In the general model, replacing $P(Z)$ by a linear probability approximation of it (e.g., $E(D \mid Z) = \pi Z = J(Z)$) is not guaranteed to produce positive weights for $\Delta^{\text{MTE}}(x, u_D)$ or $\Delta^{\text{LATE}}(x, u'_D, u_D)$, or to replicate the weights based on the correctly specified $P(Z)$.

net effect of the policy on mean outcomes. Only if the effect of migration on outcomes induced by the subsidy on outcomes is the question of interest, and not the effect of the subsidy, does uniformity emerge as an interesting condition.

4.7. Some examples of weights in the generalized Roy model and the extended Roy model

It is useful to develop intuition about the properties of the IV estimator and the structure of the weights for two prototypical choice models. We develop the weights for a generalized Roy model where unobserved cost components are present and an extended Roy model where cost components are observed but there are no unobserved cost components. The extended Roy model is used to generate Figure 1 and was introduced at the end of Section 2.

Table 3 presents the IV estimand for the generalized Roy model used to generate Figures 2A and 2B using $P(Z)$ as the instrument. The model generating $D = \mathbf{1}[Z\gamma \geqslant V]$ is given at the base of Figure 2B (Z is a scalar, γ is 1, V is normal, $U_D = \Phi(\frac{V}{\sigma_V})$). We compare the IV estimand with the policy relevant treatment effect for a policy precisely defined at the base of Table 3. This policy has the structure that if $Z > 0$, persons get a bonus Zt for participation in the program, where $t > 0$. The decision rule for program participation for $Z > 0$ is $D = \mathbf{1}[Z(1+t) \geqslant V]$. People are not forced into participation in the program but are rather induced into it by the bonus. Given the assumed distribution of Z, and the other parameters of the model, we obtain the policy relevant treatment parameter weight $\omega_{\text{PRTE}}(u_D)$ as plotted in Figures 4A–4C (the scales of the ordinates differ across the graphs, but the weight is the same). We use the per capita PRTE and consider three instruments. Table 5 presents estimands for the three instruments shown in the table for the generalized Roy model in three environments.

The first instrument we consider for this example is $P(Z)$, which assumes that there is no policy in place ($t = 0$). It is identified (estimated) on a sample with no policy in place but otherwise the model is the same as the one with the policy in place. The weight on this instrument is plotted in Figure 4A. That figure also displays the OLS weight as well as the MTE that is being weighted to generate the estimate. It also shows the weight used to generate PRTE. The IV weights for $P(Z)$ and the weights for Δ^{PRTE} differ. This is as it should be because Δ^{PRTE} is making a comparison across regimes but the IV in this case makes comparisons within a no policy regime. Given the shape of $\Delta^{\text{MTE}}(u_D)$, it is not surprising that the estimand for IV based on $P(Z)$ is so much above the Δ^{PRTE} which weights a lower-valued segment of $\Delta^{\text{MTE}}(u_D)$ more heavily.[63]

The second instrument we consider exploits the variation induced by the policy in place and fits it on samples where the policy is in place (i.e., the t is the same as that

[63] Heckman and Vytlacil (2005) show how to construct the proper instrument for such policies using a pre-policy sample.

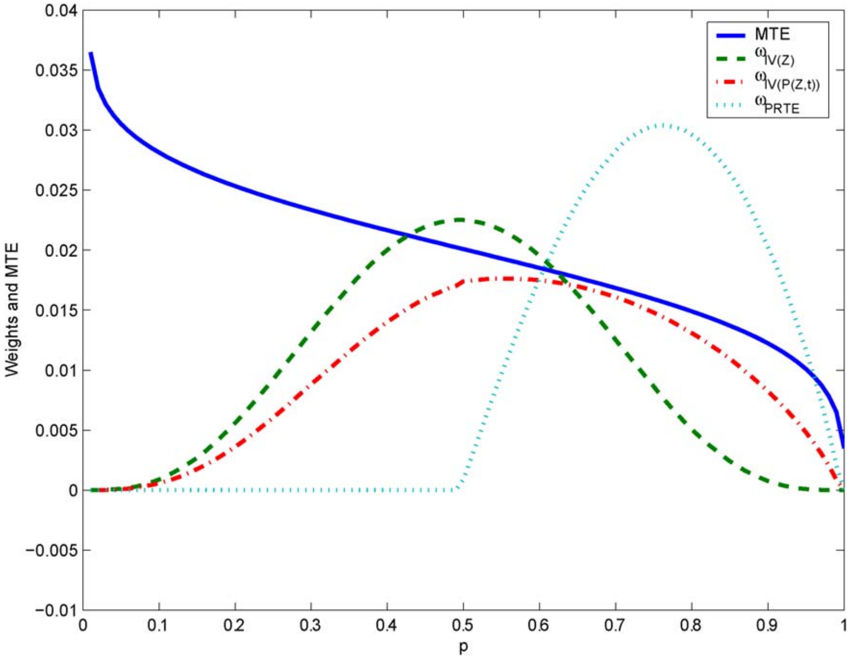


Figure 4B. MTE vs. linear IV with $P(Z(1 + t(\mathbf{1}[Z > 0]))) = \tilde{P}(z, t)$ as an instrument, and policy relevant treatment effect weights for the policy defined at the base of Table 3. The model parameters are given at the base of Figure 2. *Source*: Heckman and Vytlacil (2005).

used to generate the PRTE). On intuitive grounds, this instrument might be thought to work well in identifying the PRTE, but in fact it does not. The instrument is $\tilde{P}(Z, t) = P(Z(1 + t\mathbf{1}[Z > 0]))$ which jumps in value when Z switches from $Z < 0$ to $Z > 0$. This is the choice probability in the regime with the policy in place. Figure 4B plots the weight for this IV along with the weight for $P(Z)$ as an IV and the weight for PRTE (repeated from Figure 4A).[64] While this weight looks a bit more like the weight for Δ^{PRTE} than the previous instrument, it is clearly different.

Figure 4C plots the weight for an ideal instrument for PRTE: a randomization of eligibility. This compares the outcomes in one population where the policy is in place with outcomes in a regime where the policy is not in place. Thus we use an instrument B such that

$$B = \begin{cases} 1 & \text{if a person is eligible to participate in the program,} \\ 0 & \text{otherwise.} \end{cases}$$

Persons for whom $B = 1$, make their participation choices under the policy with a jump in Z, $t\mathbf{1}(Z > 0)$, in their choice sets.[65] If $B = 0$, persons are embargoed from

[64] Remember that the scales are different across the two graphs.

[65] Recall that, in this example, we set $\gamma = 1$.

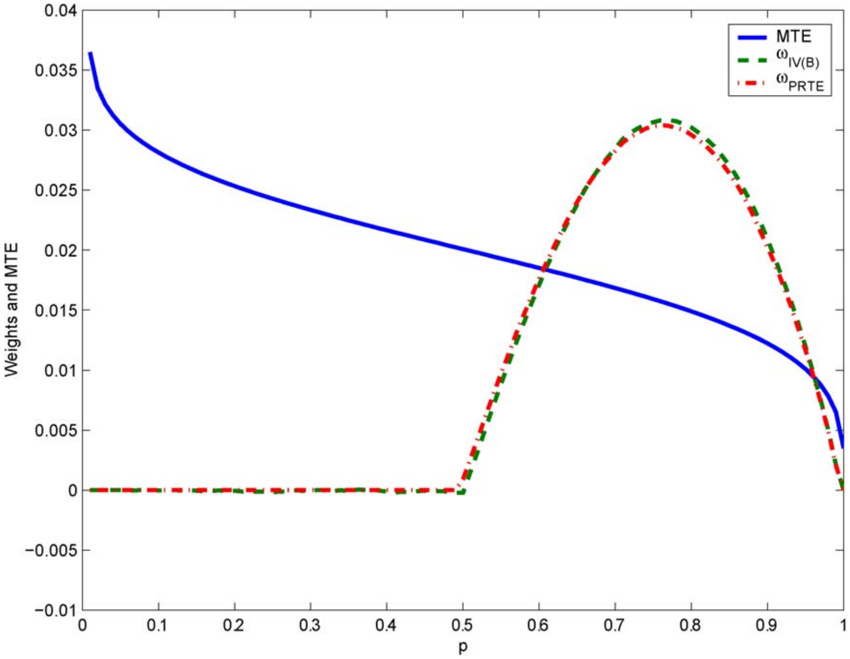


Figure 4C. MTE vs. IV policy and policy relevant treatment effect weights for the policy defined at the base of Table 3. *Source*: Heckman and Vytlacil (2005).

the policy and cannot receive a bonus. The $B = 0$ case is a prepolicy regime. We assume $\Pr[B = 1 \mid Y_0, Y_1, V, Z] = \Pr[B = 1] = 0.5$, so all persons are equally likely to receive or not receive eligibility for the bonus and assignment does not depend on model unobservables in the outcome equation.

The Wald estimator in this case is

$$\frac{E(Y \mid B = 1) - E(Y \mid B = 0)}{\Pr(D = 1 \mid B = 1) - \Pr(D = 1 \mid B = 0)}.$$

Table 5
Linear instrumental variable estimands and the policy relevant treatment effect

Using propensity score $P(Z)$ as the instrument	0.2013
Using propensity score $P(Z(1 + t(\mathbf{1}[Z > 0])))$ as the instrument	0.1859
Using a dummy B as an instrument[a]	0.1549
Policy relevant treatment effect (PRTE)	0.1549

Source: Heckman and Vytlacil (2005).

[a] The dummy B is such that $B = 1$ if an individual belongs to a randomly assigned eligible population, 0 otherwise.

The IV weight for this estimator is a special case of Equation (4.13):

$$\omega_{\mathrm{IV}}(u_D \mid B) = \frac{E(B - E(B) \mid \hat{P}(Z) \geqslant u_D)\Pr(\hat{P}(Z) \geqslant u_D)}{\operatorname{Cov}(B, \hat{P}(Z))},$$

where $\hat{P}(Z) = P(Z(1+t\mathbf{1}[Z > 0]))^B P(Z)^{(1-B)}$. Here, the IV is eligibility for a policy and IV is equivalent to a social experiment that identifies the mean gain per participant who switches to participation in the program. It is to be expected that this IV weight and ω_{PRTE} are identical.

4.7.1. Further examples within the extended Roy model

To gain a further understanding of how to construct the weights, and to understand how negative weights can arise, it is useful to return to the policy adoption model presented at the end of Section 2. The only unobservables in this model are in the outcome equations. To simplify the analysis, we use an extended Roy model where the only unobservables are the unmeasured gains.

In this framework, the cost C of adopting the policy is the same across all countries. Countries choose to adopt the policy if $D^* > 0$ where D^* is the net benefit of adoption: $D^* = (Y_1 - Y_0 - C)$ and ATE $= E(\beta) = E(Y_1 - Y_0) = \mu_1 - \mu_0$, while treatment on the treated is $E(\beta \mid D = 1) = E(Y_1 - Y_0 \mid D = 1) = \mu_1 - \mu_0 + E(U_1 - U_0 \mid D = 1)$.

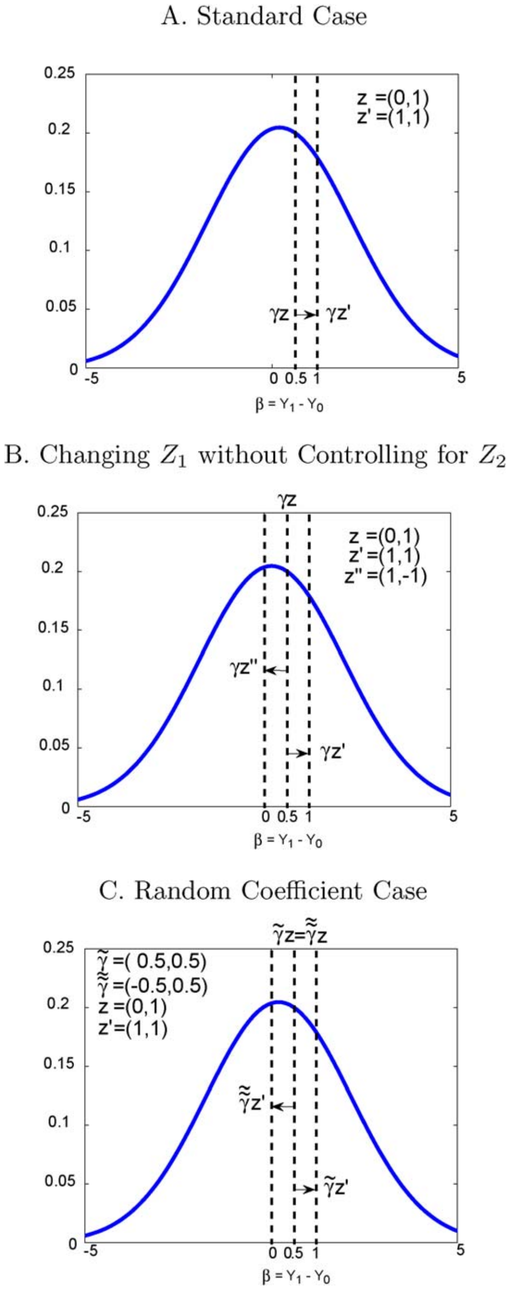
In this setting, the gross return to the country at the margin is C, i.e., $E(Y_1 - Y_0 \mid D^* = 0) = E(Y_1 - Y_0 \mid Y_1 - Y_0 = C) = C$. Recall that Figure 1 presents the standard treatment parameters for the values of the choice parameter presented at the base of the figure. Countries that adopt the policy are above average. In a model where the cost varies (the generalized Roy model with $U_C \neq 0$), and C is negatively correlated with the gain, adopting countries could be below average.[66] We consider cases with discrete instruments and cases with continuous instruments. We first turn to the discrete case.

4.7.2. Discrete instruments and weights for LATE

Consider what instrumental variables identify in the model of country policy adoption presented below Figure 5. That figure presents three cases that we analyze in this section. Let cost $C = Z\gamma$ where instrument $Z = (Z_1, Z_2)$. Higher values of Z reduce the probability of adopting the policy if $\gamma \geqslant 0$, component by component.

Consider the "standard" case depicted in Figure 5A. Increasing both components of discrete-valued Z raises costs and hence raises the benefit observed for the country at the margin by eliminating adoption in low return countries. It also reduces the probability that countries adopt the policy. In general a different country is at the margin when different instruments are used.

[66] See, e.g., Heckman (1976a, 1976c) and Willis and Rosen (1979).



A. Standard Case

B. Changing Z_1 without Controlling for Z_2

C. Random Coefficient Case

Figure 5. Monotonicity: The extended Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

Outcomes

$$Y_1 = \alpha + \bar{\beta} + U_1$$
$$Y_0 = \alpha + U_0$$

Choice model

$$D = \begin{cases} 1 & \text{if } Y_1 - Y_0 - \gamma Z \geqslant 0, \\ 0 & \text{if } Y_1 - Y_0 - \gamma Z < 0 \end{cases}$$

with $\gamma Z = \gamma_1 Z_1 + \gamma_2 Z_2$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \boldsymbol{\Sigma}),\ \ \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix},\ \alpha = 0.67,\ \bar{\beta} = 0.2,\ \gamma = (0.5, 0.5) \text{ (except in Case C)}$$

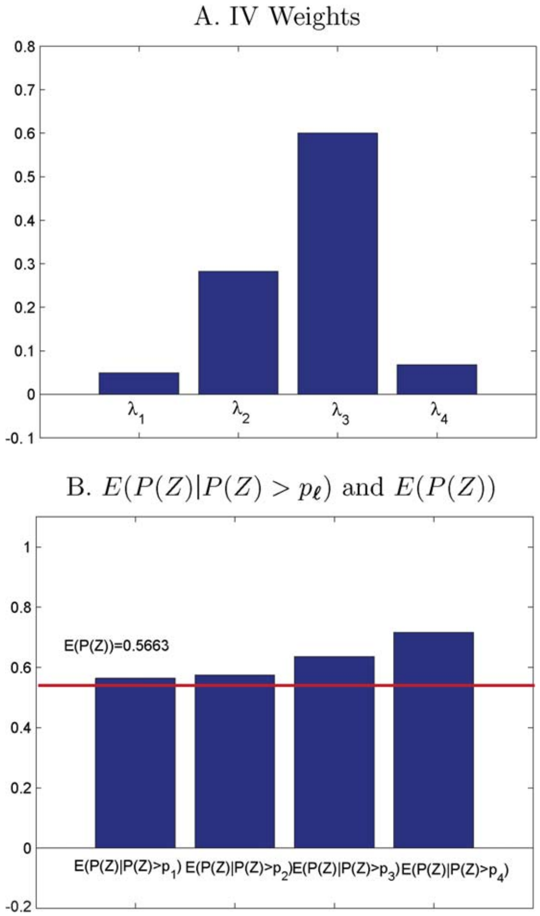
$$Z_1 = \{-1, 0, 1\} \text{ and } Z_2 = \{-1, 0, 1\}$$

A. Standard case	B. Changing Z_1 without controlling for Z_2	C. Random coefficient case
$z \to z'$ $z = (0, 1)$ and $z' = (1, 1)$	$z \to z'$ or $z \to z''$ $z = (0, 1), z' = (1, 1)$ and $z'' = (1, -1)$	$z \to z'$ $z = (0, 1)$ and $z' = (1, 1)$
		γ is a random vector $\tilde{\gamma} = (0.5, 0.5)$ and $\tilde{\tilde{\gamma}} = (-0.5, 0.5)$ where $\tilde{\gamma}$ and $\tilde{\tilde{\gamma}}$ are two realizations of γ
$D(\gamma z) \geqslant D(\gamma z')$	$D(\gamma z) \geqslant D(\gamma z')$ or $D(\gamma z) < D(\gamma z'')$	$D(\tilde{\tilde{\gamma}} z) \geqslant D(\tilde{\tilde{\gamma}} z')$ and $D(\tilde{\gamma} z) < D(\tilde{\gamma} z')$
For all individuals	Depending on the value of z' or z''	Depending on value of γ

Figure 5. (*Continued*)

Figure 6A plots the weights and Figure 6B plots the components of the weights for the LATE values using $P(Z)$ as an instrument for the distribution of discrete Z values shown at the base of the figure. Figure 6C presents the LATE parameter derived using $P(Z)$ as an instrument. The weights are positive as predicted from Equation (4.5) when $J(Z) = P(Z)$. Thus, the monotonicity condition for the weights in terms of u_D is satisfied. The outcome and choice parameters are the same as those used to generate Figures 1 and 5. The LATE parameters for each interval of P values are presented in a table just below the figures. There are four LATE parameters corresponding to the five distinct values of the propensity score for that value. The LATE parameters exhibit the declining pattern with u_D predicted by the Roy model.

A case producing negative weights is depicted in Figure 5B. In that graph, the same Z is used to generate the choices as is used to generate Figure 1B. However, in this case, the analyst uses Z_1 as the instrument. Z_1 and Z_2 are negatively dependent and $E(Z_1 \mid P(Z) > u_D)$ is not monotonic in u_D. This nonmonotonicity is evident in Figure 7B. It produces the pattern of negative weights shown in Figure 7A. These are associated with two way flows. Increasing Z_1 controlling for Z_2 reduces the probability of country policy adoption. However, we do not condition on Z_2 in constructing this figure. Z_2 is floating. Two way flows are induced by uncontrolled variation in Z_2. For some units, the strength of the associated variation in Z_2 offsets the increase in Z_1 and for other units it does not. Observe that the LATE parameters defined using $P(Z)$ are the same in both examples. They are just weighted differently. We discuss the random coefficient choice model generating Figure 5C in Section 4.10.


A. IV Weights

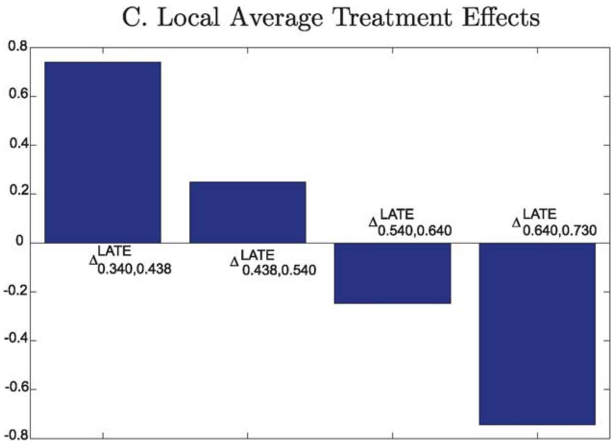
B. $E(P(Z)|P(Z) > p_{\ell})$ and $E(P(Z))$

Figure 6. IV weights and its components under discrete instruments when $P(Z)$ is the instrument, the extended Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

The IV estimator does not identify ATE, TT or TUT (given at the bottom of Figure 6C). Conditioning on Z_2 produces positive weights. This is illustrated in the weights shown in Table 6 that condition on Z_2 using the same model that generated Figure 6. Conditioning on Z_2 effectively converts the problem back into one with a scalar instrument and the weights are positive for that case.

From Yitzhaki's analysis, for any sample size, a regression of Y on P identifies a weighted average of slopes based on ordered regressors:

$$\frac{E(Y_\ell \mid p_\ell) - E(Y_{\ell-1} \mid p_{\ell-1})}{p_\ell - p_{\ell-1}},$$



The model is the same as the one presented below Figure 5.

$$\text{ATE} = 0.2,\ \text{TT} = 0.5942,\ \text{TUT} = -0.4823 \text{ and } \Delta^{\text{IV}}_{P(Z)} = \sum_{\ell=1}^{K-1} \Delta^{\text{LATE}}(p_\ell, p_{\ell+1})\lambda_\ell = -0.09$$

$$\begin{aligned}\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) &= \frac{E(Y \mid P(Z) = p_{\ell+1}) - E(Y \mid P(Z) = p_\ell)}{p_{\ell+1} - p_\ell} \\ &= \frac{\bar{\beta}(p_{\ell+1} - p_\ell) + \sigma_{U_1 - U_0}(\phi(\Phi^{-1}(1 - p_{\ell+1})) - \phi(\Phi^{-1}(1 - p_\ell)))}{p_{\ell+1} - p_\ell}\end{aligned}$$

$$\begin{aligned}\lambda_\ell &= (p_{\ell+1} - p_\ell)\frac{\sum_{i=1}^{K}(p_i - E(P(Z)))\sum_{t>\ell}^{K} f(p_i, p_t)}{\operatorname{Cov}(Z_1, D)} \\ &= (p_{\ell+1} - p_\ell)\frac{\sum_{t>\ell}^{K}(p_t - E(P(Z)))f(p_t)}{\operatorname{Cov}(Z_1, D)}\end{aligned}$$

Joint probability distribution of (Z_1, Z_2) and the propensity score (joint probabilities in ordinary type $(\Pr(Z_1 = z_1, Z_2 = z_2))$; propensity score in italics $(\Pr(D = 1 \mid Z_1 = z_1, Z_2 = z_2))$)

$Z_1 \backslash Z_2$	-1	0	1
-1	0.02	0.02	0.36
	0.7309	*0.6402*	*0.5409*
0	0.3	0.01	0.03
	0.6402	*0.5409*	*0.4388*
1	0.2	0.05	0.01
	0.5409	*0.4388*	*0.3408*

$\operatorname{Cov}(Z_1, Z_2) = -0.5468$

Figure 6. (*Continued*)

where $p_\ell > p_{\ell-1}$ and the weights are the positive Yitzhaki–Imbens–Angrist weights derived in Yitzhaki (1989, 1996) or in Yitzhaki and Schechtman (2004). The weights are positive whether or not monotonicity condition (IV-3) holds. If monotonicity holds, IV is a weighted average of LATEs. Otherwise it is just a weighted average of ordered

Table 6
The conditional instrumental variable estimator ($\Delta^{\mathrm{IV}}_{Z_1|Z_2=z_2}$) and conditional local average treatment effect ($\Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1} \mid Z_2 = z_2)$) when Z_1 is the instrument (given $Z_2 = z_2$)

The extended Roy economy			
	$Z_2 = -1$	$Z_2 = 0$	$Z_2 = 1$
$P(-1, Z_2) = p_3$	0.7309	0.6402	0.5409
$P(0, Z_2) = p_2$	0.6402	0.5409	0.4388
$P(1, Z_2) = p_1$	0.5409	0.4388	0.3408
λ_1	0.8418	0.5384	0.2860
λ_2	0.1582	0.4616	0.7140
$\Delta^{\mathrm{LATE}}(p_1, p_2)$	−0.2475	0.2497	0.7470
$\Delta^{\mathrm{LATE}}(p_2, p_3)$	−0.7448	−0.2475	0.2497
$\Delta^{\mathrm{IV}}_{Z_1\|Z_2=z_2}$	−0.3262	0.0202	0.3920

The model is the same as the one presented below Figure 2.

$$\Delta^{\mathrm{IV}}_{Z_1|Z_2=z_2} = \sum_{\ell=1}^{I-1} \Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1} \mid Z_2 = z_2)\lambda_{\ell|Z_2=z_2} = \sum_{\ell=1}^{I-1} \Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1} \mid Z_2 = z_2)\lambda_{\ell|Z_2=z_2}$$

$$\Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1} \mid Z_2 = z_2) = \frac{E(Y \mid P(Z) = p_{\ell+1}, Z_2 = z_2) - E(Y \mid P(Z) = p_\ell, Z_2 = z_2)}{p_{\ell+1} - p_\ell}$$

$$\lambda_{\ell|Z_2=z_2} = (p_{\ell+1} - p_\ell)\frac{\sum_{i=1}^{I}(z_{1,i} - E(Z_1 \mid Z_2 = z_2))\sum_{t>\ell}^{I} f(z_{1,i}, p_t \mid Z_2 = z_2)}{\mathrm{Cov}(Z_1, D)}$$
$$= (p_{\ell+1} - p_\ell)\frac{\sum_{t>\ell}^{I}(z_{1,t} - E(Z_1 \mid Z_2 = z_2))f(z_{1,t}, p_t \mid Z_2 = z_2)}{\mathrm{Cov}(Z_1, D)}$$

Probability distribution of Z_1 conditional on Z_2 ($\Pr(Z_1 = z_1 \mid Z_2 = z_2)$)

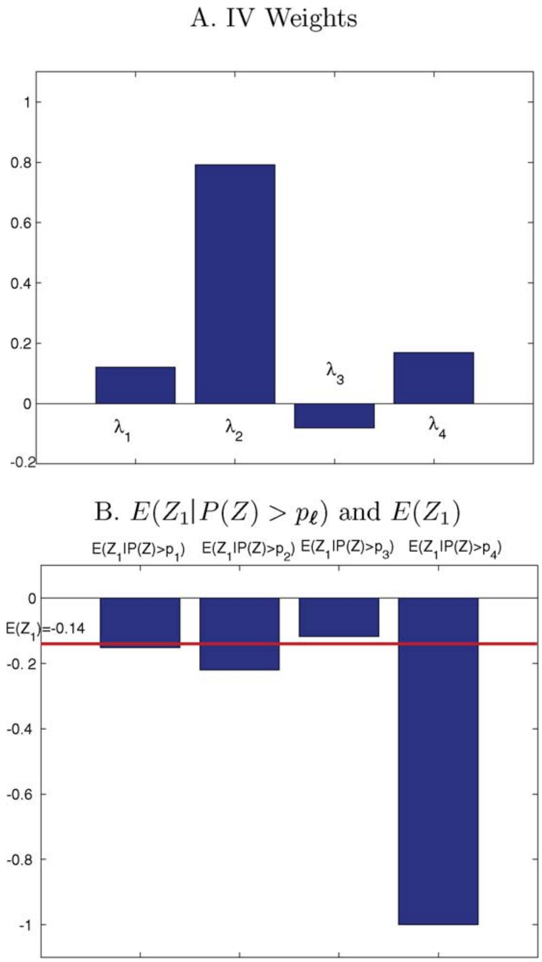
z_1	$\Pr(Z_1 = z_1 \mid Z_2 = -1)$	$\Pr(Z_1 = z_1 \mid Z_2 = 0)$	$\Pr(Z_1 = z_1 \mid Z_2 = 1)$
−1	0.0385	0.25	0.9
0	0.5769	0.125	0.075
1	0.3846	0.625	0.025

Source: Heckman, Urzua and Vytlacil (2006).

(by p_ℓ) estimators consistent with two-way flows. We next discuss continuous instruments.

4.7.3. Continuous instruments

For the case of continuous Z, we present a parallel analysis for the weights associated with the MTE. Figure 8 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 I and II, $\beta \perp\!\!\!\perp D$,


A. IV Weights

B. $E(Z_1|P(Z) > p_\ell)$ and $E(Z_1)$

The model is the same as the one presented below Figure 5. The values of the treatment parameters are the same as the ones presented below Figure 6.

Figure 7. IV weights and its components under discrete instruments when Z_1 is the instrument, the extended Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

so $\Delta^{\text{MTE}}(u_D)$ is constant in u_D. In case I, this is trivial since β is a constant. In case II, β is random but selection into D does not depend on β. Case III is the model with essential heterogeneity ($\beta \not\perp\!\!\!\perp D$). The graph (Figure 8A) depicts $E(Y \mid P(Z))$ in the three cases. Cases I and II make $E(Y \mid P(Z))$ linear in $P(Z)$. Case III is nonlinear in $P(Z)$. This arises when $\beta \not\perp\!\!\!\perp D$. The derivative of $E(Y \mid P(Z))$ is presented in Figure 8B. It is a constant for cases I and II (flat MTE) but declining in $U_D = P(Z)$ for the case

$$\Delta^{\mathrm{IV}}_{Z_1} = \sum_{\ell=1}^{K-1} \Delta^{\mathrm{LATE}}(p_\ell, p_{\ell+1})\lambda_\ell = 0.1833$$

$$\lambda_\ell = (p_{\ell+1} - p_\ell)\frac{\sum_{i=1}^{I}(z_{1,i} - E(Z_1))\sum_{t>\ell}^{K} f(z_{1,i}, p_t)}{\mathrm{Cov}(Z_1, D)}$$

Joint probability distribution of (Z_1, Z_2) and the propensity score (joint probabilities in ordinary type $(\Pr(Z_1 = z_1, Z_2 = z_2))$; propensity score in italics $(\Pr(D = 1 \mid Z_1 = z_1, Z_2 = z_2)))$

$Z_1 \backslash Z_2$	-1	0	1
-1	0.02	0.02	0.36
	0.7309	*0.6402*	*0.5409*
0	0.3	0.01	0.03
	0.6402	*0.5409*	*0.4388*
1	0.2	0.05	0.01
	0.5409	*0.4388*	*0.3408*

$\mathrm{Cov}(Z_1, Z_2) = -0.5468$

Figure 7. (*Continued*)

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 I and II ($\beta \perp\!\!\!\perp D$) or case III ($\beta \not\perp\!\!\!\perp D$).[67] These cases are the extended Roy counterparts to $E(Y \mid P(Z) = p)$ and MTE shown for the generalized Roy model in Figures 3A and 3B.

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 8 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 ($P(Z)$).

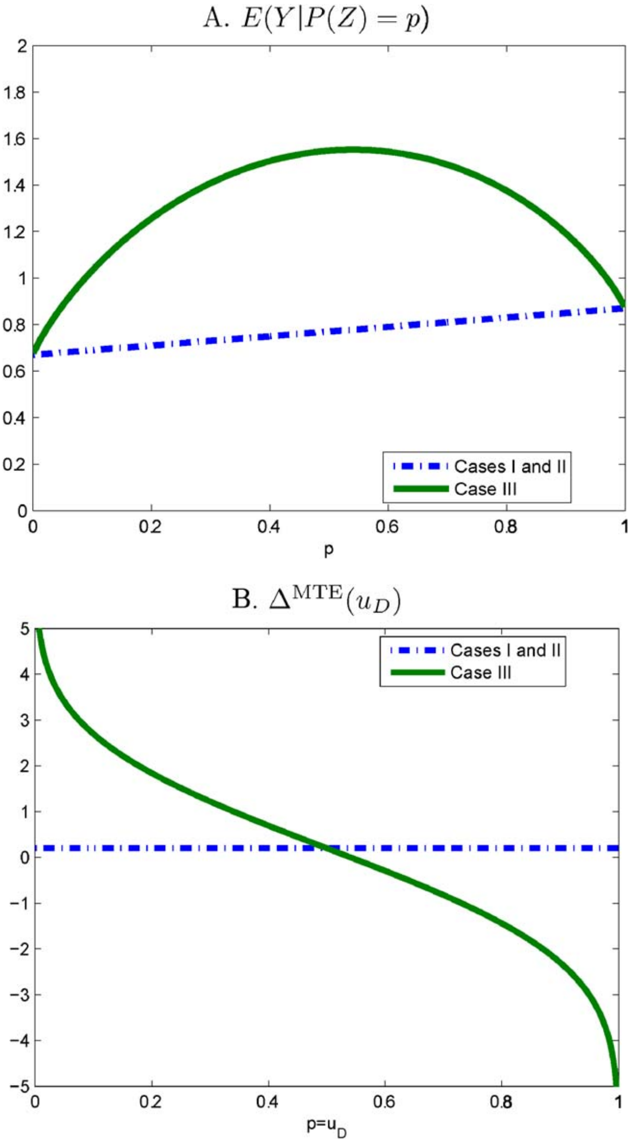
Figure 9 plots MTE and LATE for different intervals of u_D using the model generating Figure 8. LATE is the chord of $E(Y \mid P(Z))$ evaluated at different points. The relationship between LATE and MTE is depicted in Figure 9B. LATE is the integral under the MTE curve divided by the difference between the upper and lower limits.

The treatment parameters associated with case III are plotted in Figure 10. The MTE is the same as that presented in Figure 8. ATE has the same value for all p. The effect of treatment on the treated for $P(Z) = p$, $\Delta^{\mathrm{TT}}(p) = E(Y_1 - Y_0 \mid D = 1, P(Z) = p)$ declines in p (equivalently it declines in u_D). Treatment on the untreated given p, $\mathrm{TUT}(p) = \Delta^{\mathrm{TUT}}(p) = E(Y_1 - Y_0 \mid D = 0, P(Z) = p)$ also declines in p,

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

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

[67] Recall that we keep the conditioning on X implicit.



A. $E(Y|P(Z)=p)$

B. $\Delta^{\mathrm{MTE}}(u_D)$

Figure 8. Conditional expectation of Y on $P(Z)$ and the MTE, the extended Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

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

Matching on $P = p$ (which is equivalent to nonparametric regression given $P = p$) produces a biased estimator of $\mathrm{TT}(p)$. Matching assumes a flat MTE (average return

Outcomes	Choice model
$Y_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \alpha + U_0$	$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{if } D^* < 0 \end{cases}$

Case I	Case II	Case III
$U_1 = U_0$ $\bar{\beta} = \text{ATE} = \text{TT} = \text{TUT} = \text{IV}$	$U_1 - U_0 \perp\!\!\!\perp D$ $\bar{\beta} = \text{ATE} = \text{TT} = \text{TUT} = \text{IV}$	$U_1 - U_0 \not\perp\!\!\!\perp D$ $\bar{\beta} = \text{ATE} \neq \text{TT} \neq \text{TUT} \neq \text{IV}$

Parameterization

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

Figure 8. (*Continued*)

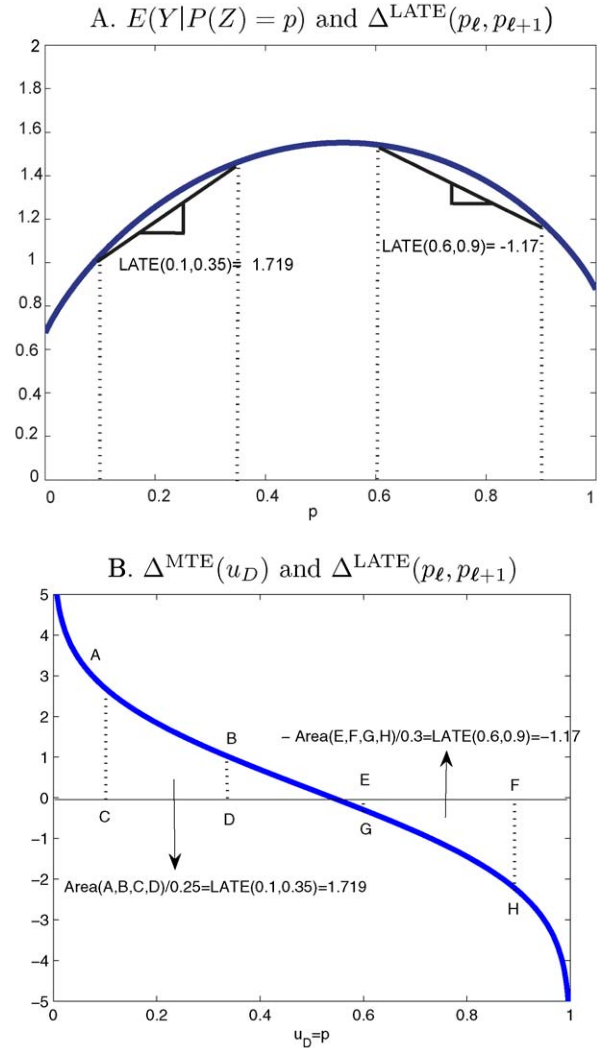
equals marginal return).[68] Therefore it is systematically biased for $\Delta^{\text{TT}}(p)$ in a model with essential heterogeneity. Making observables alike makes the unobservables dissimilar. Holding p constant across treatment and control groups understates $\text{TT}(p)$ for low values of p and overstates it for high values of p. We develop this point further when we discuss matching in Section 8.

Figure 11 plots the MTE (as a function of u_D where $u_D = F_V(v)$), the weights for ATE, TT and TUT and the IV weights using Z_1 as the instrument for the model used to generate Figure 9. The distribution of the Z is assumed to be normal with generating parameters given at the base of Figure 9. The IV weight for normal Z is always nonnegative even if we use only one coordinate of vector Z. This is a consequence of the monotonicity of $E(Z_j \mid P(Z) \geqslant u_D)$ in u_D for any component of vector Z, which is a property of normal selection models.[69]

Panel A of Figure 11 plots the treatment weights derived by Heckman and Vytlacil (1999, 2001b) and the IV weight (4.14), along with the MTE. The ATE = Δ^{ATE} weight is flat (= 1). TT oversamples the low u_D agents (those more likely to adopt the policies). TUT oversamples the high u_D agents. The IV weight is positive as it must be when the Z are normally distributed. IV is far from any of the standard treatment parameters. Panel B decomposes the weight into its numerator components $E(Z_1 \mid P(Z) \geqslant u_D)$ and $E(Z_1)$, and the weight itself. The difference $E(Z_1 \mid P(Z) \geqslant u_D) - E(Z_1)$ multiplied by $\Pr(P(Z) \geqslant u_D)$ and normalized by $\text{Cov}(Z_1, D)$ is the weight (see Equation (4.13)). The weight is plotted as the dotted line in Figure 9B.

[68] See Heckman and Vytlacil (2005) and Section 8.

[69] See Heckman and Honoré (1990). In a broad class of models [see, e.g., Heckman, Tobias and Vytlacil (2003)] $E(R \mid S > c)$ is monotonic in c for vector R. The normal model is one member of this family.



A. $E(Y|P(Z)=p)$ and $\Delta^{\text{LATE}}(p_{\ell}, p_{\ell+1})$

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

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

$$\Delta^{\text{LATE}}(0.6, 0.9) = -1.17$$

$$\Delta^{\text{LATE}}(0.1, 0.35) = 1.719$$

Outcomes

$$Y_1 = \alpha + \bar{\beta} + U_1$$
$$Y_0 = \alpha + U_0$$

Choice model

$$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{if } D^* < 0 \end{cases}$$

with $D^* = Y_1 - Y_0 - \gamma Z$

Figure 9. The local average treatment effect, the extended Roy economy. *Source*: Heckman, Urzua and Vytlacil (2006).

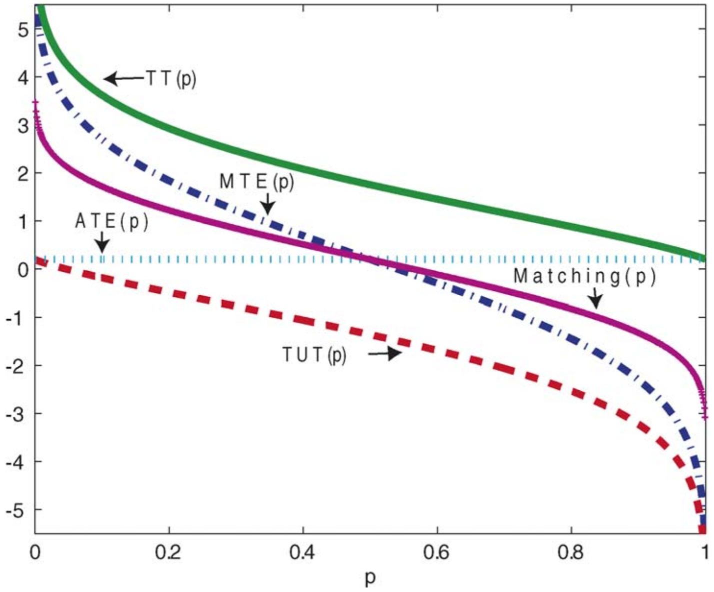
Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \text{ and } Z \sim N(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \boldsymbol{\mu}_Z = (2, -2) \text{ and } \boldsymbol{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$$

$$\alpha = 0.67, \ \bar{\beta} = 0.2, \ \gamma = (0.5, 0.5)$$

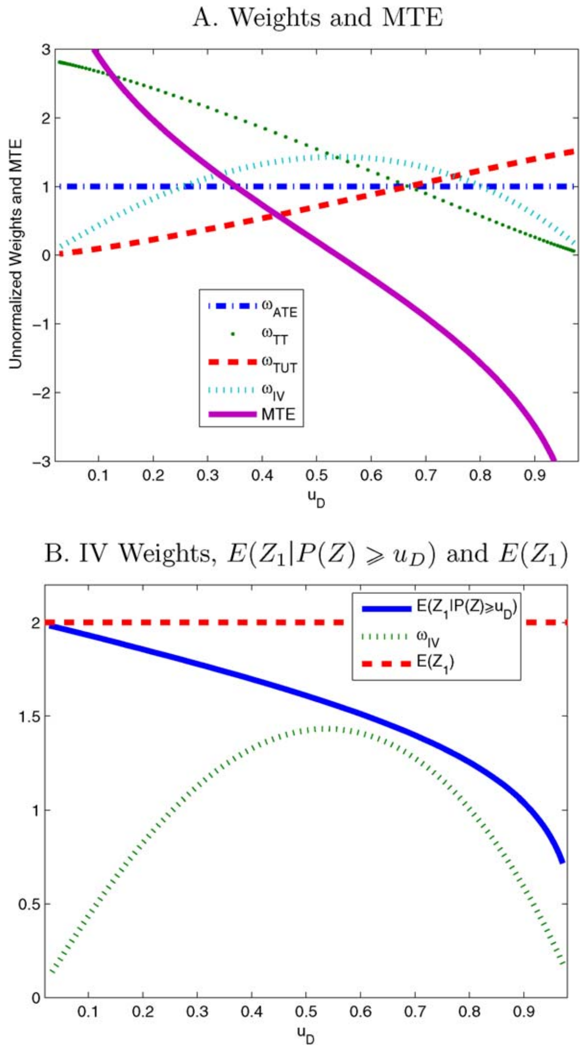
Figure 9. (*Continued*)



Parameter	Definition	Under assumptions (*)
Marginal treatment effect	$E[Y_1 - Y_0 \mid D^* = 0, P(Z) = p]$	$\bar{\beta} + \sigma_{U_1-U_0}\Phi^{-1}(1-p)$
Average treatment effect	$E[Y_1 - Y_0 \mid P(Z) = p]$	$\bar{\beta}$
Treatment on the treated	$E[Y_1 - Y_0 \mid D^* \geqslant 0, P(Z) = p]$	$\bar{\beta} + \sigma_{U_1-U_0}\frac{\phi(\Phi^{-1}(1-p))}{p}$
Treatment on the untreated	$E[Y_1 - Y_0 \mid D^* < 0, P(Z) = p]$	$\bar{\beta} - \sigma_{U_1-U_0}\frac{\phi(\Phi^{-1}(1-p))}{1-p}$
OLS/Matching on $P(Z)$	$E[Y_1 \mid D^* \geqslant 0, P(Z) = p]$ $- E[Y_0 \mid D^* < 0, P(Z) = p]$	$\bar{\beta} + \left(\frac{\sigma_{U_1}^2 - \sigma_{U_1,U_0}}{\sqrt{\sigma_{U_1-U_0}}}\right)\left(\frac{1-2p}{p(1-p)}\right)\phi(\Phi^{-1}(1-p))$

(*): The model in this case is the same as the one presented below Figure 9.
Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively.
$\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.

Figure 10. Treatment parameters and OLS/Matching as a function of $P(Z) = p$.
Source: Heckman, Urzua and Vytlacil (2006).



A. Weights and MTE

B. IV Weights, $E(Z_1|P(Z) \geqslant u_D)$ and $E(Z_1)$

Parameter	Under assumptions (*)
ATE	0.2
TT	1.1878
TUT	−0.9132
IV_{Z_1}	0.0924

(*) The model in this case is the same as the one presented below Figure 9.

Figure 11. Treatment weights, IV weights using Z_1 as the instrument and the MTE. *Source*: Heckman, Urzua and Vytlacil (2004).

Suppose that instead of assuming normality for the regressors, instrument Z is assumed to be a random vector with a distribution function given by a mixture of two normals:

$$Z \sim P_1 N(\kappa_1, \Sigma_1) + P_2 N(\kappa_2, \Sigma_2),$$

where P_1 is the proportion in population 1, P_2 is the proportion in population 2, and $P_1 + P_2 = 1$. This produces a model with continuous instruments, where $E(\tilde{J}(Z) \mid P(Z) \geqslant u_D)$ need not be monotonic in u_D where $\tilde{J}(Z) = J(Z) - E(J(Z))$. Such a data generating process for the instrument could arise from an ecological model in which two different populations are mixed (e.g., rural and urban populations).[70]

Appendix E derives the instrumental variable weights on Δ^{MTE} when Z_1 (the first element of Z) is used as the instrument, i.e., $J(Z) = Z_1$. For simplicity, we assume that there are no X regressors. The probability of selection is generated using $\mu_D(Z) = Z\gamma$. The joint distribution of $(Z_1, Z\gamma)$ is normal within each group.

In our example, the dependence between Z_1 and $Z\gamma$ $(= F_V(Z\gamma) = P(Z))$ is negative in one population and positive in another. Thus in one population, as Z_1 increases $P(Z)$ increases. In the other population, as Z_1 increases $P(Z)$ decreases. If this second population is sufficiently big (P_1 is small) or the negative correlation in the second population is sufficiently big, the weights can become negative because $E(\tilde{J}(Z) \mid P(Z) \geqslant u_D)$ is not monotonic in u_D.

We present examples for a conventional normal outcome selection model generated by the parameters presented at the base of Figure 12. The discrete choice equation is a conventional probit: $\Pr(D = 1 \mid Z = z) = \Phi(\frac{z\gamma}{\sigma_V})$. The outcome equations are linear normal equations. Thus $\Delta^{\text{MTE}}(v) = E(Y_1 - Y_0 \mid V = v)$, is linear in v:

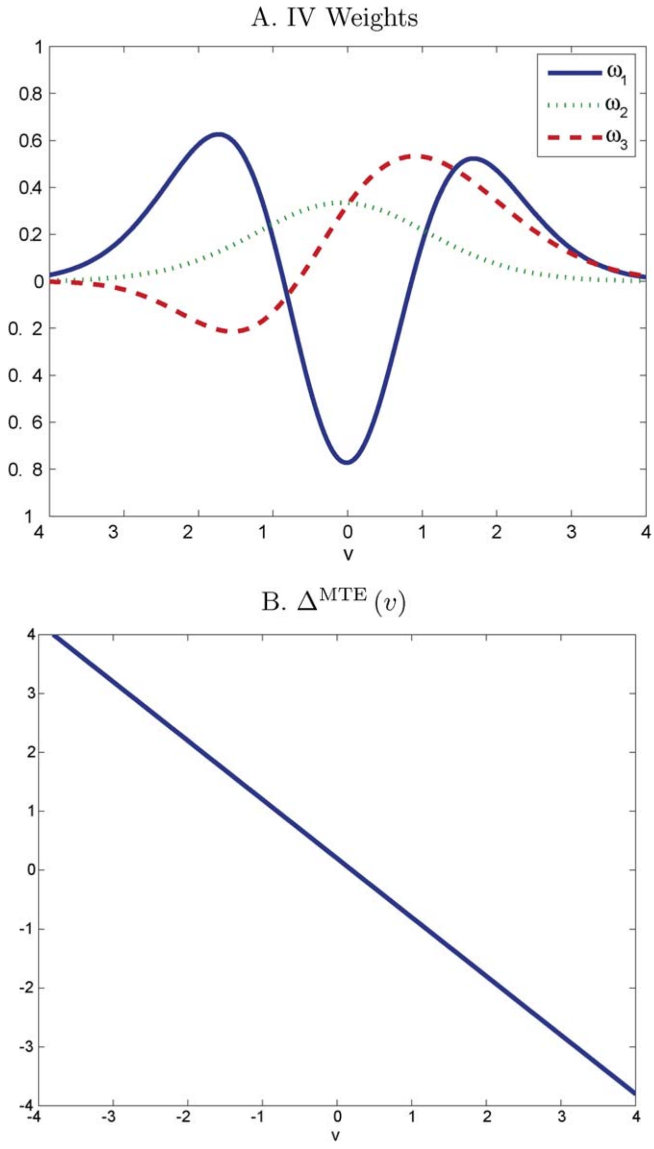
$$E(Y_1 - Y_0 \mid V = v) = \mu_1 - \mu_0 + \frac{\text{Cov}(U_1 - U_0, V)}{\text{Var}(V)} v.$$

At the base of the figure, we define $\bar{\beta} = \mu_1 - \mu_0$ and $\alpha = \mu_0$. The average treatment effects are the same for all different distributions of the Z.

In each of the following examples, we show results for models with vector Z that satisfies (IV-1) and (IV-2) and with $\gamma > 0$ componentwise where γ is the coefficient of Z in the cost equation. We vary the weights and means of the instruments. *Ceteris paribus*, an increase in each component of Z increases $\Pr(D = 1 \mid Z = z)$. Table 7 presents the parameters treatment on the treated $(E(Y_1 - Y_0 \mid D = 1))$, treatment on the untreated $(E(Y_1 - Y_0 \mid D = 0))$, and the average treatment effect $(E(Y_1 - Y_0))$ produced by our model for different distributions of the regressors.

In standard IV analysis, under assumptions (IV-1) and (IV-2) the distribution of Z does not affect the probability limit of the IV estimator. It only affects its sampling distribution. Figure 12A shows three weights corresponding to the perturbations of the variances of the instruments in the second component population Σ_2 and the means

[70] Observe that $E(Z) = P_1\kappa_1 + P_2\kappa_2$.


A. IV Weights

B. $\Delta^{\mathrm{MTE}}(v)$

Figure 12. MTE and IV weights using Z_1 as the instrument when $Z = (Z_1, Z_2) \sim p_1 N(\kappa_1, \Sigma_1) + p_2 N(\kappa_2, \Sigma_2)$ for different values of Σ_2.
Source: Heckman, Urzua and Vytlacil (2006).

(κ_1, κ_2) shown at the table at the base of the figure. The Δ_V^{MTE} used in all of our examples are plotted in Figure 12B. The MTE has the familiar shape, reported in Heckman

Outcomes	Choice model
$Y_1 = \alpha + \bar{\beta} + U_1$	$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{if } D^* < 0 \end{cases}$
$Y_0 = \alpha + U_0$	$D^* = Y_1 - Y_0 - \gamma Z$ and $V = -(U_1 - U_0)$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \alpha = 0.67, \bar{\beta} = 0.2$$

$$Z = (Z_1, Z_2) \sim p_1 N(\kappa_1, \Sigma_1) + p_2 N(\kappa_2, \Sigma_2)$$

$$p_1 = 0.45, p_2 = 0.55; \Sigma_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{bmatrix}$$

$$\mathrm{Cov}(Z_1, \gamma Z) = \gamma \Sigma_1^1 = 0.98; \gamma = (0.2, 1.4)$$

Figure 12. (*Continued*)

Table 7
The IV estimator and $\mathrm{Cov}(Z_2, \gamma Z)$ associated with each value of Σ_2

Weights	Σ_2	κ_1	κ_2	IV	ATE	TT	TUT	$\mathrm{Cov}(Z_2, \gamma Z) = \gamma \Sigma_2^1$
ω_1	$\begin{bmatrix} 0.6 & -0.5 \\ -0.5 & 0.6 \end{bmatrix}$	[0 0]	[0 0]	0.434	0.2	1.401	−1.175	−0.58
ω_2	$\begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.6 \end{bmatrix}$	[0 0]	[0 0]	0.078	0.2	1.378	−1.145	0.26
ω_3	$\begin{bmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{bmatrix}$	[0 −1]	[0 1]	−2.261	0.2	1.310	−0.859	−0.30

Source: Heckman, Urzua and Vytlacil (2006).

(2001) and Heckman, Tobias and Vytlacil (2003) that returns are highest for those with values of v that make them more likely to get treatment (i.e., low values of v).

The weights ω_1 and ω_3 plotted in Figure 12A correspond to the case where $E(Z_1 - E(Z_1) \mid P(Z) \geqslant u_D)$ is not monotonic in u_D. In these cases, the sign of the covariance between Z_1 and $Z\gamma$ (i.e., $P(Z)$) is not the same in the two subpopulations. The IV estimates reported in the table at the base of the figure range all over the place even though the parameters of the outcome and choice model are the same.[71]

Different distributions of Z critically affect the probability limit of the IV estimator in the model of essential heterogeneity. The model of outcomes and choices is the same across all of these examples. The MTE and ATE parameters are the same. Only the distribution of the instrument differs. The instrumental variable estimand is sometimes positive and sometimes negative, and oscillates wildly in magnitude depending on the distribution of the instruments. The estimated "effect" is often way off the mark for any

[71] Since TT and TUT depend on the distribution of $P(Z)$, they are not invariant to changes in the distribution of the Z.

desired treatment parameter. These examples show how uniformity in Z does not translate into uniformity in $J(Z)$ (Z_1 in this example). This sensitivity is a phenomenon that does not appear in the conventional homogeneous response model but is a central feature of a model with essential heterogeneity.[72] We now compare selection and IV models.

4.8. Comparing selection and IV models

We now show that local IV identifies the derivatives of a selection model. Making the X explicit, in the standard selection model, U_1 and U_0 are scalar random variables that are additively separable in the outcome equations, $Y_1 = \mu_1(X) + U_1$ and $Y_0 = \mu_0(X) + U_0$. The control function approach conditions on Z and D. As a consequence of index sufficiency, this is equivalent to conditioning on $P(Z)$ and D:

$$\begin{aligned} E(Y \mid X, D, Z) = \mu_0(X) &+ \bigl[\mu_1(X) - \mu_0(X)\bigr]D \\ &+ K_1\bigl(P(Z), X\bigr)D + K_0\bigl(P(Z), X\bigr)(1 - D), \end{aligned}$$

where the control functions are

$$\begin{aligned} K_1\bigl(P(Z), X\bigr) &= E\bigl(U_1 \bigm| D = 1, X, P(Z)\bigr), \\ K_0\bigl(P(Z), X\bigr) &= E\bigl(U_0 \bigm| D = 0, X, P(Z)\bigr). \end{aligned}$$

The IV approach does not condition on D. It works with

$$\begin{aligned} E(Y \mid X, Z) = \mu_0(X) &+ \bigl[\mu_1(X) - \mu_0(X)\bigr]P(Z) + K_1\bigl(P(Z), X\bigr)P(Z) \\ &+ K_0\bigl(P(Z), X\bigr)\bigl(1 - P(Z)\bigr), \end{aligned} \tag{4.17}$$

the population mean outcome given X, Z.

From index sufficiency, $E(Y \mid X, Z) = E(Y \mid X, P(Z))$. The MTE is the derivative of this expression with respect to $P(Z)$, which we have defined as LIV:

$$\left.\frac{\partial E(Y \mid X, P(Z))}{\partial P(Z)}\right|_{P(Z)=p} = \mathrm{LIV}(X, p) = \mathrm{MTE}(X, p).[73]$$

The distribution of $P(Z)$ and the relationship between $J(Z)$ and $P(Z)$ determine the weight on MTE.[74] Under assumptions (A-1)–(A-5), along with rank and limit conditions [Heckman and Robb (1985a), Heckman (1990)], one can identify $\mu_1(X)$, $\mu_0(X)$, $K_1(P(Z), X)$, and $K_0(P(Z), X)$.

[72] We note parenthetically that if we assume $P_1 = 0$ (or $P_2 = 0$), the weights are positive even if we only use Z_1 as an instrument and Z_1 and Z_2 are negatively correlated. This follows from the monotonicity of $E(R \mid S > c)$ in c for vector R. See Heckman and Honoré (1990). This case is illustrated in Figure 11.

[73] Björklund and Moffitt (1987) analyze this marginal effect for a parametric generalized Roy model.

[74] Because LIV does not condition on D, it discards information. Lost in taking derivatives are the constants in the model that do not interact with $P(Z)$ in Equation (4.17).

The selection (control function) estimator identifies the conditional means

$$E\big(Y_1 \mid X, P(Z), D = 1\big) = \mu_1(X) + K_1\big(X, P(Z)\big) \tag{4.18a}$$

and

$$E\big(Y_0 \mid X, P(Z), D = 0\big) = \mu_0(X) + K_0\big(X, P(Z)\big). \tag{4.18b}$$

These can be identified from nonparametric regressions of Y_1 and Y_0 on X, Z in each population. To decompose these means and separate $\mu_1(X)$ from $K_1(X, P(Z))$ without invoking functional form or curvature assumptions, it is necessary to have an exclusion (a Z not in X).[75] In addition, there must exist a limit set for Z given X such that $K_1(X, P(Z)) = 0$ for Z in that limit set. Otherwise, without functional form or curvature assumptions, it is not possible to disentangle $\mu_1(X)$ from $K_1(X, P(Z))$ which may contain constants and functions of X that do not interact with $P(Z)$ [see Heckman (1990)]. A parallel argument for Y_0 shows that we require a limit set for Z given X such that $K_0(X, P(Z)) = 0$. Selection models operate by identifying the components of (4.18a) and (4.18b) and generating the treatment parameters from these components. Thus they work with levels of the Y.

The local IV method works with derivatives of (4.17) and not levels and cannot directly recover the constant terms in (4.18a) and (4.18b). Using our analysis of LIV but applied to $YD = Y_1 D$ and $Y(1 - D) = Y_0(1 - D)$, it is straightforward to use LIV to estimate the components of the MTE separately. Thus we can identify

$$\mu_1(X) + E(U_1 \mid X, U_D = u_D)$$

and

$$\mu_0(X) + E(U_0 \mid X, U_D = u_D)$$

separately. This corresponds to what is estimated from taking the derivatives of expressions (4.18a) and (4.18b) multiplied by $P(Z)$ and $(1 - P(Z))$, respectively:[76]

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

and

$$\begin{aligned}&\big(1 - P(Z)\big)E(Y_0 \mid X, Z, D = 0)\\ &\quad = \big(1 - P(Z)\big)\mu_0(X) + \big(1 - P(Z)\big)K_0\big(X, P(Z)\big).\end{aligned}$$

Thus the control function method works with levels, whereas the LIV approach works with slopes of combinations of the same basic functions. Constants that do not depend

[75] See Heckman and Navarro (2007) for use of semiparametric curvature restrictions in identification analysis that do not require functional form assumptions.

[76] Björklund and Moffitt (1987) use the derivative of a selection model in levels to define the marginal treatment effect.

on $P(Z)$ disappear from the estimates of the model. The level parameters are obtained by integration using the formulae in Table 2B.

Misspecification of $P(Z)$ (either its functional form or its arguments) and hence of $K_1(P(Z), X)$ and $K_0(P(Z), X)$, in general, produces biased estimates of the parameters of the model under the control function approach even if semiparametric methods are used to estimate μ_0, μ_1, K_0 and K_1. To implement the method, we need to know all of the arguments of Z. The terms $K_1(P(Z), X)$ and $K_0(P(Z), X)$ can be nonparametrically estimated so it is only necessary to know $P(Z)$ up to a monotonic transformation.[77] The distributions of U_0, U_1 and V do not need to be specified to estimate control function models [see Powell (1994)].

These problems with control function models have their counterparts in IV models. If we use a misspecified $P(Z)$ to identify the MTE or its components, in general, we do not identify MTE or its components. Misspecification of $P(Z)$ plagues both approaches.

One common criticism of selection models is that without invoking functional form assumptions, identification of $\mu_1(X)$ and $\mu_0(X)$ requires that $P(Z) \to 1$ and $P(Z) \to 0$ in limit sets.[78] Identification in limit sets is sometimes called "identification at infinity". In order to identify $\text{ATE} = E(Y_1 - Y_0 \mid X)$, IV methods also require that $P(Z) \to 1$ and $P(Z) \to 0$ in limit sets, so an identification at infinity argument is implicit when IV is used to identify this parameter.[79] The LATE parameter avoids this problem by moving the goal posts and redefining the parameter of interest away from a level parameter like ATE or TT to a slope parameter like LATE which differences out the unidentified constants. Alternatively, if we define the parameter of interest to be LATE or MTE, we can use the selection model without invoking identification at infinity.

The IV estimator is model dependent, just like the selection estimator, but in application, the model does not have to be fully specified to obtain Δ^{IV} using Z (or $J(Z)$). However, the distribution of $P(Z)$ and the relationship between $P(Z)$ and $J(Z)$ generates the weights. The interpretation placed on Δ^{IV} in terms of weights on Δ^{MTE} depends crucially on the specification of $P(Z)$. In both control function and IV approaches for the general model of heterogeneous responses, $P(Z)$ plays a central role.

Two economists using the same instrument will obtain the same point estimate using the same data. Their *interpretation* of that estimate will differ depending on how they specify the arguments in $P(Z)$, even if neither uses $P(Z)$ as an instrument. By conditioning on $P(Z)$, the control function approach makes the dependence of estimates on the specification of $P(Z)$ explicit. The IV approach is less explicit and masks the assumptions required to economically interpret the empirical output of an IV estimation. We now turn to some empirical examples of LIV.

[77] See Heckman et al. (1998).

[78] See Imbens and Angrist (1994). Heckman (1990) establishes the identification in the limit argument for ATE in selection models. See Heckman and Navarro (2007) for a generalization to multiple outcome models.

[79] Thus if the support of $P(Z)$ is not full, we cannot identify treatment on the treated or the average treatment effect. We can construct bounds. See Heckman and Vytlacil (1999, 2001a, 2001b).

4.9. *Empirical examples: "The effect" of high school graduation on wages and using IV to estimate "the effect" of the GED*

The previous examples illustrate logical possibilities. This subsection shows that these logical possibilities arise in real data. We analyze two examples: (a) the effect of graduating high school on wages, and (b) the effect of obtaining a GED on wages. We first analyze the effect of graduating high school on wages.

4.9.1. *Empirical example based on LATE: Using IV to estimate* "the effect" *of high school graduation on wages*

We first study the effects of graduating from high school on wages using data from the National Longitudinal Survey of Youth 1979 (NLSY79). This survey gathers information at multiple points in time on the labor market activities for men and women born in the years 1957–1964. We estimate LATE using log hourly wages at age 30 as the outcome measure. Following a large body of research [see Mare (1980)], we use the number of siblings and residence in the south at age 14 as instruments.

Figure 13 plots the weights on LATE using the estimated $P(Z)$. The procedure used to derive the estimates is explained in Heckman, Urzua and Vytlacil (2006). The weights are derived from Equation (4.16). The LATE parameters are both positive and negative. The weights using siblings as an instrument are both positive and negative. The weights using $P(Z)$ as an instrument are positive, as they must be following the analysis of Yitzhaki (1989). The two IV estimates differ from each other because the weights are different. The overall IV estimate is a crude summary of the underlying component LATEs that are both large and positive and large and negative. We next turn to analysis of the GED.

4.9.2. *Effect of the GED on wages*

The GED test is used to certify high school dropouts as high school equivalents. Numerous studies document that the economic return to the GED is low [see Cameron and Heckman (1993), Heckman and LaFontaine (2007)]. It is estimated by the method described in Heckman, Urzua and Vytlacil (2006). In this example, we study the effect of the GED on the wages of recipients compared to wages of dropouts. We use data from the National Longitudinal Survey of Youth 1979 (NLSY79) which gathers information at multiple points in time on the labor market activities for men and women born in the years 1957–1964.

We estimate the MTE for the GED and also consider the IV weights for various instruments for a sample of males at age 25. Figure 14 shows the sample support of $P(Z)$ for both GEDs and high school dropouts. It is not possible to estimate the MTE over its full support. Thus the average treatment effect (ATE) and treatment on the treated (TT) cannot be estimated from these data. The list of Z variables is presented in Table 8 along with IV estimates. The IV estimates fluctuate from positive to negative.

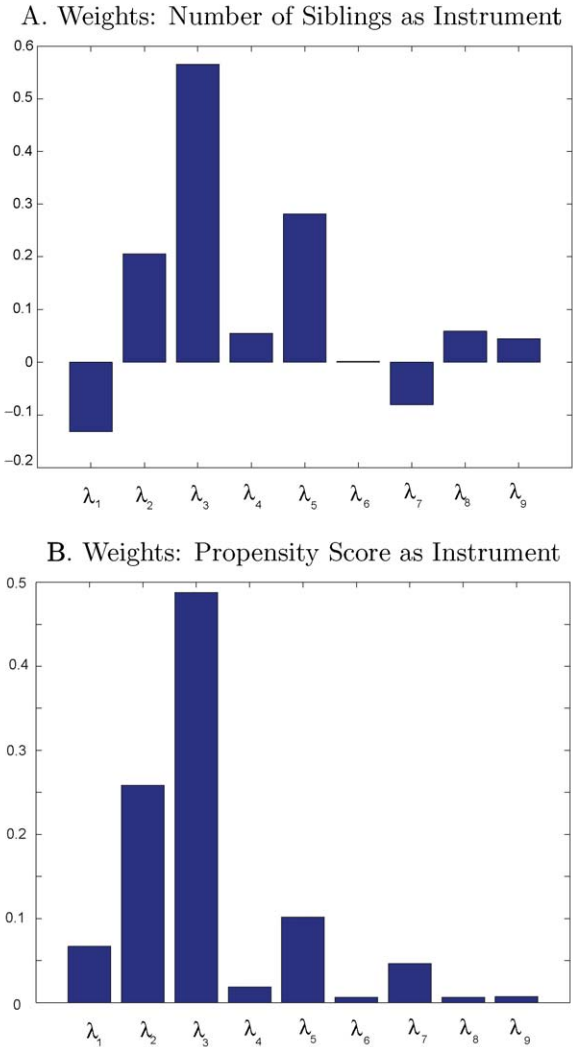
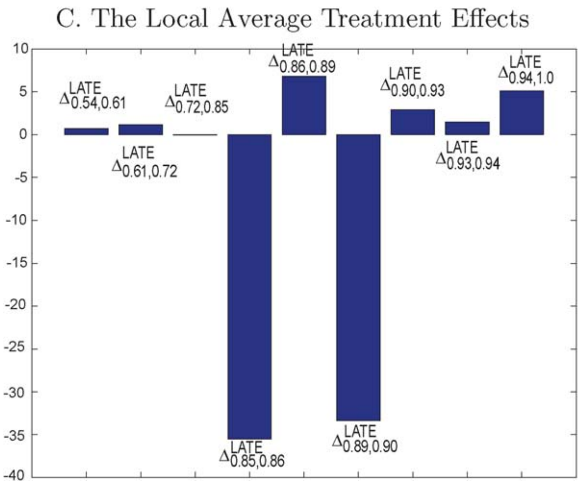


Figure 13. IV weights – the effect of graduating from high school – sample of high school dropouts and high school graduates. *Source*: Heckman, Urzua and Vytlacil (2006).

Using $P(Z)$ as an instrument, the GED effect on log wages is in general negative.[80] For other instruments, the signs and magnitudes vary.

[80] In this example, we use the log of the average nonmissing hourly wages reported between ages 24 and 26. Using the hourly wage reported at age 25 leads to roughly the same results (negative IV weights, and positive and negative IV estimates), but an increase in the standard errors.



Y = Log per-hour wage at age 30, Z_1 = number of siblings in 1979, Z_2 = mother is a high school graduate

$$D = \begin{cases} 1 & \text{if high school graduate,} \\ 0 & \text{if high school dropout} \end{cases}$$

IV estimates
(bootstrap std. errors in parentheses – 100 replications)

Instrument	Value
Number of siblings in 1979	0.115 (0.695)
Propensity score	0.316 (0.110)

Joint probability distribution of (Z_1, Z_2) and the propensity score
(joint probabilities $\Pr(Z_1 = z_1, Z_2 = z_2)$ in ordinary type;
propensity score $\Pr(D = 1 \mid Z_1 = z_1, Z_2 = z_2)$ in italics)

$Z_2 \backslash Z_1$	0	1	2	3	4
0	0.07 *1.0*	0.03 *0.54*	0.47 *0.86*	0.121 *0.72*	0.06 *0.61*
1	0.039 *0.94*	0.139 *0.89*	0.165 *0.90*	0.266 *0.85*	0.121 *0.93*

$\text{Cov}(Z_1, Z_2) = -0.066$, number of observations = 1,702

Figure 13. (*Continued*)

Figure 15 plots the estimated MTE. Details of the nonparametric estimation procedure used to produce these estimates are shown in an appendix in Heckman, Urzua and Vytlacil (2006). Local linear regression is used to estimate the MTE implementing Equation (4.9). While the standard error band is large, the estimated Δ^{MTE} is in general negative, suggesting a negative marginal treatment effect for most participants. However, we observe that for small values of u_D the point estimates of the marginal effect

Table 8

Instrumental variables estimates[a]: Sample of GED and dropouts – males at age 25[b]

Instruments	IV–MTE
Father's highest grade completed	0.146 (0.251)
Mother's highest grade completed	−0.052 (0.179)
Number of siblings	−0.052 (0.160)
GED cost	−0.053 (0.156)
Family income in 1979	−0.047 (0.177)
Dropout's local wage at age 17	−0.013 (0.218)
High school graduate's local wage at age 17	−0.049 (0.182)
Dropout's local unemployment rate at age 17	0.443 (1.051)
High school graduate's local unemployment rate at age 17	−0.563 (0.577)
Propensity score[c]	−0.058 (0.164)

Notes:

[a] The IV estimates are computed by taking the weighted sum of the MTE. The standard deviations (in parentheses) are computed using bootstrapping (50 draws).

[b] We excluded the oversample of poor whites and the military sample. The cost of the GED corresponds to the average testing fee per GED battery by state between 1993 and 2000. (*Source*: GED Statistical Report.) Average local wage for dropouts and high school graduates correspond to the average in the place of residence for each group, respectively, and local unemployment rate corresponds to the unemployment rate in the place of residence. Average local wages, local unemployment rates, mother's and father's education refer to the level at age 17.

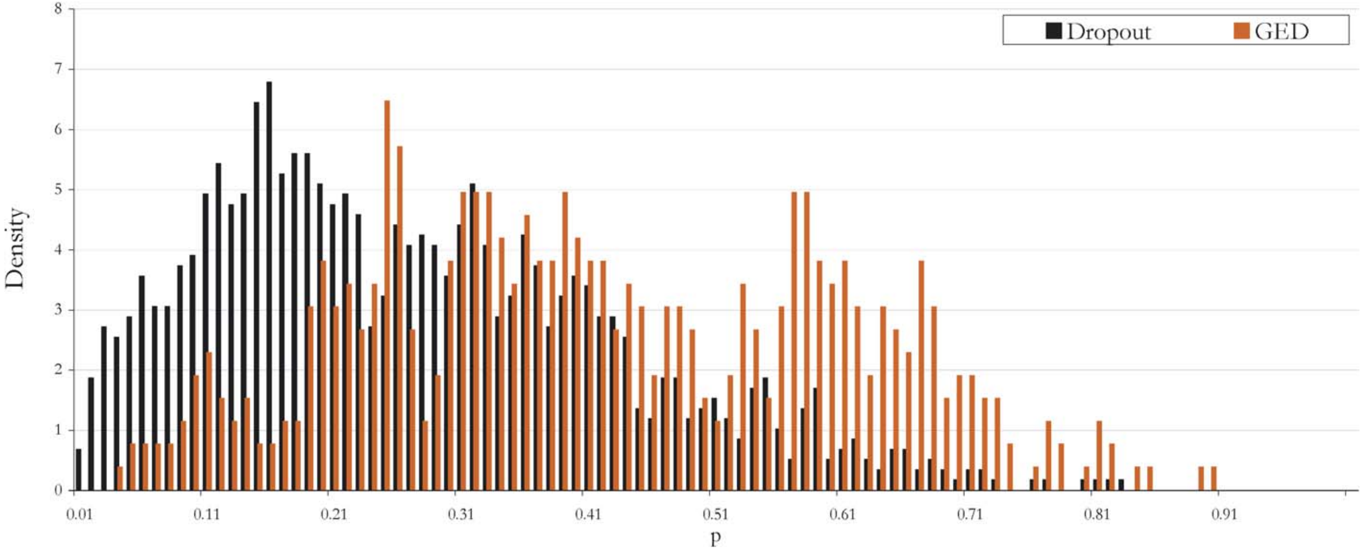
[c] The propensity score ($P(D = 1 \mid Z = z)$) is computed using as controls the instruments presented in the table, as well as two dummy variables controlling for the place of residence at age 14 (south and urban), and a set of dummy variables controlling for the year of birth (1957–1963).

Source: Heckman, Urzua and Vytlacil (2004).

are positive. This analysis indicates that, for people who are more likely to take the GED exam in terms of their unobservables (i.e., for people at the margin of indifference associated with a small u_D), the marginal effect is in fact positive.

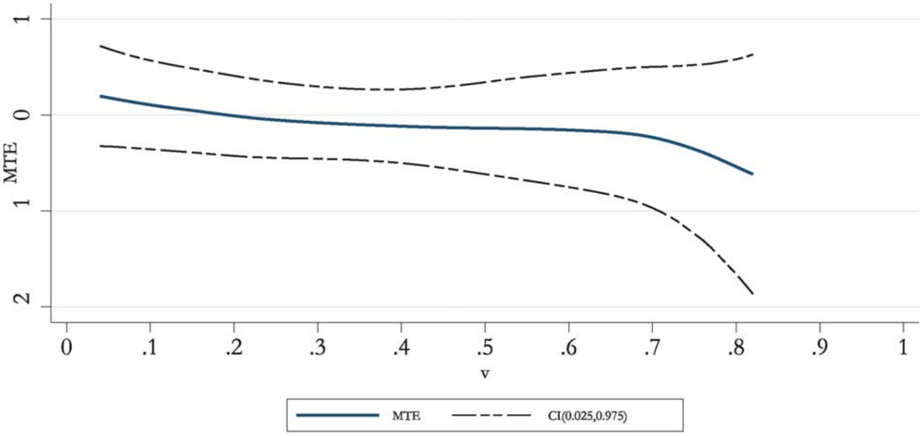
It is instructive to examine the various IV estimates using the one instrument at a time strategy favored by many applied economists who like to do sensitivity analysis.[81]

[81] See, e.g., Card (2001).



Note: The propensity score ($\mathrm{P}(D = 1 \mid Z)$) is computed using as controls (Z): Father's highest grade completed, mother's highest grade completed, number of siblings, GED testing fee by state between 1993 and 2000, family income in 1979, dropout's local wage at age 17, and high school graduate's local unemployment at age 17. We also include two dummy variables controlling for the place of residence at age 14 (south and urban), and a set of dummies controlling for the year of birth (1957–1963).

Figure 14. Frequency of the propensity score by final schooling decision: Dropouts and GEDs, NLSY males at age 25. *Source*: Heckman, Urzua and Vytlacil (2004).



Note: The dependent variable in the outcome equation is the log of the average hourly wage reported between ages 24 and 26. The controls in the outcome equations are tenure, tenure squared, experience, corrected AFQT, black (dummy), Hispanic (dummy), marital status, and years of schooling. Let $D = 0$ denote dropout status and $D = 1$ denote GED status. The model for D (choice model) includes as controls the corrected AFQT, number of siblings, father's education, mother's education, family income at age 17, local GED costs, broken home at age 14, average local wage at age 17 for dropouts and high school graduates, local unemployment rate at age 17 for dropouts and high school graduates, the dummy variables for black and Hispanic, and a set of dummy variables controlling for year of birth. We also include two dummy variables controlling for the place of residence at age 14 (south and urban). The choice model is estimated using a probit model. In computing the MTE, the bandwidths are selected using the "leave one out" cross-validation method. We use biweight kernel functions. The confidence interval is computed from bootstrapping using 50 draws.

Figure 15. MTE of the GED with confidence interval: Dropouts and GEDs, males of the NLSY at the age 25. *Source*: Heckman, Urzua and Vytlacil (2004).

Many of the variables used in the analysis are determined by age 17. Both father's highest grade completed and local unemployment rate among high school dropouts produce positive (if not precisely determined) IV estimates. A negative MTE weighted by negative IV weights produces a positive IV. A naive application of IV could produce the wrong causal inference, i.e., that GED certification raises wages. Our estimates show that our theoretical examples have real world counterparts.[82]

Carneiro, Heckman and Vytlacil (2006) present an extensive empirical analysis of the wage returns to college attendance. They show how to unify and interpret diverse instruments within a common framework using the MTE and the weights derived in Heckman and Vytlacil (1999, 2001a, 2005). They show negative weights on the MTE for commonly used instruments. Basu et al. (2007) use the MTE and the derived weights to identify the ranges of the MTE identified by different instruments in their analysis of the costs of breast cancer. We next discuss the implications of relaxing separability in the choice equations.

[82] We discuss the GED further in Section 7.

4.10. Monotonicity, uniformity, nonseparability, independence and policy invariance: The limits of instrumental variables

The analysis of this section and the entire recent literature on instrumental variables estimators for models with heterogeneous responses (i.e., models with outcomes of the forms (3.1) and (3.2)) relies critically on the assumption that the treatment choice equation has a representation in the additively separable form (3.3). From Vytlacil (2002), we know that under assumptions (A-1)–(A-5), separability is equivalent to the assumption of monotonicity or uniformity, (IV-3).

This uniformity condition imparts an asymmetry to the entire instrumental variable enterprise. Responses are permitted to be heterogeneous in a general way, but choices of treatment are not. In this section, we relax the assumption of additive separability in (3.3). We establish that in the absence of additive separability or uniformity, the entire instrumental variable identification strategy in this section and the entire recent literature collapses. Parameters can be defined as weighted averages of an MTE. MTE and the derived parameters cannot be identified using any instrumental variable strategy. Appendix B presents a comprehensive discussion, which we summarize in this subsection.

One natural benchmark nonseparable model is a random coefficient model of choice $D = \mathbf{1}[Z\gamma \geqslant 0]$, where γ is a random coefficient vector and $\gamma \perp\!\!\!\perp (Z, U_0, U_1)$. If γ is a random coefficient with a nondegenerate distribution and with components that take both positive and negative values, uniformity is clearly violated. However, it can be violated even when all components of γ are of the same sign if Z is a vector.[83]

Relax the additive separability assumption of Equation (3.3) to consider a more general case

$$D^* = \mu_D(Z, V), \tag{4.19a}$$

where $\mu_D(Z, V)$ is not necessarily additively separable in Z and V, and V is not necessarily a scalar.[84] In the random coefficient example, $V = \gamma$ and $\mu_D = z\gamma$.

$$D = \mathbf{1}\big[D^* \geqslant 0\big]. \tag{4.19b}$$

We maintain assumptions (A-1)–(A-5) and (A-7).

In special cases, (4.19a) can be expressed in an additively separable form. For example, if D^* is weakly separable in Z and V, $D^* = \mu_D(\theta(Z), V)$ for any V where $\theta(Z)$ is a scalar function, μ_D is increasing in $\theta(Z)$, and V is a scalar, then we can write (4.19b) in the same form as (3.3):

$$D = \mathbf{1}\big[\theta(Z) \geqslant \tilde{V}\big],$$

[83] Thus, if γ is a vector with positive components, a change from $Z = z$ to $Z = z'$ can produce different effects on choice if γ varies in the population and if components of Z are of different signs.

[84] The additively separable latent index model is more general than it may at first appear. It is shown in Vytlacil (2006a) that a wide class of threshold crossing models without the additive structure on the latent index will have a representation with the additively separable structure on the latent index.

where $\tilde{V} = \mu_D^{-1}(0; V)$ and $\tilde{V} \perp\!\!\!\perp Z \mid X$, and the inverse function is expressed with respect to the first argument [see Vytlacil (2006a)]. Vytlacil (2002) shows that any model that does not satisfy uniformity (or "monotonicity") will not have a representation in this form.[85]

In the additively separable case, the MTE (3.4) has three equivalent interpretations. (i) $U_D = F_V(V)$ is the only unobservable in the first stage decision rule, and MTE is the average effect of treatment given the unobserved characteristics in the decision rule $(V = v)$. (ii) A person with $V = v$ would be indifferent between treatment or not if $P(Z) = u_D$, where $P(Z)$ is a mean scale utility function. Thus, the MTE is the average effect of treatment given that the individual would be indifferent between treatment or not if $P(Z) = u_D$. (iii) One can also view the additively separable form (3.3) as intrinsic in the way we are defining the parameter and interpret the MTE (Equation (3.4)) as an average effect conditional on the additive error term from the first stage choice model. Under all interpretations of the MTE and under the assumptions used in the preceding sections of this chapter, MTE can be identified by LIV; the MTE does not depend on Z and hence it is policy invariant and the MTE integrates up to generate all treatment effects, policy effects and all IV estimands.

The three definitions are not the same in the general nonseparable case (4.19a). Heckman and Vytlacil (2001b) extend MTE in the nonseparable case using interpretation (i). MTE defined this way is policy invariant to changes in Z. Appendix B, which summarizes their work, shows that LIV is a weighted average of the MTE with possibly negative weights and does not identify MTE. If uniformity does not hold, the definition of MTE allows one to integrate MTE to obtain all of the treatment effects, but the instrumental variables estimator breaks down.

Alternatively, one could define MTE based on (ii):

$$\Delta_{\mathrm{ii}}^{\mathrm{MTE}}(z) = E\big(Y_1 - Y_0 \bigm| V \in \{v\colon \mu_D(z, v) = 0\}\big).$$

This is the average treatment effect for individuals who would be indifferent between treatment or not at a given value of z (recall that we keep the conditioning on X implicit). Heckman and Vytlacil (2001b) show that in the nonseparable case LIV does not identify this MTE and that MTE does not change when the distribution of Z changes, provided that the support of MTE does not change.[86] In general, this definition of MTE does not allow one to integrate up MTE to obtain the treatment parameters.

A third possibility is to force the index rule into an additive form by taking $\mu_D^*(Z) = E(\mu_D(Z, V) \mid Z)$, defining $V^* = \mu_D(Z, V) - E(\mu_D(Z, V) \mid Z)$ and define MTE as $E(Y_1 - Y_0 \mid V^* = v^*)$. Note that V^* is not independent of Z, is not policy invariant and is not structural. LIV does not estimate this MTE. With this definition of the MTE it is not possible, in general, to integrate up MTE to obtain the various treatment effects.

[85] In the random coefficient case where $Z = (1, Z_1)$ where Z_1 is a scalar, and $\gamma = (\gamma_0, \gamma_1)$ if $\gamma_1 > 0$ for all realizations, we can write the choice rule in the form of (3.3): $Z_1\gamma_1 \geqslant -\gamma_0 \Rightarrow Z_1 \geqslant -\frac{\gamma_0}{\gamma_1}$ and $\tilde{V} = -\frac{\gamma_0}{\gamma_1}$. This trick does not work in the general case.

[86] If the support of Z changes, then the MTE must be extended to a new support.

For any version of the nonseparable model, except those that can be transformed to separability, index sufficiency fails. To see this, assume that $\mu_D(Z, V)$ is continuous.[87] Define $\Omega(z) = \{v\colon \mu_D(z, v) \geqslant 0\}$. In the additively separable case, $P(z) \equiv \Pr(D = 1 \mid Z = z) = \Pr(U_D \in \Omega(z))$, $P(z) = P(z') \Leftrightarrow \Omega(z) = \Omega(z')$. This produces index sufficiency. In the more general case of (4.19a), it is possible to have (z, z') such that $P(z) = P(z')$ and $\Omega(z) \neq \Omega(z')$ so index sufficiency does not hold.

4.10.1. Implications of nonseparability

This section develops generalization (i), leaving development of the other interpretations for later research. We focus on an analysis of PRTE, comparing two policies $p, p' \in \mathcal{P}$. Here "p" denotes a policy and not a realization of $P(Z)$ as in the previous sections. This is our convention when we discuss PRTE. The analysis of the other treatment parameters follows by parallel arguments.

For any v in the support of the distribution of V, define $\Omega = \{z\colon \mu_D(z, v) \geqslant 0\}$. For example, in the random coefficient case, with $V \equiv \gamma$ and $D = \mathbf{1}[Z\gamma \geqslant 0]$, we have $\Omega_g = \{z\colon zg \geqslant 0\}$, where g is a realization of γ. Define $\mathbf{1}_{\mathcal{A}}(t)$ to be the indicator function for the event $t \in \mathcal{A}$. Then, making the X explicit, Appendix B derives the result that

$$\begin{aligned} &E(Y_p) - E(Y_{p'}) \\ &\quad = E\big[E(Y_p \mid X) - E(Y_{p'} \bigm| X)\big] \\ &\quad = \int \bigg[\int E(\Delta^{\text{MTE}} \mid X = x, V = v) \\ &\qquad \times \big(\Pr[Z_p \in \Omega \mid X = x] - \Pr[Z_{p'} \in \Omega \mid X = x]\big)\, dF_{V|X}(v \mid x)\bigg]\, dF_X(x). \end{aligned} \tag{4.20}$$

Thus, without additive separability, we can still derive an expression for PRTE and by similar reasoning the other treatment parameters. However, to evaluate the expression requires knowledge of MTE, of $\Pr[Z_p \in \Omega \mid X = x]$ and $\Pr[Z_{p'} \in \Omega \mid X = x]$ for every (v, x) in the support of the distribution of (V, X), and of the distribution of V. In general, if no structure is placed on the μ_D function, one can normalize V to be unit uniform (or a vector of unit uniform random variables) so that $F_{V|X}$ will be known.

However, in this case, the $\Omega = \{z\colon \mu_D(z, v) \geqslant 0\}$ sets will not in general be identified. If structure is placed on the μ_D function, one might be able to identify the $\Omega = \{z\colon \mu_D(z, v) \geqslant 0\}$ sets but then one needs to identify the distribution of V (conditional on X). If structure is placed on μ_D, one cannot in general normalize the distribution of V to be unit uniform without undoing the structure being imposed on μ_D.

In particular, consider the random coefficient model $D = \mathbf{1}[Z\gamma \geqslant 0]$ where $V = \gamma$ is a random vector, so that $\Omega_\gamma = \{z\colon z\gamma \geqslant 0\}$. In this case, if all of the other assumptions

[87] Absolutely continuous with respect to Lebesgue measure.

hold, including $Z \perp\!\!\!\perp \gamma \mid X$, and the policy change does not affect (Y_1, Y_0, X, γ), the PRTE is given by

$$\begin{aligned} E(Y_p) - E(Y_{p'}) &= E\big[E(Y_p \mid X) - E(Y_{p'} \mid X)\big] \\ &= \int \bigg[\int E(\Delta^{\text{MTE}} \mid X = x, \gamma = g)\big(\Pr[Z_p \in \Omega_g \mid X = x] \\ &\quad - \Pr[Z_{p'} \in \Omega_g \mid X = x]\big)\, dF_{\gamma|X}(g \mid x)\bigg]\, dF_X(x). \end{aligned}$$

Because structure has been placed on the $\mu_D(Z, \gamma)$ function, the sets Ω_γ are known. However, evaluating the function requires knowledge of the distribution of γ which will not in general be identified without further assumptions.[88] Normalizing the distribution of γ to be a vector of unit uniform random variables produces the distribution of γ but eliminates the assumed linear index structure on μ_D and results in Ω_γ sets that are not identified.

Even if the weights are identified, Heckman and Vytlacil (2001b) show that it is not possible to use LIV to identify MTE without additive separability between Z and V in the selection rule index. Appendix F develops this point for the random coefficient model. Without additive separability in the latent index for the selection rule, we can still create an expression for PRTE (and the other treatment parameters) but both the weights and the MTE function are no longer identified using instrumental variables.

One superficially plausible way to avoid these problems would be to define $\tilde{\mu}_D(Z) = E(\mu_D(Z, V) \mid Z)$ and $\tilde{V} = \mu_D(Z, V) - E(\mu_D(Z, V) \mid Z)$, producing the model $D = \mathbf{1}[\tilde{\mu}_D(Z) + \tilde{V} \geqslant 0]$. We keep the conditioning on X implicit. One could redefine MTE using $\tilde{V}$ and proceed as if the true model possessed additive separability between observables and unobservables in the latent index. This is the method pursued in approach (iii).

For two reasons, this approach does not solve the problem of providing an adequate generalization of MTE. First, with this definition, $\tilde{V}$ is a function of (Z, V), and a policy that changes Z will then also change $\tilde{V}$. Thus, policy invariance of the MTE no longer holds. Second, this approach generates a $\tilde{V}$ that is no longer statistically independent of Z so that assumption (A-1) no longer holds when $\tilde{V}$ is substituted for V even when (A-1) is true for V. Lack of independence between observables and unobservables in the latent index both invalidates our expression for PRTE (and the expressions for the other treatment effects) and causes LIV to no longer identify MTE.

The nonseparable model can also restrict the support of $P(Z)$. For example, consider a standard normal random coefficient model with a scalar regressor ($Z = (1, Z_1)$). Assume $\gamma_0 \sim N(0, \sigma_0^2)$, $\gamma_1 \sim N(\bar{\gamma}_1, \sigma_1^2)$, and $\gamma_0 \perp\!\!\!\perp \gamma_1$. Then

$$P(z_1) = \Phi\left(\frac{\bar{\gamma}_1 z_1}{\sqrt{\sigma_0^2 + \sigma_1^2 z_1^2}}\right),$$

[88] See, e.g., Ichimura and Thompson (1998) for conditions for identifying the distribution of γ in a random coefficient discrete choice model when $Z \perp\!\!\!\perp \gamma$.

where Φ is the standard cumulative normal distribution. If the support of z_1 is $\mathbb{R}$, then in the standard additive model, $\sigma_1^2 = 0$ and $P(z_1)$ has support [0, 1]. When $\sigma_1^2 > 0$, the support is strictly within the unit interval.[89] In the special case when $\sigma_0^2 = 0$, the support is one point ($P(z) = \Phi(\frac{\bar{\gamma}_1}{\sigma_1})$). We cannot, in general, identify ATE, TT or any treatment effect requiring the endpoints 0 or 1.

Thus the general models of nonuniformity presented in this section do not satisfy the index sufficiency property, and the support of the treatment effects and estimators is, in general, less than full. The random coefficient model for choice may explain the empirical support problems for $P(Z)$ found in Heckman et al. (1998) and many other evaluation studies.

4.10.2. Implications of dependence

We next consider relaxing the independence assumption (A-1) to allow $Z \not\perp\!\!\!\perp V \mid X$ while maintaining the assumption that $Z \perp\!\!\!\perp (Y_0, Y_1) \mid (X, V)$. We maintain the other assumptions, including additive separability between Z and V in the latent index for the selection rule (Equation (3.3)) and the assumption that the policy changes Z but does not change (V, Y_0, Y_1, X). Thus we assume that the policy shift does not change the MTE function (policy invariance). Given these assumptions, we derive in Appendix C the following expression for PRTE in the nonindependent case for policies $p, p' \in \mathcal{P}$:

$$\begin{aligned}
&E(Y_p) - E(Y_{p'}) \\
&\quad = E\big[E(Y_p \mid X) - E(Y_{p'} \mid X)\big] \\
&\quad = \int \bigg[\int E(\Delta^{\text{MTE}} \mid X = x, V = v)\big(\Pr\big[\mu_D(Z_{p'}) < v \;\big|\; X = x, V = v\big] \\
&\qquad\quad - \Pr\big[\mu_D(Z_p) < v \;\big|\; X = x, V = v\big]\big)\, dF_{V|X}(v \mid x) \bigg] dF_X(x). \qquad (4.21)
\end{aligned}$$

Notice that "p" denotes a policy and not a realized value of $P(Z)$. Although we can derive an expression for PRTE without requiring independence between Z and V, to evaluate this expression requires knowledge of MTE and of $\Pr[\mu_D(Z_{p'}) < v \mid X = x, V = v]$ and of $\Pr[\mu_D(Z_p) < v \mid X = x, V = v]$ for every (x, v) in the support of the distribution of (X, V). This requirement is stronger than what is needed in the case of independence since the weights no longer depend only on the distribution of $P_p(Z_p)$ and $P_{p'}(Z_{p'})$ conditional on X. To evaluate these weights requires knowledge of the function μ_D and of the joint distribution of (V, Z_p) and $(V, Z_{p'})$ conditional on X, and these will in general not be identified without further assumptions.

Even if the weights are identified, Heckman and Vytlacil (2001b) show that it is not possible to use LIV to identify MTE without independence between Z and V conditional on X. Thus, without conditional independence between Z and V in the latent

[89] The interval is $[\Phi(\frac{-|\bar{\gamma}_1|}{\sigma_1}), \Phi(\frac{|\bar{\gamma}_1|}{\sigma_1})]$.

index for the decision rule, we can still create an expression for PRTE but both the weights and the MTE function are no longer identified without invoking further assumptions.

One superficially appealing way to avoid these problems is to define $\tilde{V} = F_{V|X,Z}(V)$ and $\tilde{\mu}_D(Z) = F_{V|X,Z}(\mu_D(Z))$, so $D = \mathbf{1}[\mu_D(Z) - V \geqslant 0] = \mathbf{1}[\tilde{\mu}_D(Z) - \tilde{V} \geqslant 0]$ with $\tilde{V} \sim \text{Unif}[0,1]$ conditional on X and Z and so $\tilde{V}$ is independent of X and Z. It might seem that the previous analysis would carry over. However, by defining $\tilde{V} = F_{V|X,Z}(V)$, we have defined $\tilde{V}$ in a way that depends functionally on Z and X, and hence we violate invariance of the MTE with respect to the shifts in the distribution of Z given X.

4.10.3. The limits of instrumental variable estimators

The treatment effect literature focuses on a class of policies that move treatment choices in the same direction for everyone. General instruments do not have universally positive weights on Δ^{MTE}. They are not guaranteed to shift everyone in the same direction. They do not necessarily estimate gross treatment effects. However, the effect of treatment is not always the parameter of policy interest. Thus, in the housing subsidy example developed in Section 4.6, migration is the vehicle through which the policy operates. One might be interested in the effect of migration (the treatment effect) or the effect of the policy (the housing subsidy). These are separate issues unless the policy is the treatment.

Generalizing the MTE to the case of a nonseparable choice equation that violates the monotonicity condition, we can define but cannot identify the policy parameters of interest using ordinary instrumental variables or our extension LIV. If we make the model symmetrically heterogeneous in outcome and choice equations, the method of instrumental variables and our extensions of it break down in terms of estimating economically interpretable parameters. Vytlacil and Yildiz (2006) and Vytlacil, Santos and Shaikh (2005) restore symmetry in the IV analysis of treatment choice and outcome equations by imposing uniformity on both outcome and choice equations. The general case of heterogeneity in both treatment and choice equations is beyond the outer limits of the entire IV literature, although it captures intuitively plausible phenomena. More general structural methods are required.[90]

5. Regression discontinuity estimators and LATE

Campbell (1969) developed the regression discontinuity design which is now widely used. [See an early discussion of this estimator in econometrics by Barnow, Cain and

[90] The framework of Carneiro, Hansen and Heckman (2003) can be generalized to allow for random coefficient models in choice equations, and lack of policy invariance in the sense of assumption (A-7). However, a fully semiparametric analysis of treatment and choice equations with random coefficients remains to be developed.

Goldberger (1980).] Hahn, Todd and Van der Klaauw (2001) present an exposition of the regression discontinuity estimator within a LATE framework. This section exposits the regression discontinuity method within our MTE framework.

Suppose assumptions (A-1)–(A-5) hold except that we relax independence assumption (A-1) to assume that $(Y_1 - Y_0, U_D)$ is independent of Z conditional on X. We *do not* impose the condition that Y_0 is independent of Z conditional on X. Relaxing the assumption that Y_0 is independent of Z conditional on X causes the standard LIV estimand to differ from the MTE. We show that the LIV estimand in this case equals MTE plus a bias term that depends on $\frac{\partial}{\partial p} E(Y_0 \mid X = x, P(Z) = p)$. Likewise, we show that the discrete-difference IV formula will no longer correspond to LATE, but will now correspond to LATE plus a bias term.

A regression discontinuity design allows analysts to recover a LATE parameter at a particular value of Z. If $E(Y_0 \mid X = x, Z = z)$ is continuous in z, while $P(z)$ is discontinuous in z at a particular point, then it will be possible to use a regression discontinuity design to recover a LATE parameter. While the regression discontinuity design does have the advantage of allowing Y_0 to depend on Z conditional on X, it only recovers a LATE parameter at a particular value of Z and cannot in general be used to recover either other treatment parameters such as the average treatment effect or the answers to policy questions such as the PRTE. The following discussion is motivated by the analysis of Hahn, Todd and Van der Klaauw (2001).

For simplicity, assume that Z is a scalar random variable. First, consider LIV while relaxing independence assumption (A-1) to assume that $(Y_1 - Y_0, U_D)$ is independent of Z conditional on X but without imposing that Y_0 is independent of Z conditional on X. In order to make the comparison with the regression discontinuity design easier, we will condition on Z instead of $P(Z)$. Using $Y = Y_0 + D(Y_1 - Y_0)$, we obtain

$$\begin{aligned}
&E(Y \mid X = x, Z = z) \\
&\quad = E(Y_0 \mid X = x, Z = z) + E\big(D(Y_1 - Y_0) \bigm| X = x, Z = z\big) \\
&\quad = E(Y_0 \mid X = x, Z = z) + \int_0^{P(z)} E(Y_1 - Y_0 \mid X = x, U_D = u_D)\, du_D.
\end{aligned}$$

So

$$\begin{aligned}
\frac{\frac{\partial}{\partial z} E(Y \mid X = x, Z = z)}{\frac{\partial}{\partial z} P(z)} = {} & \frac{\frac{\partial}{\partial z} E(Y_0 \mid X = x, Z = z)}{\frac{\partial}{\partial z} P(z)} \\
& + E\big(Y_1 - Y_0 \bigm| X = x, U_D = P(z)\big)
\end{aligned}$$

where we have assumed that $\frac{\partial}{\partial z} P(z) \neq 0$ and that $E(Y_0 \mid X = x, Z = z)$ is differentiable in z. Notice that under our stronger independence condition (A-1), $\frac{\partial}{\partial z} E(Y_0 \mid X = x, Z = z) = 0$ so that we identify MTE as before. With Y_0 possibly dependent on Z conditional on X, we now get MTE plus the bias term that depends on $\frac{\partial}{\partial z} E(Y_0 \mid X = x, Z = z)$. Likewise, if we consider the discrete change form of IV:

$$\begin{aligned}
&\frac{E(Y \mid X=x, Z=z) - E(Y \mid X=x, Z=z')}{P(z)-P(z')} \\
&\quad = \underbrace{\frac{E(Y_0 \mid X=x, Z=z) - E(Y_0 \mid X=x, Z=z')}{P(z)-P(z')}}_{\text{Bias for LATE}} \\
&\qquad + \underbrace{E\big(Y_1 - Y_0 \bigm| X=x, P(z) > U_D > P(z')\big)}_{\text{LATE}}
\end{aligned}$$

so that we now recover LATE plus a bias term.

Now consider a regression discontinuity design. Suppose that there exists an evaluation point z_0 for Z such that $P(\cdot)$ is discontinuous at z_0, and suppose that $E(Y_0 \mid X=x, Z=z)$ is continuous at z_0. Suppose that $P(\cdot)$ is increasing in a neighborhood of z_0. Let

$$\begin{aligned}
P(z_0-) &= \lim_{\epsilon \downarrow 0} P(z_0 - \epsilon), \\
P(z_0+) &= \lim_{\epsilon \downarrow 0} P(z_0 + \epsilon),
\end{aligned}$$

and note that the conditions that $P(\cdot)$ is increasing in a neighborhood of z_0 and discontinuous at z_0 imply that $P(z_0+) > P(z_0-)$. Let

$$\begin{aligned}
\mu(x, z_0-) &= \lim_{\epsilon \downarrow 0} E(Y \mid X=x, Z=z_0-\epsilon), \\
\mu(x, z_0+) &= \lim_{\epsilon \downarrow 0} E(Y \mid X=x, Z=z_0+\epsilon),
\end{aligned}$$

and note that

$$\begin{aligned}
\mu(x, z_0-) &= E(Y_0 \mid X=x, Z=z_0) \\
&\quad + \int_0^{P(z_0-)} E(Y_1 - Y_0 \mid X=x, U_D=u_D)\, du_D
\end{aligned}$$

and

$$\begin{aligned}
\mu(x, z_0+) &= E(Y_0 \mid X=x, Z=z_0) \\
&\quad + \int_0^{P(z_0+)} E(Y_1 - Y_0 \mid X=x, U_D=u_D)\, du_D,
\end{aligned}$$

where we use the fact that $E(Y_0 \mid X=x, Z=z)$ is continuous at z_0. Thus,

$$\begin{aligned}
&\mu(x, z_0+) - \mu(x, z_0-) = \int_{P(z_0-)}^{P(z_0+)} E(Y_1 - Y_0 \mid X=x, U_D=u_D)\, du_D \\
&\Rightarrow \quad \frac{\mu(x, z_0+) - \mu(x, z_0-)}{P(z_0+) - P(z_0-)} = E\big(Y_1 - Y_0 \bigm| X=x, P(z_0+) \geqslant U_D > P(z_0-)\big)
\end{aligned}$$

so that we now recover a LATE parameter for a particular point of evaluation. Note that if $P(z)$ is only discontinuous at z_0, then we only identify $E(Y_1 - Y_0 \mid X = x,\, P(z_0+) \geqslant U_D > P(z_0-))$ and not any LATE or MTE at any other evaluation points. While this discussion assumes that Z is a scalar, it is straightforward to generalize the discussion to allow for Z to be a vector. For more discussion of the regression discontinuity design estimator and an example, see Hahn, Todd and Van der Klaauw (2001).

6. Policy evaluation, out-of-sample policy forecasting, forecasting the effects of new policies and structural models based on the MTE

We have thus far focused on policy problem P-1, the problem of "internal validity". We have shown how to identify a variety of parameters but have not put them to use in evaluating policies. This section discusses policy evaluation and out-of-sample forecasting. We discuss two distinct evaluation and forecasting problems. The first problem uses the MTE to develop a cost benefit analysis. Corresponding to the gross benefit parameters analyzed in Sections 3–4, there is a parallel set of cost parameters that emerge from the economics of the generalized Roy model. This part of our analysis works in the domain of problem P-1 to construct a cost-benefit analysis for programs in place. However, these tools can be extended to new environments using the other results established in this section.

The second topic is the problem of constructing the PRTE in new environments in a more general way. This addresses policy problems P-2 and P-3 and considers large scale changes in policies and forecasts of new policies.

6.1. Econometric cost benefit analysis based on the MTE

This section complements the analysis of Section 3. There we developed gross outcome measures for a generalized Roy model. Here we define a parallel set of treatment parameters for the generalized Roy model corresponding to the average cost of participating in a program. The central feature of the generalized Roy model is that the agent chooses treatment if the benefit exceeds the subjective cost perceived by the agent. This creates a simple relationship between the cost and benefit parameters that can be exploited for identifying or bounding the cost parameters by adapting the results of the previous sections. The main result of this section is that cost parameters in the generalized Roy model can be identified or bounded without direct information on the costs of treatment. Our analysis complements and extends the analysis of Björklund and Moffitt (1987) who first noted this duality.

Assume the outcomes (Y_0, Y_1) are generated by the additively separable system (2.2). Let C denote the individual-specific subjective cost of selecting into treatment. We assume that C is generated by: $C = \mu_C(W) + U_C$, where W is a (possibly vector-valued) observed random variable and U_C is an unobserved random variable. We assume that the agent selects into treatment if the benefit exceeds the cost, using the structure of the generalized Roy model where $D = \mathbf{1}[Y_1 - Y_0 \geqslant C]$ and $C = \mu_C(W) + U_C$, where $\mu_C(W)$ is nondegenerate and integrable; U_C is continuous and $Z = (W, X)$ is independent of (U_C, U_0, U_1).[91]

We do not assume any particular functional form for the functions μ_0, μ_1 and μ_C, and we do not assume that the distribution of U_0, U_1, or U_C is known.[92] Let $V \equiv U_C - (U_1 - U_0)$ and let F_V denote the distribution function of V. As before, we use the convention that U_D is the probability integral transformation of the latent variable generating choices so that $U_D = F_V(V)$. Let $P(z) \equiv \Pr(D = 1 \mid Z = z)$ so that $P(z) = F_V(\mu_1(x) - \mu_0(x) - \mu_C(w))$. For convenience, we will assume that F_V is strictly increasing so that F_V will be invertible, though this assumption is not required. We work with $U_D = F_V(V)$ instead of working directly with V to link our analysis to that in Section 3. In this section we make explicit the conditioning on X, Z, and W because it plays an important role in the analysis.

Corresponding to the treatment parameters defined in Section 2 and Tables 2A and 2B, we can define analogous cost parameters. We define the marginal cost of treatment for a person with characteristics $W = w$ and $U_D = u_D$ as

$$C^{\mathrm{MTE}}(w, u_D) \equiv E(C \mid W = w, U_D = u_D).$$

This is a cost version of the marginal treatment effect. Likewise, we have an analogue average cost:

$$\begin{aligned} C^{\mathrm{ATE}}(w) &\equiv E(C \mid W = w) \\ &= \int_0^1 E(C \mid W = w, U_D = u_D)\, du_D, \end{aligned} \tag{6.1}$$

recalling that $dF_{U_D}(u_D) = du_D$ because U_D is uniform. This is the mean subjective cost of treatment as perceived by the average agent. We next consider

$$\begin{aligned} C^{\mathrm{TT}}\big(w, P(z)\big) &\equiv E\big(C \,\big|\, W = w, P(Z) = P(z), D = 1\big) \\ &= \frac{1}{P(z)} \int_0^{P(z)} E(C \mid W = w, U_D = u_D)\, du_D. \end{aligned}$$

This is the mean subjective cost of treatment as perceived by the treated with a given value of $P(z)$. Removing the conditioning on $P(z)$,

$$\begin{aligned} C^{\mathrm{TT}}(w) &\equiv E(C \mid W = w, D = 1) \\ &= \int_0^1 E(C \mid W = w, U_D = u_D) g_w(u_D)\, du_D, \end{aligned}$$

[91] We require that U_C be absolutely continuous with respect to Lebesgue measure.

[92] Recall that the original Roy model (1951) assumes that $U_C = 0$, that there are no observed X and W regressors, that $(U_0, U_1) \sim N(0, \Sigma)$ and that only $Y = DY_1 + (1 - D)Y_0$ is observed, but not both components of the sum at the same time.

where $g_w(u_D) = \frac{1-F_{P(Z)|W=w}(u_D)}{\int(1-F_{P(Z)|W=w}(t))\,dt}$ and $F_{P(Z)|W=w}$ denotes the distribution of $P(Z)$ conditional on $W = w$. This is the mean subjective cost of treatment for the treated. Finally, we can derive a LATE version of the cost:

$$C^{\mathrm{LATE}}\bigl(w, P(z), P(z')\bigr) \equiv \frac{1}{P(z) - P(z')} \int_{P(z')}^{P(z)} E(C \mid W = w, U_D = u_D)\, du_D.$$

This is the mean subjective cost of switching states for those induced to switch status by a change in the instrument.

The generalized Roy model makes a tight link between the cost of treatment and the benefit of treatment. Thus one might expect a relationship between the gross benefit and cost parameters. We show that the benefit and cost parameters coincide for MTE. This relationship can be used to infer information on the subjective cost of treatment by the use of local instrumental variables.

Define $\Delta^{\mathrm{LIV}}(x, P(z))$ as in Equation (4.9):

$$\Delta^{\mathrm{LIV}}\bigl(x, P(z)\bigr) \equiv \frac{\partial E(Y \mid X = x, P(Z) = P(z))}{\partial P(z)}.$$

Under assumptions (A-1)–(A-5), LIV identifies MTE:

$$\Delta^{\mathrm{LIV}}\bigl(x, P(z)\bigr) = \Delta^{\mathrm{MTE}}\bigl(x, P(z)\bigr).$$

Note that

$$\begin{aligned}\Delta^{\mathrm{MTE}}\bigl(x, P(z)\bigr) &= E\bigl(\Delta \bigm| X = x, U_D = P(z)\bigr)\\ &= E\bigl(\Delta \bigm| X = x, \Delta(x) = C(w)\bigr)\\ &= E\bigl(\Delta(x) \bigm| \Delta(x) = C(w)\bigr),\end{aligned} \tag{6.2}$$

where $\Delta(x) = \mu_1(x) - \mu_0(x) + U_1 - U_0$, and $C(w) = \mu_C(w) + U_C$. ($\Delta(x)$ and $C(w)$ are, respectively, the benefit and cost for the agent if the X and W are externally set to x and w without changing (U_1, U_0, U_D) values.) We thus obtain

$$\begin{aligned}E\bigl(\Delta(x) \bigm| \Delta(x) = C(w)\bigr) &= E\bigl(C(w) \bigm| \Delta(x) = C(w)\bigr)\\ &= E\bigl(C(w) \bigm| W = w, U_D = P(z)\bigr)\\ &= C^{\mathrm{MTE}}\bigl(w, P(z)\bigr).\end{aligned} \tag{6.3}$$

Thus,

$$\Delta^{\mathrm{LIV}}\bigl(x, P(z)\bigr) = \Delta^{\mathrm{MTE}}\bigl(x, P(z)\bigr) = C^{\mathrm{MTE}}\bigl(w, P(z)\bigr), \tag{6.4}$$

where $\Delta^{\mathrm{LIV}}(w, P(z))$ is $\Delta^{\mathrm{LIV}}(x, P(z))$ defined for the support where $\Delta(x) = C(w)$. The benefit and cost parameters coincide for the MTE parameter because at the margin, the marginal cost should equal the marginal benefit. The benefit to treatment for an agent indifferent between treatment and no treatment is equal to the cost of treatment, and thus the two parameters coincide.

Suppose that one has access to a large sample of (Y, D, X, W) observations. Since $\Delta^{\text{LIV}}(x, P(z)) = \frac{\partial E(Y|X=x, P(Z)=P(z))}{\partial P(z)}$, $\Delta^{\text{LIV}}(x, P(z))$ can be identified for any $(x, P(z))$ in the support of $(X, P(Z))$, and thus the corresponding $\Delta^{\text{MTE}}(x, P(z))$ and $C^{\text{MTE}}(w, P(z))$ parameters can also be identified.[93] One can thus identify the marginal cost parameter without direct information on the cost of treatment by using the structure of the Roy model and by identifying the marginal benefit parameter.

Heckman and Vytlacil (1999) establish conditions under which Δ^{LIV} can be used to identify Δ^{ATE} and Δ^{TT} given large support conditions, and to bound those parameters without large support conditions if the outcome variables are bounded. We review their results on bounds in Section 10. We surveyed their results on identification of Δ^{ATE} and Δ^{TT} in Sections 3 and 4. From (6.1) and (6.4), we can use the same arguments to use C^{MTE} to identify or bound C^{ATE} and C^{TT}. Thus, C^{MTE} can be used to identify $C^{\text{ATE}}(w)$ if the support of $P(Z)$ conditional on $W = w$ is the full unit interval. If the support of $P(Z)$ conditional on $W = w$ is a proper subset of the full unit interval, then C^{MTE} can be used to bound $C^{\text{ATE}}(x)$ if C is bounded. One can thus identify or bound the average cost of treatment or the cost of treatment on the treated without direct information on the cost of treatment.

We next consider what information is available on the underlying benefit functions μ_0 and μ_1 and the underlying cost function $\mu_C(w)$. From the definitions,

$$\begin{aligned}\Delta^{\text{MTE}}\bigl(x, P(z)\bigr) &= E\bigl(\Delta \bigm| X = x, U_D = P(z)\bigr)\\ &= \mu_1(x) - \mu_0(x) + \Upsilon\bigl(P(z)\bigr)\end{aligned} \tag{6.5}$$

with $\Upsilon(P(z)) = E(U_1 - U_0 \mid U_D = P(z))$. Likewise,

$$\begin{aligned}C^{\text{MTE}}\bigl(w, P(z)\bigr) &= E\bigl(C \bigm| W = w, U_D = P(z)\bigr)\\ &= \mu_C(w) + \Gamma\bigl(P(z)\bigr),\end{aligned} \tag{6.6}$$

with $\Gamma(P(z)) = E(U_C \mid U_D = P(z))$. Let $\Delta^{\text{LIV}}(z) = \Delta^{\text{LIV}}(x, P(z))$, and recall from the preceding analysis that $\Delta^{\text{LIV}}(z) = \Delta^{\text{MTE}}(x, P(z)) = C^{\text{MTE}}(w, P(z))$. Consider two points of evaluation (z, z') such that $P(z) = P(z')$. Using Equation (6.4), we obtain

$$\begin{aligned}\Delta^{\text{LIV}}(z) - \Delta^{\text{LIV}}(z') &= \bigl(\mu_1(x) - \mu_0(x)\bigr) - \bigl(\mu_1(x') - \mu_0(x')\bigr)\\ &= \mu_C(w) - \mu_C(w').\end{aligned}$$

Assuming that X and W each have at least one component not in the other, we can identify $\mu_C(w)$ up to constants within the support of W conditional on $P(Z) = P(z)$ using $\Delta^{\text{LIV}}(z)$. Shifting z while conditioning on $P(z)$ shifts $(\mu_1(x)-\mu_0(x))$ and $\mu_C(w)$ along the line $(\mu_1(x)-\mu_0(x))-\mu_C(w) = F_V^{-1}(p)$. Thus, conditional on $P(z)$, a shift in the benefit, $\mu_1(X)-\mu_0(X)$, is associated with the same shift in the cost, $\mu_C(w)$. For any $p \in (0, 1)$, let $\Omega_p = \{z\colon\ P(z) = p\} = \{(w, x)\colon\ (\mu_1(x)-\mu_0(x))-\mu_C(w) = F_V^{-1}(p)\}$.

[93] Formally, these parameters are identified in the limit points of the set.

As we vary z within the set Ω_p, we trace out changes in $\mu_C(w)$ and $\mu_1(x) - \mu_0(x)$, where the changes in $\mu_C(w)$ equal the changes in $\mu_1(x) - \mu_0(x)$.

For the special case of the generalized Roy model where U_C is degenerate, $\Delta^{\mathrm{LIV}}(z) = \mu_C(w)$. Thus, in the case of a deterministic cost function, LIV identifies $\mu_C(w)$. We plot this case in Figures 5A–5C for the country policy adoption example where the cost C is a constant across all countries.

In the case where U_C is nondegenerate but $U_1 - U_0$ is degenerate, $Y_1 - Y_0 = \mu_1(X) - \mu_0(X)$ ($\beta = \bar{\beta}$ in the context of the model of Section 2), and there is no variation in the gross benefit from participating in the program conditional on X. In that case, $\Delta^{\mathrm{LIV}}(z) = \mu_1(x) - \mu_0(x) = \bar{\beta}$, where we keep the conditioning on X implicit in defining $\Delta^{\mathrm{LIV}}(z)$. Thus, in the case of a deterministic benefit from participation, LIV identifies the benefit function. If U_D and $U_1 - U_0$ are both degenerate, then $\Delta^{\mathrm{LIV}}(z)$ is not well defined.[94]

In summary, the generalized Roy model structure can be exploited to identify cost parameters without direct information on the cost of treatment. The MTE parameter for cost is immediately identified within the proper support, and can be used to identify or bound the average cost of treatment and the cost of treatment on the treated. In addition, the MTE parameter allows one to infer how the cost function shifts in response to a change in observed covariates, and to completely identify the cost function if the cost of treatment is deterministic conditional on observable covariates. Thus we can compute the costs and benefits of alternative programs for various population averages. Heckman and Vytlacil (2007) develop this analysis to consider marginal extensions of the policy relevant treatment effect (PRTE).

6.2. Constructing the PRTE in new environments

In this section, we present conditions for constructing PRTE for new environments and for new programs using historical data for general changes in policies and environments. We consider general changes in the environment and policies and not just the marginal perturbations of the $P(Z)$ considered in the previous section. We address policy problems P-2, forecasting the effects of existing policies to new environments and P-3, forecasting the effects of new policies, never previously implemented.

Let $p \in \mathcal{P}$ denote a policy characterized by random vector Z_p. The usage of "p" in this section is to be distinguished from a realized value of $P(Z)$ as in most other sections in this chapter. Let $e \in \mathcal{E}$ denote an environment characterized by random vector X_e. A history, $\mathcal{H}$, is a collection of policy–environment (p, e) pairs that have been experienced and documented. We assume that the environment is autonomous so

[94] In this case, $E(Y_1 - Y_0 \mid Z = z, D = 0)$ is well defined for $z = (w, x)$ such that $\mu_1(x) - \mu_0(x) \leqslant \mu_C(w)$, in which case $E(Y_1 - Y_0 \mid Z = z, D = 0) = \mu_1(x) - \mu_0(x) \leqslant \mu_C(w)$. Likewise $E(Y_1 - Y_0 \mid Z = z, D = 1)$ is well defined for $z = (w, x)$ such that $\mu_1(x) - \mu_0(x) \geqslant \mu_C(w)$, in which case $E(Y_1 - Y_0 \mid Z = z, D = 1) = \mu_1(x) - \mu_0(x) \geqslant \mu_C(w)$.

the choice of p does not affect X_e. Letting $X_{e,p}$ denote the value of X_e under policy p, autonomy requires that

(A-8) $X_{e,p} = X_e, \ \forall p, e$ (*Autonomy*).

Autonomy is a more general notion than the no-feedback assumption introduced in (A-6). They are the same when the policy is a treatment. General equilibrium feedback effects can cause a failure of autonomy. In this section, we will assume autonomy, in accordance with the partial equilibrium tradition in the treatment effect literature.[95] Autonomy is a version of Hurwicz's policy invariance postulate but for a random variable and not a function.

Evaluating a particular policy p' in environment e' is straightforward if $(p', e') \in \mathcal{H}$. One simply looks at the associated outcomes and treatment effects formed in that policy environment and applies the methods previously discussed to obtain internally valid estimates. The challenge comes in forecasting the impacts of policies (p') in environments (e') for (p', e') not in $\mathcal{H}$.

We show how Δ^{MTE} plays the role of a policy-invariant functional that aids in creating counterfactual states never previously experienced. We focus on the problem of constructing the policy relevant treatment effect Δ^{PRTE} but our discussion applies more generally to the other treatment parameters.

Given the assumptions invoked in Section 3, Δ^{MTE} can be used to evaluate a whole menu of policies characterized by different conditional distributions of $P_{p'}$. In addition, given our assumptions, we can focus on how policy p', which is characterized by $Z_{p'}$, produces the distribution $F_{P_{p'}|X}$ which weights an invariant Δ^{MTE} without having to conduct a new investigation of (Y, X, Z) relationships for each proposed policy.[96]

6.2.1. Constructing weights for new policies in a common environment

The problem of constructing Δ^{PRTE} for policy p' (compared to baseline policy $\bar{p}$) in environment e when $(p', e) \notin \mathcal{H}$ entails constructing $E(\Upsilon(Y_{p'}))$. We maintain the assumption that the baseline policy is observed, so $(\bar{p}, e) \in \mathcal{H}$. We also postulate instrumental variable assumptions (A-1)–(A-5), presented in Section 3, and the policy invariance assumption (A-7), presented in Section 3.2 and embedded in assumption (A-8). We use separable choice Equation (3.3) to characterize choices. The policy is assumed not to change the distribution of (Y_0, Y_1, U_D) conditional on X. Under these conditions, Equation (3.6) is a valid expression for PRTE and constructing PRTE only requires identification of Δ^{MTE} and constructing $F_{P_{p'}|X_e}$ from the policy histories $\mathcal{H}_e$, defined as the elements of $\mathcal{H}$ for a particular environment e, $\mathcal{H}_e = \{p\colon\ (p, e) \in \mathcal{H}\}$.

[95] See Heckman, Lochner and Taber (1998) for an example of a nonautonomous treatment model.

[96] Ichimura and Taber (2002) present a discussion of local policy analysis in a more general framework without the MTE structure, using a framework developed by Hurwicz (1962). We review the Hurwicz framework in Chapter 70.

Associated with the policy histories $p \in \mathcal{H}_e$ is a collection of policy variables $\{Z_p\colon p \in \mathcal{H}_e\}$. Suppose that a new policy p' can be written as $Z_{p'} = T_{p',j}(Z_j)$ for some $j \in \mathcal{H}_e$, where $T_{p',j}$ is a known deterministic transformation and $Z_{p'}$ has the same list of variables as Z_j. Examples of policies that can be characterized in this way are tax and subsidy policies on wages, prices and incomes that affect unit costs (wages or prices) and transfers. Tuition might be shifted upward for everyone by the same amount, or tuition might be shifted according to a nonlinear function of current tuition, parents' income, and other observable characteristics in Z_j.

Constructing $F_{P_{p'}|X_e}$ from data in the policy history entails two distinct steps. From the definitions,

$$\Pr(P_{p'} \leqslant t \mid X_e) = \Pr\bigl(\bigl\{Z_{p'}\colon \Pr(D_{p'} = 1 \bigm| Z_{p'}, X_e) \leqslant t\bigr\} \mid X_e\bigr).$$

If (i) we know the distribution of $Z_{p'}$, and (ii) we know the function $\Pr(D_{p'} = 1 \mid Z_{p'} = z, X_e = x)$ over the appropriate support, we can then recover the distribution of $P_{p'}$ conditional on X_e. Given that $Z_{p'} = T_{p',j}(Z_j)$ for a known function $T_{p',j}(\cdot)$, step (i) is straightforward since we recover the distribution of $Z_{p'}$ from the distribution of Z_j by using the fact that $\Pr(Z_{p'} \leqslant t \mid X_e) = \Pr(\{Z_j\colon T_{p',j}(Z_j) \leqslant t\} \mid X_e)$. Alternatively, part of the specification of the policy p' might be the distribution $\Pr(Z_{p'} \leqslant t \mid X_e)$. We now turn to the second step, recovering the function $\Pr(D_{p'} = 1 \mid Z_{p'} = z, X_e = x)$ over the appropriate support.

If $Z_{p'}$ and Z_j contain the same elements though possibly with different distributions, then a natural approach to forecasting the new policy is to postulate that

$$P_j(z) = \Pr(D_j = 1 \mid Z_j = z, X_e) \tag{6.7}$$

$$= \Pr(D_{p'} = 1 \mid Z_{p'} = z, X_e) = P_{p'}(z), \tag{6.8}$$

i.e., that over a common support for Z_j and $Z_{p'}$ the known conditional probability function and the desired conditional probability function agree. Condition (6.7) will hold, for example, if $D_j = \mathbf{1}[\mu_D(Z_j) - V \geqslant 0]$, $D_{p'} = \mathbf{1}[\mu_D(Z_{p'}) - V \geqslant 0]$, $Z_j \perp\!\!\!\perp V \mid X_e$, and $Z_{p'} \perp\!\!\!\perp U_D \mid X_e$, recalling that $U_D = F_{V|X}(V)$. Even if condition (6.7) is satisfied on a common support, the support of Z_j and $Z_{p'}$ may not be the same. If the support of the distribution of $Z_{p'}$ is not contained in the support of the distribution of Z_j, then some form of extrapolation is needed. Alternatively, if we strengthen our assumptions so that (6.7) holds for all $j \in \mathcal{H}_e$, we can identify $P_{p'}(z)$ for all z in $\bigcup_{j \in \mathcal{H}_e} \operatorname{Supp}(Z_j)$. However, there is no guarantee that the support of the distribution of $Z_{p'}$ will be contained in $\bigcup_{j \in \mathcal{H}_e} \operatorname{Supp}(Z_j)$, in which case some form of extrapolation is needed.

If extrapolation is required, one approach is to assume a parametric functional form for $P_j(\cdot)$. Given a parametric functional form, one can use the joint distribution of (D_j, Z_j) to identify the unknown parameters of $P_j(\cdot)$ and then extrapolate the parametric functional form to evaluate $P_j(\cdot)$ for all evaluation points in the support of $Z_{p'}$.

Alternatively, if there is overlap between the support of $Z_{p'}$ and Z_j,[97] so there is some overlap in the historical and policy p' supports of Z, we may use nonparametric methods presented in Matzkin (1994) and extended by her in Chapter 73 (Matzkin) of this Handbook, based on functional restrictions (e.g., homogeneity) to construct the desired probabilities on new supports or to bound them. Under the appropriate conditions, we may use analytic continuation to extend $\Pr(D_j = 1 \mid Z_j = z, X_e = x)$ to a new support for each $X_e = x$ [Rudin (1974)].

The approach just presented is based on the assumption stated in Equation (6.7). That assumption is quite natural when $Z_{p'}$ and Z_j both contain the same elements, say they both contain tuition and parent's income. However, in some cases $Z_{p'}$ might contain additional elements not contained in Z_j. As an example, $Z_{p'}$ might include new user fees while Z_j consists of taxes and subsidies but does not include user fees. In this case, the assumption stated in Equation (6.7) is not expected to hold and is not even well defined if $Z_{p'}$ and Z_j contain a different number of elements.

A more basic approach analyzes a class of policies that operate on constraints, prices and endowments arrayed in vector Q. Given the preferences and technology of the agent, a given $Q = q$, however arrived at, generates the same choices for the agent. Thus a wage tax offset by a wage subsidy of the same amount produces a wage that has the same effect on choices as a no-policy wage. Policy j affects Q (e.g., it affects prices paid, endowments and constraints). Define a map $\Phi_j : Z_j \to Q_j$ which maps a policy j, described by Z_j, into its consequences (Q_j) for the baseline, fixed-dimensional vector Q. A new policy p', characterized by $Z_{p'}$, produces $Q_{p'}$ that is possibly different from Q_j for all previous policies $j \in \mathcal{H}_e$.

To construct the random variable $P_{p'} = \Pr(D_{p'} = 1 \mid Z_{p'}, X_e)$, we postulate that

$$\begin{aligned}\Pr\bigl(D_j = 1 \bigm| Z_j \in \Phi_j^{-1}(q), X_e = x\bigr) &= \Pr(D_j = 1 \mid Q_j = q, X_e = x)\\ &= \Pr(D_{p'} = 1 \mid Q_{p'} = q, X_e = x)\\ &= \Pr\bigl(D_{p'} = 1 \bigm| Z_{p'} \in \Phi_{p'}^{-1}(q), X_e = x\bigr),\end{aligned}$$

where $\Phi_j^{-1}(q) = \{z\colon \Phi_j(z) = q\}$ and $\Phi_{p'}^{-1}(q) = \{z\colon \Phi_{p'}(z) = q\}$. Given these assumptions, our ability to recover $\Pr(D_{p'} = 1 \mid Z_{p'} = z, X_e = x)$ for all (z, x) in the support of $(Z_{p'}, X_e)$ depends on what Φ_j functions have been historically observed and the richness of the histories of Q_j, $j \in \mathcal{H}_e$. For each $z_{p'}$ evaluation point in the support of the distribution of $Z_{p'}$, there is a corresponding $q = \Phi_{p'}(z_{p'})$ evaluation point in the support of the distribution of $Q_j = \Phi_j(Z_j)$. If, in the policy histories, there is at least one $j \in \mathcal{H}_e$ such that $\Phi_j(z_j) = q$ for a z_j with (z_j, x) in the support of the distribution of (Z_j, X_e), then we can construct the probability of the new policy from data in the policy histories. The methods used to extrapolate $P_{p'}(\cdot)$ over new regions, discussed previously, apply here. If the distribution of $Q_{p'}$ (or $\Phi_{p'}$ and the distribution

[97] If we strengthen condition (6.7) to hold for all $j \in \mathcal{H}_e$, then the condition becomes that $\mathrm{Supp}(Z_{p'}) \cap \bigcup_{j \in \mathcal{H}_e} \mathrm{Supp}(Z_j)$ is not empty.

of $Z_{p'}$) is known as part of the specification of the proposed policy, the distribution of $F_{P_{p'}|X_e}$ can be constructed using the constructed $P_{p'}$. Alternatively, if we can relate $Q_{p'}$ to Q_j by $Q_{p'} = \Psi_{p',j}(Q_j)$ for a known function $\Psi_{p',j}$ or if we can relate $Z_{p'}$ to Z_j by $Z_{p'} = T_{p',j}(Z_j)$ for a known function $T_{p',j}$, and the distributions of Q_j and/or Z_j are known for some $j \in \mathcal{H}_e$, we can apply the method previously discussed to derive $F_{P_{p'}|X_e}$ and hence the policy weights for the new policy.

This approach assumes that a new policy acts on components of Q like a policy in $\mathcal{H}_e$, so it is possible to forecast the effect of a policy with nominally new aspects. The essential idea is to recast the new aspects of policy in terms of old aspects previously measured. Thus in a model of schooling, let $D = \mathbf{1}[Y_1 - Y_0 - B \geqslant 0]$ where $Y_1 - Y_0$ is the discounted gain in earnings from going to school and B is the tuition cost. In this example, a decrease in a unit of cost (B) has the same effect on choice as an increase in return ($Y_1 - Y_0$). Historically, we might only observe variation in $Y_1 - Y_0$ (say tuition has never previously been charged). But B is on the same footing (has the same effect on choice, except for sign) as $Y_1 - Y_0$. The identified historical variation in $Y_1 - Y_0$ can be used to nonparametrically forecast the effect of introducing B, provided that the support of $P_{p'}$ is in the historical support generated by the policy histories in $\mathcal{H}_e$. Otherwise, some functional structure (parametric or semiparametric) must be imposed to solve the support problem for $P_{p'}$. We used this basic principle in constructing our econometric cost benefit analysis in Section 6.1.

As another example, following Marschak (1953), consider the introduction of wage taxes in a world where there has never before been a tax. This example is analyzed in Heckman (2001). Let Z_j be the wage without taxes. We seek to forecast a post-tax net wage $Z_{p'} = (1 - \tau)Z_j + b$ where τ is the tax rate and b is a constant shifter. Thus $Z_{p'}$ is a known linear transformation of policy Z_j. We can construct $Z_{p'}$ from Z_j. We can forecast under (A-1) using $\Pr(D_j = 1 \mid Z_j = z) = \Pr(D_{p'} = 1 \mid Z_{p'} = z)$. This assumes that the response to after tax wages is the same as the response to wages at the after tax level. The issue is whether $P_{p'|X_e}$ lies in the historical support, or whether extrapolation is needed. Nonlinear versions of this example can be constructed.

As a final example, environmental economists use variation in one component of cost (e.g., travel cost) to estimate the effect of a new cost (e.g., a park registration fee). See Smith and Banzhaf (2004). Relating the costs and characteristics of new policies to the costs and characteristics of old policies is a standard, but sometimes controversial, method for forecasting the effects of new policies.

In the context of our model, extrapolation and forecasting are confined to constructing $P_{p'}$ and its distribution. If policy p', characterized by vector $Z_{p'}$, consists of new components that cannot be related to Z_j, $j \in \mathcal{H}_e$, or a base set of characteristics whose variation cannot be identified, the problem is intractable. Then $P_{p'}$ and its distribution cannot be formed using econometric methods applied to historical data.

When it can be applied, our approach allows us to simplify the policy forecasting problem and concentrate our attention on forecasting choice probabilities and their distribution in solving the policy forecasting problem. We can use choice theory and choice data to construct these objects to forecast the impacts of new policies, by relating new policies to previously experienced policies.

6.2.2. *Forecasting the effects of policies in new environments*

When the effects of policy p are forecast for a new environment e' from baseline environment e, and $X_e \neq X_{e'}$, in general both $\Delta^{\mathrm{MTE}}(x, u_D)$ and $F_{P_p|X_e}$ will change. In general, neither object is environment invariant.[98] The new $X_{e'}$ may have a different support than X_e or any other environment in $\mathcal{H}$. In addition, the new $(X_{e'}, U_D)$ stochastic relationship may be different from the historical (X_e, U_D) stochastic relationship. Constructing $F_{P_p|X_{e'}}$ from $F_{P_p|X_e}$ and $F_{Z_p|X_{e'}}$ from $F_{Z_p|X_e}$ can be done using (i) functional form (including semiparametric functional restrictions) or (ii) analytic continuation methods. Notice that the maps $T_{p,j}$ and Φ_p may depend on X_e and so the induced changes in these transformations must also be modeled. There is a parallel discussion for $\Delta^{\mathrm{MTE}}(x, u_D)$. The stochastic dependence between $X_{e'}$ and (U_0, U_1, U_D) may be different from the stochastic dependence between X_e and (U_0, U_1, U_D). We suppress the dependence of U_0 and U_1 on e and p only for convenience of exposition and make it explicit in the next paragraph.

Forecasting new stochastic relationships between $X_{e'}$ and (U_1, U_0, U_D) is a difficult task. Some of the difficulty can be avoided if we invoke the traditional exogeneity assumptions of classical econometrics:

(A-9) $(U_{0,e,p}, U_{1,e,p}, U_{D,e,p}) \perp\!\!\!\perp (X_e, Z_p)\ \forall e, p.$

Under (A-9), we only encounter the support problems for Δ^{MTE} and the distribution of $\Pr(D_p = 1 \mid Z_p, X_e)$ in constructing policy counterfactuals.

Conditions (A-7)–(A-9) are unnecessary if the only goal of the analysis is to establish internal validity, the standard objective of the treatment effect literature. This is problem P-1. Autonomy and exogeneity conditions become important issues if we seek external validity. An important lesson from this analysis is that as we try to make the treatment effect literature do the tasks of structural econometrics (i.e., make out-of-sample forecasts), common assumptions are invoked in the two literatures.

6.2.3. *A comparison of three approaches to policy evaluation*

Table 9 compares the strengths and limitations of the three approaches to policy evaluation that we have discussed in this Handbook chapter and our contribution in Chapter 70: the structural approach, the conventional treatment effect approach, and the approach to treatment effects based on the MTE function developed by Heckman and Vytlacil (1999, 2001b, 2005).

[98] We suppress the dependence of U_D on p for notational convenience.

Table 9
Comparison of alternative approaches to program evaluation

	Structural econometric approach	Treatment effect approach	Approach based on MTE
Interpretability	Well defined economic parameters and welfare comparisons	Link to economics and welfare comparisons obscure	Interpretable in terms of willingness to pay; weighted averages of the MTE answer well-posed economic questions
Range of questions addressed	Answers many counterfactual questions	Focuses on one treatment effect or narrow range of effects	With support conditions, generates all treatment parameters
Extrapolation to new environments	Provides ingredients for extrapolation	Evaluates one program in one environment	Can be partially extrapolated; extrapolates to new policy environments with different distributions of the probability of participation due solely to differences in distributions of Z
Comparability across studies	Policy invariant parameters comparable across studies	Not generally comparable	Partially comparable; comparable across environments with different distributions of the probability of participation due solely to differences in distributions of Z
Key econometric problems	Exogeneity, policy invariance and selection bias	Selection bias	Selection bias: Exogeneity and policy invariance if used for forecasting
Range of policies that can be evaluated	Programs with either partial or universal coverage, depending on variation in data (prices/endowments)	Programs with partial coverage (treatment and control groups)	Programs with partial coverage (treatment and control groups)
Extension to general equilibrium evaluation	Need to link to time series data; parameters compatible with general equilibrium theory	Difficult because link to economics is not precisely specified	Can be linked to nonparametric general equilibrium models under exogeneity and policy invariance

Source: Heckman and Vytlacil (2005).

The approach based on the MTE function and the structural approach share interpretability of parameters. Like the structural approach, it addresses a range of policy evaluation questions. The MTE parameter is less comparable and less easily extrapolated across environments than are structural parameters, unless nonparametric versions of invariance and exogeneity assumptions are made. However, Δ^{MTE} is comparable across populations with different distributions of P (conditional on X_e) and results from one population can be applied to another population under the conditions presented in this section. Analysts can use Δ^{MTE} to forecast a variety of policies. This invariance property is shared with conventional structural parameters. Our framework solves the problem of external validity, which is ignored in the standard treatment effect approach. The price of these advantages of the structural approach is the greater range of econometric problems that must be solved. They are avoided in the conventional treatment approach at the cost of producing parameters that cannot be linked to well-posed economic models and hence do not provide building blocks for an empirically motivated general equilibrium analysis or for investigation of the impacts of new public policies. Δ^{MTE} estimates the preferences of the agents being studied and provides a basis for integration with well posed economic models. If the goal of a study is to examine one policy in place (the problem of internal validity), the stronger assumptions invoked in this section of the chapter, and in structural econometrics, are unnecessary. Even if this is the only goal of the analysis however, our approach allows the analyst to generate all treatment effects and IV estimands from a common parameter and provides a basis for unification of the treatment effect literature.

7. Extension of MTE to the analysis of more than two treatments and associated outcomes

We have thus far analyzed models with two potential outcomes associated with receipt of binary treatments ($D = 0$ or $D = 1$). Focusing on this simple case allows us to develop main ideas. However, models with more than two outcomes are common in empirical work. Angrist and Imbens (1995) analyze an ordered choice model with a single instrument that shifts people across all margins. We generalize their analysis in several ways. We consider vectors of instruments, some of which may affect choices at all margins and some of which affect choices only at certain margins. We then analyze a general unordered choice model.

7.1. Background for our analysis of the ordered choice model

Angrist and Imbens (1995) extend their analysis of LATE to an ordered choice model with outcomes generated by a scalar instrument that can assume multiple values. From their analysis of the effect of schooling on earnings, it is unclear even under a strengthened “monotonicity” condition whether IV estimates the effect of a change of schooling on earnings for a well defined margin of choice.

To summarize their analysis, let $\bar{S}$ be the number of possible outcome states with associated outcomes Y_s and choice indicators D_s, $s = 1, \dots, \bar{S}$. The s, in their analysis, correspond to different levels of schooling. For any two instrument values $Z = z_i$ and $Z = z_j$ with $z_i > z_j$, we can define associated indicators $\{D_s(z_i)\}_{s=1}^{\bar{S}}$ and $\{D_s(z_j)\}_{s=1}^{\bar{S}}$, where $D_s(z_i) = 1$ if a person assigned instrument value z_i chooses state s. As in the two-outcome model, the instrument Z is assumed to be independent of the potential outcomes $\{Y_s\}_{s=1}^{\bar{S}}$ as well as the associated indicator functions defined by fixing Z at z_i

and z_j. Observed schooling for instrument z_j is $S(z_j) = \sum_{s=1}^{\bar{S}} s D_s(z_j)$. Observed outcomes with this instrument are $Y(z_j) = \sum_{s=1}^{\bar{S}} Y_s D_s(z_j)$.

Angrist and Imbens show that IV (with $Z = z_i$ and $Z = z_j$) applied to S in a two stage least squares regression of Y on S identifies a "causal parameter"

$$\Delta^{\mathrm{IV}} = \sum_{s=2}^{\bar{S}} \bigl\{E\bigl(Y_s - Y_{s-1} \bigm| S(z_i) \geqslant s > S(z_j)\bigr)\bigr\} \frac{\Pr(S(z_i) \geqslant s > S(z_j))}{\sum_{s=2}^{\bar{S}} \Pr(S(z_i) \geqslant s > S(z_j))}. \tag{7.1}$$

This "causal parameter" is a weighted average of the gross returns from going from $s-1$ to s for persons induced by the change in the instrument to move from *any* schooling level below s to *any* schooling level s or above. Thus the conditioning set defining the sth component of IV includes people who have schooling below $s - 1$ at instrument value $Z = z_j$ and people who have schooling above level s at instrument value $Z = z_i$. In expression (7.1), the average return experienced by some of the people in the conditioning set for each component conditional expectation does not correspond to the average outcome corresponding to the gain in the argument of the expectation. In the case where $\bar{S} = 2$, agents face only two choices and the margin of choice is well defined. Agents in each conditioning set are at different margins of choice. The weights are positive but, as noted by Angrist and Imbens (1995), persons can be counted multiple times in forming the weights. When they generalize their analysis to multiple-valued instruments, they use the Yitzhaki (1989) weights.

Whereas the weights in Equation (7.1) can be constructed empirically using nonparametric discrete choice theory (see, e.g., our analysis in Appendix B of Chapter 70 or the contribution of Matzkin to this Handbook), the terms in braces cannot be identified by any standard IV procedure.[99] We present decompositions with components that are recoverable, whose weights can be estimated from the data and that are economically interpretable.

In this section, we generalize LATE to a multiple outcome case where we can identify agents at different well-defined margins of choice. Specifically, we (1) analyze both ordered and unordered choice models; (2) analyze outcomes associated with choices at various well-defined margins; and (3) develop models with multiple instruments that can affect different margins of choice differently. With our methods, we can define and estimate a variety of economically interpretable parameters. In contrast, the Angrist–Imbens analysis produces a single "causal parameter" (7.1) that does not answer any well-defined policy question such as that posed by the PRTE. We first consider an explicit ordered choice model and decompose the IV into policy-useful (identifiable) components.

[99] It can be identified by a structural model using the methods surveyed in Chapter 72.

7.2. Analysis of an ordered choice model

Ordered choice models arise in many settings. In schooling models, there are multiple grades. One has to complete grade $s-1$ to proceed to grade s. The ordered choice model has been widely used to fit data on schooling transitions [Harmon and Walker (1999), Cameron and Heckman (1998)]. Its nonparametric identifiability has been studied [Carneiro, Hansen and Heckman (2003), Cunha, Heckman and Navarro (2007)]. It can also be used as a duration model for dynamic treatment effects with associated outcomes as in Cunha, Heckman and Navarro (2007). It also represents the "vertical" model of the choice of product quality [Prescott and Visscher (1977), Shaked and Sutton (1982), Bresnahan (1987)].[100]

Our analysis generalizes the analysis for the binary model in a parallel way. Write potential outcomes as

$$Y_s = \mu_s(X, U_s), \quad s = 1, \dots, \bar{S}.$$

The $\bar{S}$ could be different schooling levels or product qualities. We define latent variables $D_S^* = \mu_D(Z) - V$ where

$$D_s = \mathbf{1}\big[C_{s-1}(W_{s-1}) < \mu_D(Z) - V \leqslant C_s(W_s)\big], \quad s = 1, \dots, \bar{S},$$

and the cutoff values satisfy

$$C_{s-1}(W_{s-1}) \leqslant C_s(W_s), \quad C_0(W_0) = -\infty \quad \text{and} \quad C_{\bar{S}}(W_{\bar{S}}) = \infty.$$

The cutoffs used to define the intervals are allowed to depend on observed (by the economist) regressors W_s. In Appendix G we extend the analysis presented in the text to allow the cutoffs to depend on unobserved regressors as well, following structural analysis along these lines by Carneiro, Hansen and Heckman (2003) and Cunha, Heckman and Navarro (2007). Observed outcomes are: $Y = \sum_{s=1}^{\bar{S}} Y_s D_s$. The Z shift the index generally; the W_s affect s-specific transitions. Thus, in a schooling example, Z could include family background variables while W_s could include college tuition or opportunity wages for unskilled labor.[101] Collect the W_s into $W = (W_1, \dots, W_{\bar{S}})$, and the U_s into $U = (U_1, \dots, U_{\bar{S}})$. Larger values of $C_s(W_s)$ make it more likely that $D_s = 1$. The inequality restrictions on the $C_s(W_s)$ functions play a critical role in defining the model and producing its statistical implications.

[100] Cunha, Heckman and Navarro (2007) analyze a dynamic discrete choice setting with sequential revelation of information.

[101] Many of the instruments studied by Harmon and Walker (1999) and Card (2001) are transition-specific. Card's model of schooling is not sufficiently rich to make a distinction between the Z and the W. See Heckman and Navarro (2007) and Cunha, Heckman and Navarro (2007) for more general models of schooling that make these distinctions explicit.

Analogous to the assumptions made for the binary outcome model, we assume

(OC-1) $(U_s, V) \perp\!\!\!\perp (Z, W) \mid X, s = 1, \ldots, \bar{S}$ (*Conditional independence of the instruments*);

(OC-2) $\mu_D(Z)$ *is a nondegenerate random variable conditional on* X *and* W (*Rank condition*);

(OC-3) *the distribution of* V *is continuous*[102];

(OC-4) $E(|Y_s|) < \infty, s = 1, \ldots, \bar{S}$ (*Finite means*);

(OC-5) $0 < \Pr(D_s = 1 \mid X) < 1$ *for* $s = 1, \ldots, \bar{S}$, *for all* X (*In large samples, there are some persons in each treatment state*);

(OC-6) *for* $s = 1, \ldots, \bar{S} - 1$, *the distribution of* $C_s(W_s)$ *conditional on* X, Z *and the other* $C_j(W_j)$, $j = 1, \ldots, \bar{S}$, $j \neq s$, *is nondegenerate and continuous.*[103]

Assumptions (OC-1)–(OC-5) play roles analogous to their counterparts in the two-outcome model, (A-1)–(A-5). (OC-6) is a new condition that is key to identification of the Δ^{MTE} defined below for each transition. It assumes that we can vary the choice sets of agents at different margins of schooling choice without affecting other margins of choice. A necessary condition for (OC-6) to hold is that at least one element of W_s is nondegenerate and continuous conditional on X, Z and $C_j(W_j)$ for $j \neq s$. Intuitively, one needs an instrument (or source of variability) for each transition. The continuity of the regressor allows us to differentiate with respect to $C_s(W_s)$, like we differentiated with respect to $P(Z)$ to estimate the MTE in the analysis of the two-outcome model.

The analysis of Angrist and Imbens (1995) discussed in the introduction to this section makes independence and monotonicity assumptions that generalize their earlier work. They do not consider estimation of transition-specific parameters as we do, or even transition-specific LATE. We present a different decomposition of the IV estimator where each component can be recovered from the data, and where the transition-specific MTEs answer well-defined and economically interpretable policy evaluation questions.[104]

The probability of $D_s = 1$ given X, Z and W is generated by an ordered choice model:

$$\begin{aligned}\Pr(D_s = 1 \mid Z, W, X) &\equiv P_s(Z, W, X)\\ &= \Pr\bigl(C_{s-1}(W_{s-1}) < \mu_D(Z) - V \leqslant C_s(W_s) \bigm| X\bigr).\end{aligned}$$

Analogous to the binary case, we can define $U_D = F_{V|X}(V)$ so $U_D \sim \text{Unif}[0, 1]$ under our assumption that the distribution of V is absolutely continuous with respect to Lebesgue measure. The probability integral transformation used extensively

[102] Absolutely continuous with respect to Lebesgue measure.

[103] Absolutely continuous with respect to Lebesgue measure.

[104] Vytlacil (2006b) shows that their monotonicity and independence conditions imply (and are implied by) a more general version of the ordered choice model with stochastic thresholds, which appears in Heckman, LaLonde and Smith (1999), Carneiro, Hansen and Heckman (2003), and Cunha, Heckman and Navarro (2007), and is analyzed in Appendix G.

in the binary choice model is somewhat less useful for analyzing ordered choices, so we work with both U_D and V in this section of the chapter. Monotonic transformations of V induce monotonic transformations of $\mu_D(Z) - C_s(W_s)$, but one is not free to form arbitrary monotonic transformations of $\mu_D(Z)$ and $C_s(W_s)$ separately. Using the probability integral transformation, the expression for choice s is $D_s = \mathbf{1}[F_{V|X}(\mu_D(Z) - C_{s-1}(W_{s-1})) > U_D \geqslant F_{V|X}(\mu_D(Z) - C_s(W_s))]$. Keeping the conditioning on X implicit, we define $P_s(Z, W) = F_V(\mu_D(Z) - C_{s-1}(W_{s-1})) - F_V(\mu_D(Z) - C_s(W_s))$. It is convenient to work with the probability that $S > s$, $\pi_s(Z, W_s) = F_V(\mu_D(Z) - C_s(W_s)) = \Pr(\sum_{j=s+1}^{\bar{S}} D_j = 1 \mid Z, W_s)$, $\pi_{\bar{S}}(Z, W_{\bar{S}}) = 0$, $\pi_0(Z, W_0) = 1$ and $P_s(Z, W) = \pi_{s-1}(Z, W_{s-1}) - \pi_s(Z, W_s)$.

The transition-specific Δ^{MTE} for the transition from s to $s + 1$ is defined in terms of U_D:

$$\Delta^{\text{MTE}}_{s,s+1}(x, u_D) = E(Y_{s+1} - Y_s \mid X = x, U_D = u_D), \quad s = 1, \ldots, \bar{S} - 1.$$

Alternatively, one can condition on V. Analogous to the analysis of the earlier sections of this chapter, when we set $u_D = \pi_s(Z, W_s)$, we obtain the mean return to persons indifferent between s and $s + 1$ at mean level of utility $\pi_s(Z, W_s)$.

In this notation, keeping X implicit, the mean outcome Y, conditional on (Z, W), is the sum of the mean outcomes conditional on each state weighted by the probability of being in each state summed over all states:

$$\begin{aligned} E(Y \mid Z, W) &= \sum_{s=1}^{\bar{S}} E(Y_s \mid D_s = 1, Z, W) \Pr(D_s = 1 \mid Z, W) \\ &= \sum_{s=1}^{\bar{S}} \int_{\pi_s(Z, W_s)}^{\pi_{s-1}(Z, W_{s-1})} E(Y_s \mid U_D = u_D)\, du_D, \end{aligned} \tag{7.2}$$

where we use conditional independence assumption (OC-1) to obtain the final expression. Analogous to the result for the binary outcome model, we obtain the index sufficiency restriction $E(Y \mid Z, W) = E(Y \mid \pi(Z, W))$, where $\pi(Z, W) = [\pi_1(Z, W_1), \ldots, \pi_{\bar{S}-1}(Z, W_{\bar{S}-1})]$. The choice probabilities encode all of the influence of (Z, W) on outcomes.

We can identify $\pi_s(z, w_s)$ for (z, w_s) in the support of the distribution of (Z, W_s) from the relationship $\pi_s(z, w_s) = \Pr(\sum_{j=s+1}^{\bar{S}} D_j = 1 \mid Z = z, W_s = w_s)$. Thus $E(Y \mid \pi(Z, W) = \pi)$ is identified for all π in the support of $\pi(Z, W)$. Assumptions (OC-1), (OC-3), and (OC-4) imply that $E(Y \mid \pi(Z, W) = \pi)$ is differentiable in π. So $\frac{\partial}{\partial \pi} E(Y \mid \pi(Z, W) = \pi)$ is well defined.[105] Thus analogous to the result obtained in

[105] For almost all π that are limit points of the support of distribution of $\pi(Z, W)$, we use the Lebesgue theorem for the derivative of an integral. Under assumption (OC-6), all points in the support of the distribution of $\pi(Z, W)$ will be limit points of that support, and we thus have that $\frac{\partial}{\partial \pi} E(Y \mid \pi(Z, W) = \pi)$ is well defined and is identified for (a.e.) π.

the binary case

$$\begin{aligned}\frac{\partial E(Y \mid \pi(Z, W) = \pi)}{\partial \pi_s} &= \Delta^{\mathrm{MTE}}_{s,s+1}(U_D = \pi_s) \\ &= E(Y_{s+1} - Y_s \mid U_D = \pi_s).\end{aligned} \tag{7.3}$$

Equation (7.3) is the basis for identification of the transition-specific MTE from data on (Y, Z, X).

From index sufficiency, we can express (7.2) as

$$\begin{aligned}E\bigl(Y \bigm| \pi(Z, W) = \pi\bigr) &= \sum_{s=1}^{\bar{S}} E(Y_s \mid \pi_s \leqslant U_D < \pi_{s-1})(\pi_{s-1} - \pi_s) \\ &= \sum_{s=1}^{\bar{S}-1} \bigl[E(Y_{s+1} \mid \pi_{s+1} \leqslant U_D < \pi_s) \\ &\qquad - E(Y_s \mid \pi_s \leqslant U_D < \pi_{s-1})\bigr]\pi_s \\ &\qquad + E(Y_1 \mid \pi_1 \leqslant U_D < 1) \\ &= \sum_{s=1}^{\bar{S}-1} \bigl\{m_{s+1}(\pi_{s+1}, \pi_s) - m_s(\pi_s, \pi_{s-1})\bigr\}\pi_s \\ &\quad + E(Y_1 \mid \pi_1 \leqslant U_D < 1),\end{aligned} \tag{7.4}$$

where $m_s(\pi_s, \pi_{s-1}) = E[Y_s \mid \pi_s \leqslant U_D < \pi_{s-1}]$. In general, this expression is a nonlinear function of (π_s, π_{s-1}). This model has a testable restriction of index sufficiency in the general case: $E(Y \mid \pi(Z, W) = \pi)$ is a nonlinear function that is additive in functions of (π_s, π_{s-1}) so there are no interactions between π_s and $\pi_{s'}$ if $|s - s'| > 1$, i.e.,

$$\frac{\partial^2 E(Y \mid \pi(Z, W) = \pi)}{\partial \pi_s \partial \pi_{s'}} = 0 \quad \text{if } |s - s'| > 1.$$

Observe that if $U_D \perp\!\!\!\perp U_s$ for $s = 1, \ldots, \bar{S}$,

$$\begin{aligned}E\bigl(Y \bigm| \pi(Z, W) = \pi\bigr) &= \sum_{s=1}^{\bar{S}} E(Y_s)(\pi_{s-1} - \pi_s) \\ &= \sum_{s=1}^{\bar{S}-1} \bigl[E(Y_{s+1}) - E(Y_s)\bigr]\pi_s + E(Y_1).\end{aligned}$$

Defining $E(Y_{s+1}) - E(Y_s) = \Delta^{\mathrm{ATE}}_{s,s+1}$, $E(Y \mid \pi(Z, W) = \pi) = \sum_{s=1}^{\bar{S}-1} \Delta^{\mathrm{ATE}}_{s,s+1}\pi_s + E(Y_1)$. Thus, under full independence, we obtain linearity of the conditional mean of Y in the π_s, $s = 1, \ldots, \bar{S}$. This result generalizes the test for the presence of essential heterogeneity presented in Section 4 to the ordered case. We can ignore the complexity induced by the model of essential heterogeneity if $E(Y \mid \pi(Z, W) = \pi)$ is linear in the π_s and can use conventional IV estimators to identify well-defined treatment effects.[106]

7.2.1. The policy relevant treatment effect for the ordered choice model

The policy relevant treatment effect compares the mean outcome under one policy regime p with the mean outcome under policy regime p'. It is defined analogously to the way it is defined in the binary case in Section 3.2 and in Heckman and Vytlacil (2001c, 2005). Policies (p, p') are assumed to induce different distributions of (Z, W), $F^p(Z, W)$. Forming $E_p(Y) = \int E(Y \mid Z = z, W = w)\, dF^p_{Z,W}(z, w)$ for each policy p, the policy relevant treatment effect is $E_{p'}(Y) - E_p(Y)$.

We can represent the PRTE as a weighted average of pairwise MTE:

$$\Delta^{\text{PRTE}}_{p,p'} = E_{p'}(Y) - E_p(Y) = \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s \mid V = v)\omega_{p,p'}(v)\, dF(v). \tag{7.5}$$

The weights are known functions of the data. See Appendix H for a derivation of the weights and expression (7.5). Using the probability integral transform, we can alternatively express this in terms of $U_D = F_{V|X}(V)$.

7.2.2. What do instruments identify in the ordered choice model?

We now characterize what scalar instrument $J(Z, W)$ identifies. When Y is log earnings, it is common practice to regress Y on S where S is completed years of schooling and call the coefficient on S a rate of return.[107] We seek an expression for the instrumental variables estimator of the effect of S on Y in the ordered choice model:

$$\frac{\operatorname{Cov}(J(Z, W), Y)}{\operatorname{Cov}(J(Z, W), D)}, \tag{7.6}$$

where $S = \sum_{s=1}^{\bar{S}} sD_s$ is the number of years of schooling attainment. We keep the conditioning on X implicit. We now analyze the weights for IV. Their full derivation is presented in Appendix I.

Define $K_s(v) = E(\tilde{J}(Z, W) \mid \mu_D(Z) - C_s(W_s) > v) \Pr(\mu_D(Z) - C_s(W_s) > v)$, where $\tilde{J}(Z, W) = J(Z, W) - E(J(Z, W))$. Thus,

$$\begin{aligned}\Delta^{\text{IV}}_J &= \frac{\operatorname{Cov}(J, Y)}{\operatorname{Cov}(J, S)} \\ &= \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s \mid V = v)\omega(s, v) f_V(v)\, dv, \end{aligned} \tag{7.7}$$

[106] Notice that if $U_D \not\perp\!\!\!\perp U_s$ for some s, then we obtain an expression with nonlinearities in (π_s, π_{s-1}) in expression (7.4).

[107] Heckman, Lochner and Todd (2006) present conditions under which this economic interpretation is valid.

where

$$\omega(s, v) = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}} s \int [K_{s-1}(v) - K_s(v)] f_V(v)\, dv} = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}-1} \int K_s(v) f_V(v)\, dv},$$

and clearly $\sum_{s=1}^{\bar{S}-1} \int \omega(s, v) f_V(v)\, dv = 1$, $\omega(0, v) = 0$, and $\omega(\bar{S}, v) = 0$. We can rewrite this result in terms of the MTE, expressed in terms of u_D

$$\Delta_{s,s+1}^{\mathrm{MTE}}(u_D) = E(Y_{s+1} - Y_s \mid U_D = u_D)$$

so that

$$\frac{\operatorname{Cov}(J, Y)}{\operatorname{Cov}(J, S)} = \sum_{s=1}^{\bar{S}-1} \int_0^1 \Delta_{s,s+1}^{\mathrm{MTE}}(u_D) \tilde{\omega}(s, u_D)\, du_D,$$

where

$$\tilde{\omega}(s, u_D) = \frac{\tilde{K}_s(u_D)}{\sum_{s=1}^{\bar{S}} s \int_0^1 [\tilde{K}_{s-1}(u_D) - \tilde{K}_s(u_D)]\, du_D} = \frac{\tilde{K}_s(u_D)}{\sum_{s=1}^{\bar{S}-1} \int_0^1 \tilde{K}_s(u_D)\, du_D} \tag{7.8}$$

and

$$\tilde{K}_s(u_D) = E\big(\tilde{J}(Z, W) \bigm| \pi_s(Z, W_s) \geqslant u_D\big) \Pr\big(\pi_s(Z, W_s) \geqslant u_D\big). \tag{7.9}$$

Compare Equations (7.8) and (7.9) for the ordered choice model to Equations (4.13) and (4.14) for the binary choice model. The numerator of the weights for the Δ^{MTE} in the ordered choice model for a particular transition is exactly the numerator of the weights for the binary choice model, substituting $\pi_s(Z, W_s) = \Pr(S > s \mid Z, W_s)$ for $P(Z) = \Pr(D = 1 \mid Z)$. The numerator for the weights for IV in the binary choice model is driven by the connection between the instrument and $P(Z)$. The numerator for the weights for IV in the ordered choice model for a particular transition is driven by the connection between the instrument and $\pi_s(Z, W_s)$. The denominator of the weights is the covariance between the instrument and D (or S) for the binary (or ordered) case, respectively. However, in the binary case the covariance between the instrument and D is completely determined by the covariance between the instrument and $P(Z)$, while in the ordered choice case the covariance with S depends on the relationship between the instrument and the full vector $[\pi_1(Z, W_1), \ldots, \pi_{\bar{S}-1}(Z, W_{\bar{S}-1})]$. Comparing our decomposition of Δ^{IV} to decomposition (7.1), ours corresponds to weighting up marginal outcomes across well-defined and adjacent boundary values experienced by agents having their instruments manipulated whereas the Angrist–Imbens decomposition corresponds to outcomes not experienced by some of the persons whose instruments are being manipulated.

From Equation (7.9), the IV estimator using $J(Z, W)$ as an instrument satisfies the following properties. (a) The numerator of the weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$ is nonnegative for all u_D if $E(J(Z, W_s) \mid \pi_s(Z, W_s) \geqslant \pi_s)$ is weakly monotonic in π_s. For example, if $\operatorname{Cov}(\pi_s(Z, W_s), S) > 0$, setting $J(Z, W) = \pi_s(Z, W_s)$ will lead to nonnegative weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$, though it may lead to negative weights on other transitions. A second property (b) is that the support of the weights on $\Delta^{\text{MTE}}_{s,s+1}$ using $\pi_s(Z, W_s)$ as the instrument is $(\pi_s^{\text{Min}}, \pi_s^{\text{Max}})$ where π_s^{Min} and π_s^{Max} are the minimum and maximum values in the support of $\pi_s(Z, W_s)$, respectively, and the support of the weights on $\Delta^{\text{MTE}}_{s,s+1}$ using any other instrument is a subset of $(\pi_s^{\text{Min}}, \pi_s^{\text{Max}})$. A third property (c) is that the weights on $\Delta^{\text{MTE}}_{s,s+1}$ implied by using $J(Z, W)$ as an instrument are the same as the weights on $\Delta^{\text{MTE}}_{s,s+1}$ implied by using $E(J(Z, W) \mid \pi_s(Z, W_s))$ as the instrument.

Our analysis generalizes that of Imbens and Angrist (1994) and Angrist and Imbens (1995) by considering multiple instruments and by introducing both transition-specific instruments (W) and general instruments (Z) across all transitions. In general, the method of linear instrumental variables applied to S does not estimate anything that is economically interpretable. It is not guaranteed to estimate a positive number even if the MTE is everywhere positive since the weights can be negative. In contrast, we can use our generalization of LIV presented in Equation (7.3) under conditions (OC-1)–(OC-6) to apply LIV to identify Δ^{MTE} for each transition, which can be used to build up Δ^{PRTE} using weights that can be estimated.

7.2.3. *Some theoretical examples of the weights in the ordered choice model*

Suppose that the distributions of W_s, $s = 1, \dots, \bar{S}$, are degenerate so that the C_s are constants satisfying $C_1 < \cdots < C_{\bar{S}-1}$. This is the classical ordered choice model. In this case, $\pi_s(Z, W_s) = F_V(\mu_D(Z) - C_s)$ for any $s = 1, \dots, \bar{S}$. For this special case, using J as an instrument will lead to nonnegative weights on all transitions if $J(Z, W)$ is a monotonic function of $\mu_D(Z)$. For example, note that $\mu_D(Z) - C_s > v$ can be written as $\mu_D(Z) > C_s + F_V^{-1}(u_D)$. Using $\mu_D(Z)$ as the instrument leads to weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$ of the form specified above with $\tilde{K}_s(u_D) = [E(\mu_D(Z) \mid \mu_D(Z) > F_V^{-1}(u_D) + C_s) - E(\mu_D(Z))] \Pr(\mu_D(Z) > F_V^{-1}(u_D) + C_s)$. Clearly, these weights will be nonnegative for all points of evaluation and will be strictly positive for any evaluation point u_D such that $1 > \Pr(\mu_D(Z) > F_V^{-1}(u_D) + C_s) > 0$.

Next consider the case where $C_s(W_s) = W_s$, a scalar, for $s = 1, \dots, \bar{S} - 1$, and where $\mu_D(Z) = 0$. Consider $J(Z, W) = W_s$, a purely transition-specific instrument. In this case, the weight on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$ is of the form given above, with

$$\tilde{K}_s(u_D) = \bigl[E\bigl(W_s \bigm| W_s > F_V^{-1}(u_D)\bigr) - E(W_s)\bigr] \Pr\bigl(W_s > F_V^{-1}(u_D)\bigr),$$

which will be nonnegative for all evaluation points and strictly positive for any evaluation point such that $1 > \Pr(W_s > F_V^{-1}(u_D)) > 0$.

What are the implied weights on $\Delta^{\text{MTE}}_{s',s'+1}(u_D)$ for $s' \neq s$? First, consider the case where W_s is independent of $W_{s'}$ for $s \neq s'$. This independence of W_s and $W_{s'}$ is not in conflict with the requirement $W_s > W_{s'}$ for $s > s'$ if the supports do not overlap for any $s' \neq s$. In this case, the weight on $\Delta^{\text{MTE}}_{s',s'+1}(u_D)$ for $s' \neq s$ is of the form given above with

$$\tilde{K}_{s'}(u_D) = \big[E\big(W_s \bigm| W_{s'} > F_V^{-1}(u_D)\big) - E(W_s)\big] \Pr\big(W_{s'} > F_V^{-1}(u_D)\big) = 0.$$

Thus, in this case, the instrument only weights the Δ^{MTE} for the s to $s+1$ transition. Note that this result relies critically on the assumption that W_s is independent of $W_{s'}$ for $s' \neq s$.

Consider another version of this example where $C_s(W_s) = W_s$, $s = 1, \dots, \bar{S}-1$, with W_s a scalar, but now allow $\mu_D(Z)$ to have a nondegenerate distribution and allow there to be dependence across the W_s. In particular, consider the case where $W = (W_1, \dots, W_{\bar{S}-1})$ is a continuous random vector with a density given by

$$\frac{\prod_{i=1}^{\bar{S}-1} f_i(w_i)\mathbf{1}[w_1 < w_2 < \cdots < w_{\bar{S}-1}]}{\int \cdots \int [\mathbf{1}[w_1 < w_2 < \cdots < w_{\bar{S}-1}] \prod_{i=1}^{\bar{S}-1} f_i(w_i)]\, dw_1 \cdots dw_{\bar{S}-1}}$$

for some marginal density functions $f_1(w_1), f_2(w_2), \dots, f_{\bar{S}-1}(w_{\bar{S}-1})$. In this case, using W_j as the instrument, we have

$$\begin{aligned}
\omega(s, v) = \Bigg(& \underset{-\infty < w_1 < \cdots < w_{\bar{S}-1} < \infty}{\int \cdots \int} \big(w_j - E(w_j)\big)\big(1 - F_{\mu_D(Z)}(w_s + v)\big) \\
& \times f_1(w_1) \cdots f_{\bar{S}-1}(w_{\bar{S}-1})\, dw_1 \cdots dw_{\bar{S}-1}\, f_V(v)\, dv \Bigg) \\
& \times \Bigg(\sum_{s=1}^{\bar{S}-1} \int \underset{-\infty < w_1 < \cdots < w_{\bar{S}-1} < \infty}{\int \cdots \int} \big(w_j - E(w_j)\big)\big(1 - F_{\mu_D(Z)}(w_s + v)\big) \\
& \times f_1(w_1) \cdots f_{\bar{S}-1}(w_{\bar{S}-1})\, dw_1 \cdots dw_{\bar{S}-1}\, f_V(v)\, dv \Bigg)^{-1}.
\end{aligned}$$

In the special case where $\mu_D(Z) \sim \text{Unif}(-K, K)$, with $Z \perp\!\!\!\perp W_s$ for $s = 1, \dots, \bar{S}-1$, assuming $-K < w_s + v < K$ for all w_s, v in the support of W_s and V, respectively, the numerator is

$$\begin{aligned}
& \underset{-\infty < w_1 < \cdots < w_{\bar{S}-1} < \infty}{\int \cdots \int} \big(w_j - E(w_j)\big) \\
& \times \frac{(w_s + v + K)}{2K} f_1(w_1) \cdots f_{\bar{S}-1}(w_{\bar{S}-1})\, dw_1 \cdots dw_{\bar{S}-1}\, f_V(v)\, dv
\end{aligned}$$

$$= \frac{1}{2K} \operatorname{Cov}(W_j, W_s \mid W_1 < \cdots < W_{\bar{S}-1}).$$

Observe that when the latent W_j, W_s are independently distributed for all j, s, by Bickel's Theorem (1967), we know that this expression is positive. (This is trivial when $j = s$.) The ordering $W_1 < \cdots < W_{\bar{S}-1}$ implies that W_l is stochastically increasing in W_j for $l < j$ (the lower boundary is shifted to the right). Hence, because of the order on the W implied by the ordered discrete choice model, a positive weighting is produced. This result can be overturned when $F(w)$ has a general structure. The positive dependence induced by the order on the components of W can be reversed by negative dependence in the structure of $F(w)$. We present examples of these phenomena in our discussions in Figures 19 and 20 below.

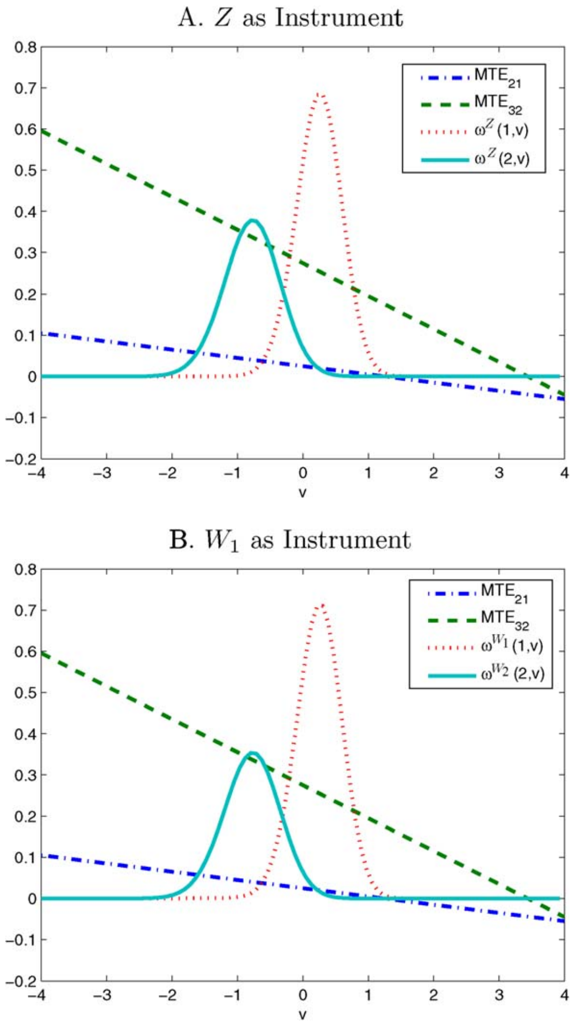
7.2.4. Some numerical examples of the IV weights

Figures 16–18 plot the transition-specific MTEs and the IV weights for the models and distributions of the weights at the base of each of the figures. We consider a three outcome ($\bar{S} = 3$) model with common instruments (Z) and transition-specific (W_s) instruments. The Z and W_s, $s = 1, \ldots, \bar{S}$, are assumed to be independent. The exact specification is given in the notes below Figure 16. In this example, D_s can be interpreted as an indicator of schooling. Y_1 is the potential earnings of the person as a dropout, Y_2 is the potential earnings of the person as a high school graduate, and Y_3 is the potential earnings of the person as a college graduate. There are two transitions: 1 → 2 and 2 → 3. The IV estimates using Z_1 and W_1 as instruments are reported transition by transition and overall decomposing IV representation (7.7) into its transition-specific components. The IV weights are defined by Equations (7.8) and (7.9). In particular, when the first element of Z, Z_1, is used as the instrument, we can decompose IV^{Z_1} as

$$\begin{aligned}
\mathrm{IV}^{Z_1} &= \sum_{s=1}^{2} \int E(Y_{s+1} - Y_s \mid V = v) \omega^{Z_1}(s, v) f_V(v)\, dv \\
&= \int \Delta_{12}^{\mathrm{MTE}}(v) \omega^{Z_1}(1, v) f_V(v)\, dv + \int \Delta_{23}^{\mathrm{MTE}}(v) \omega^{Z_1}(2, v) f_V(v)\, dv \\
&= \mathrm{IV}_{21}^{Z_1} + \mathrm{IV}_{32}^{Z_1}.
\end{aligned}$$

The same logic applies for the decomposition of IV^P which uses $P(Z)$ as an instrument. These decompositions show in this case that an important component of the total values of IV^Z and IV^{W_1} comes from the 2 → 3 transition. The bottom table presents the transition-specific treatment parameters. In Figure 16, the shape of the IV weights for Z_1 and W_1 are nearly identical. The IV estimates reflect this. The bottom table reveals that the IV estimates are far from standard treatment parameters.

In Figure 17, the IV weights for the Z_1 and W_1 are very different. So, correspondingly, are the IV estimates produced from each instrument, which are far off the mark of the standard treatment parameters shown in the bottom of the table. Observe that



Outcomes	Choice model
$Y_1 = \alpha + \beta_1 + U_1$	$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s]$,
$Y_2 = \alpha + \beta_2 + U_2$	$s = 1, 2, 3$
$Y_3 = \alpha + \beta_3 + U_3$	

Figure 16. Treatment parameters and IV – the generalized ordered choice Roy model under normality (Z, W_1). *Source*: Heckman, Urzua and Vytlacil (2004).

the IV weight for W_1 in the second transition is negative for an interval of values. This accounts for the dramatically lower IV estimate based on W_1 as the instrument. Figure 18 shows a different configuration of (Z_1, W_1, W_2). This produces negative weights

Parameterization

$$(U_1, U_2, U_3, V) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{UV}), (Z, W_1, W_2) \sim N(\boldsymbol{\mu}_{ZW}, \boldsymbol{\Sigma}_{ZW}) \text{ and } W_0 = -\infty; W_3 = \infty$$

$$\boldsymbol{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix}, \ \boldsymbol{\mu}_{ZW} = (-0.6, -1.08, 0.08)$$

$$\text{and } \boldsymbol{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0.09 \\ 0 & 0.09 & 0.25 \end{bmatrix}$$

$$\mathrm{Cov}(U_2 - U_1, V) = -0.02, \mathrm{Cov}(U_3 - U_2, V) = -0.08$$
$$\beta_1 = 0; \beta_2 = 0.025; \beta_3 = 0.3; \gamma = 1$$

IV estimates and their components*

Parameter	Value
Δ^{IV_Z}	0.1487
$\Delta_{12}^{\mathrm{IV}_Z}$	*0.0120*
$\Delta_{23}^{\mathrm{IV}_Z}$	*0.1367*
$\Delta^{\mathrm{IV}_{W_1}}$	0.1406
$\Delta_{12}^{\mathrm{IV}_{W_1}}$	*0.0126*
$\Delta_{23}^{\mathrm{IV}_{W_1}}$	*0.1280*

*IV^Z is decomposed as

$$\begin{aligned} \mathrm{IV}^Z =& \int E(Y_2 - Y_1 \mid V = v)\omega^Z(1, v) f_V(v)\, dv \\ &+ \int E(Y_3 - Y_2 \mid V = v)\omega^Z(2, v) f_V(v)\, dv \\ =& \mathrm{IV}_{21}^Z + \mathrm{IV}_{32}^Z. \end{aligned}$$

An analogous decomposition applies to IV^{W_1}.

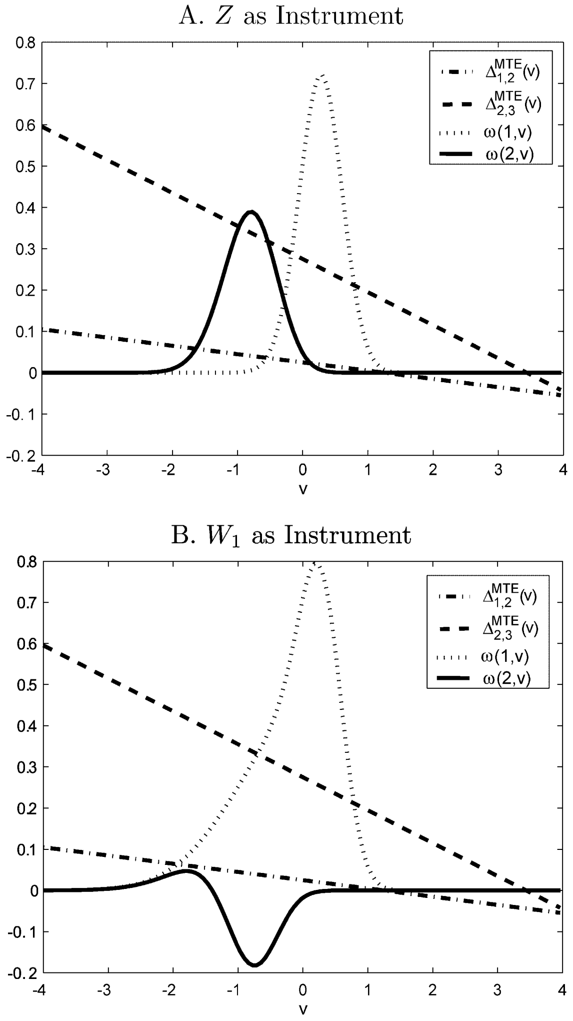
Treatment parameters and their values

Parameter	Value
$\mathrm{ATE}_{12} = E(Y_2 - Y_1)$	0.025
$\mathrm{ATE}_{23} = E(Y_3 - Y_2)$	0.275
$\mathrm{TT}_{12} = E(Y_2 - Y_1 \mid D_2 = 1)$	0.0282
$\mathrm{TT}_{23} = E(Y_3 - Y_2 \mid D_3 = 1)$	0.1908
$\mathrm{TUT}_{12} = E(Y_2 - Y_1 \mid D_1 = 1)$	0.0060
$\mathrm{TUT}_{23} = E(Y_3 - Y_2 \mid D_2 = 1)$	0.2956

Figure 16. (*Continued*)

for Z_1 for both transitions and a negative weight for W_1 in the second transition. For both instruments, IV is negative even though both MTEs are positive throughout most of their range. IV provides a misleading summary of the underlying marginal treatment effects.

Comparing Figures 16–18, it is important to recall that all are based on the same structural model. All have the same MTE and average treatment effects. But the IV estimates are very different solely as a consequence of the differences in the distributions



Outcomes	Choice model
$Y_1 = \alpha + \beta_1 + U_1$	$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s]$,
$Y_2 = \alpha + \beta_2 + U_2$	$s = 1, 2, 3$
$Y_3 = \alpha + \beta_3 + U_3$	

Figure 17. Treatment parameters and IV – the generalized ordered choice Roy model under normality (Z, W_1), Case I. *Source*: Heckman, Urzua and Vytlacil (2006).

of instruments across the examples. An alternative way to benchmark what IV estimates in the ordered choice model is to compare IV estimates to the PRTE for well-defined policy experiments. We consider two such experiments, corresponding to proportional

Parameterization

$$(U_1, U_2, U_3, V) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{UV}), (Z, W_1, W_2) \sim N(\boldsymbol{\mu}_{ZW}, \boldsymbol{\Sigma}_{ZW}) \text{ and } W_0 = -\infty; W_3 = \infty$$

$$\boldsymbol{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix}, \ \boldsymbol{\mu}_{ZW} = (-0.6, -1.08, 0.08)$$

$$\text{and } \boldsymbol{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & -0.09 \\ 0 & -0.09 & 0.25 \end{bmatrix}$$

$$\mathrm{Cov}(U_2 - U_1, V) = -0.02, \mathrm{Cov}(U_3 - U_2, V) = -0.08$$
$$\beta_1 = 0; \beta_2 = 0.025; \beta_3 = 0.3; \gamma = 1$$

IV estimates and their components*

Parameter	Value
Δ^{IV_Z}	0.1489
$\Delta_{12}^{\mathrm{IV}_Z}$	*0.0117*
$\Delta_{23}^{\mathrm{IV}_Z}$	*0.1372*
$\Delta^{\mathrm{IV}_{W_1}}$	0.0017
$\Delta_{12}^{\mathrm{IV}_{W_1}}$	*0.0325*
$\Delta_{23}^{\mathrm{IV}_{W_1}}$	*−0.0308*

*Δ^{IV_Z} is decomposed as

$$\begin{aligned}\Delta^{\mathrm{IV}_Z} &= \int E(Y_2 - Y_1 \mid V = v)\omega^Z(1, v) f_V(v)\,dv \\ &\quad + \int E(Y_3 - Y_2 \mid V = v)\omega^Z(2, v) f_V(v)\,dv \\ &= \Delta_{12}^{\mathrm{IV}_Z} + \Delta_{23}^{\mathrm{IV}_Z}.\end{aligned}$$

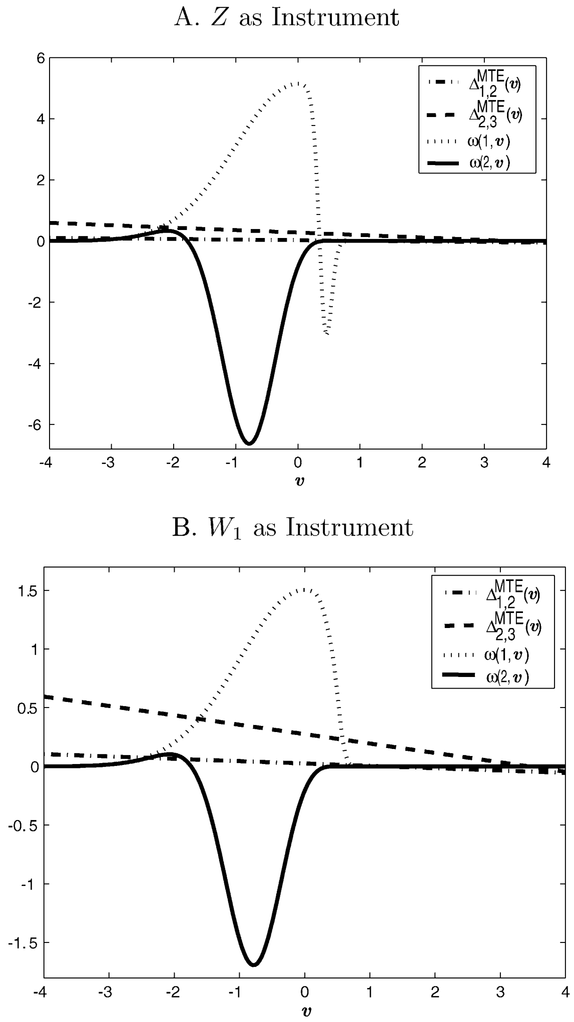
An analogous decomposition applies to $\Delta^{\mathrm{IV}_{W_1}}$.

Treatment parameters and their values

Parameter	Value
$\mathrm{ATE}_{12} = E(Y_2 - Y_1)$	0.025
$\mathrm{ATE}_{23} = E(Y_3 - Y_2)$	0.275
$\mathrm{TT}_{12} = E(Y_2 - Y_1 \mid D_2 = 1)$	0.0271
$\mathrm{TT}_{23} = E(Y_3 - Y_2 \mid D_3 = 1)$	0.1871
$\mathrm{TUT}_{12} = E(Y_2 - Y_1 \mid D_1 = 1)$	0.0047
$\mathrm{TUT}_{23} = E(Y_3 - Y_2 \mid D_2 = 1)$	0.2854

Figure 17. (*Continued*)

and fixed subsidies for attending different levels of schooling. We use the definition of the PRTE given in Equation (7.5). The baseline model is the one used to generate Figure 17. The weights can be constructed from data and are derived in Appendix H.


A. Z as Instrument

B. W_1 as Instrument

Outcomes

$$Y_1 = \alpha + \beta_1 + U_1$$
$$Y_2 = \alpha + \beta_2 + U_2$$
$$Y_3 = \alpha + \beta_3 + U_3$$

Choice model

$$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s],$$
$$s = 1, 2, 3$$

Figure 18. Treatment parameters and IV – the generalized ordered choice Roy model under normality (Z, W_1), Case II. *Source*: Heckman, Urzua and Vytlacil (2006).

Figure 19 plots the weights for the PRTE for each transition for a policy experiment. We change the economy from the benchmark economy that generates Figure 17 to an economy where W_2 is subsidized by a proportional amount τ. The PRTE weights for

Parameterization

$$(U_1, U_2, U_3, V) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{UV}), (Z, W_1, W_2) \sim N(\boldsymbol{\mu}_{ZW}, \boldsymbol{\Sigma}_{ZW}) \text{ and } W_0 = -\infty; W_3 = \infty$$

$$\boldsymbol{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix}, \quad \boldsymbol{\mu}_{ZW} = (-0.6, -1.08, 0.08)$$

$$\text{and } \boldsymbol{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0.092 & -0.036 \\ 0.092 & 0.1 & -0.09 \\ -0.036 & -0.09 & 0.25 \end{bmatrix}$$

$$\operatorname{Cov}(U_2 - U_1, V) = -0.02, \operatorname{Cov}(U_3 - U_2, V) = -0.08$$
$$\beta_1 = 0; \beta_2 = 0.025; \beta_3 = 0.3; \gamma = 1$$

IV estimates and their components*

Parameter	Value
Δ^{IV_Z}	−1.8091
$\Delta_{12}^{\mathrm{IV}_Z}$	*0.2866*
$\Delta_{23}^{\mathrm{IV}_Z}$	*−2.0957*
$\Delta^{\mathrm{IV}_{W_1}}$	−0.4284
$\Delta_{12}^{\mathrm{IV}_{W_1}}$	*0.0909*
$\Delta_{23}^{\mathrm{IV}_{W_1}}$	*−0.5193*

*See the footnote below Figure 16 for details of the decomposition of Δ^{IV_Z} and $\Delta^{\mathrm{IV}_{W_1}}$.

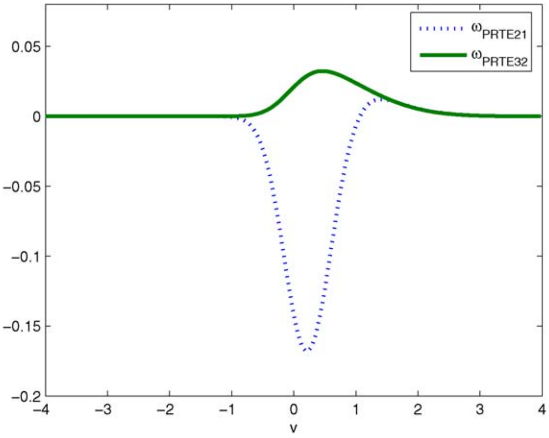
Treatment parameters and their values

Parameter	Value
$\mathrm{ATE}_{12} = E(Y_2 - Y_1)$	0.025
$\mathrm{ATE}_{23} = E(Y_3 - Y_2)$	0.275
$\mathrm{TT}_{12} = E(Y_2 - Y_1 \mid D_2 = 1)$	0.0283
$\mathrm{TT}_{23} = E(Y_3 - Y_2 \mid D_3 = 1)$	0.1754
$\mathrm{TUT}_{12} = E(Y_2 - Y_1 \mid D_1 = 1)$	0.0025
$\mathrm{TUT}_{23} = E(Y_3 - Y_2 \mid D_2 = 1)$	0.2898

Figure 18. (*Continued*)

each transition are negative over certain intervals. The overall PRTE is close to zero and can be decomposed into two components corresponding to a negative component on the second transition. The IV for the benchmark regime (p) and new regime (p') are given in the bottom table. The IV based on Z are far from the PRTE parameter. In general, the IV estimands are far off the mark from the PRTEs.

We next present a comparison between what IV estimates and the PRTE for a policy that consists of changing W_2 to $W_2 - t$ ($t = 1.2$ in the simulations). This can be thought of as a college tuition reduction policy. We compare the weights on PRTE with



Outcomes	Choice model
$Y_1 = \alpha + \beta_1 + U_1$	$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s],$
$Y_2 = \alpha + \beta_2 + U_2$	$s = 1, 2, 3$
$Y_3 = \alpha + \beta_3 + U_3$	

Parameterization

The benchmark model (regime p) is the same as the one presented below Figure 17.

Under the new regime (regime p') we define $W_1^{p'} = W_1^p(1-\tau)$ with $\tau = 0.5$. Thus, under regime p we have

$$\boldsymbol{\mu}_{ZW}^{p'} = (-0.6, -0.54, 0.08) \text{ and } \boldsymbol{\Sigma}_{ZW}^{p} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.025 & -0.045 \\ 0 & -0.045 & 0.25 \end{bmatrix}$$

The other parameters remain at the values set under the regime p

PRTE estimates and their components[1]

Parameter	Value
$\text{PRTE}^{p',p}$	0.0076
$\text{PRTE}_{21}^{p',p}$	*−0.0032*
$\text{PRTE}_{32}^{p',p}$	*0.0109*

[1] $\text{PRTE}^{p',p}$ is decomposed as

$$\begin{aligned}\text{PRTE}^{p',p} &= \int E(Y_2 - Y_1 \mid V = v)\omega^{p',p}(1, v) f_V(v)\, dv \\ &\quad + \int E(Y_3 - Y_2 \mid V = v)\omega^{p',p}(2, v) f_V(v)\, dv \\ &= \text{PRTE}_{21}^{p',p} + \text{PRTE}_{32}^{p',p}.\end{aligned}$$

Figure 19. The policy relevant treatment effect weights – the generalized ordered choice Roy model under normality. *Source*: Heckman, Urzua and Vytlacil (2004).

the weights on IV using W_1 (Figure 20) and Z (Figure 21) as instruments. The case using W_2 as an instrument is similar and for the sake of brevity is not discussed. In Figure 20A, we plot the transition-specific MTE for the values of the model presented

IV estimates and treatment parameters under different regimes[2]

Parameter	Regime p	Regime p'
IV^{Z}	0.1489	0.1521
IV_{12}^{Z}	*0.0117*	*0.0174*
IV_{23}^{Z}	*0.1372*	*0.1347*
IV^{W_1}	0.0017	0.0804
$\mathrm{IV}_{12}^{W_1}$	*0.0325*	*0.0358*
$\mathrm{IV}_{23}^{W_1}$	*−0.0308*	*0.0446*
ATE_{12}	0.0250	0.0250
ATE_{23}	0.2750	0.2750
TT_{12}	0.0271	0.0327
TT_{23}	0.1871	0.1789
TUT_{12}	0.0047	0.0103
TUT_{23}	0.2854	0.3067

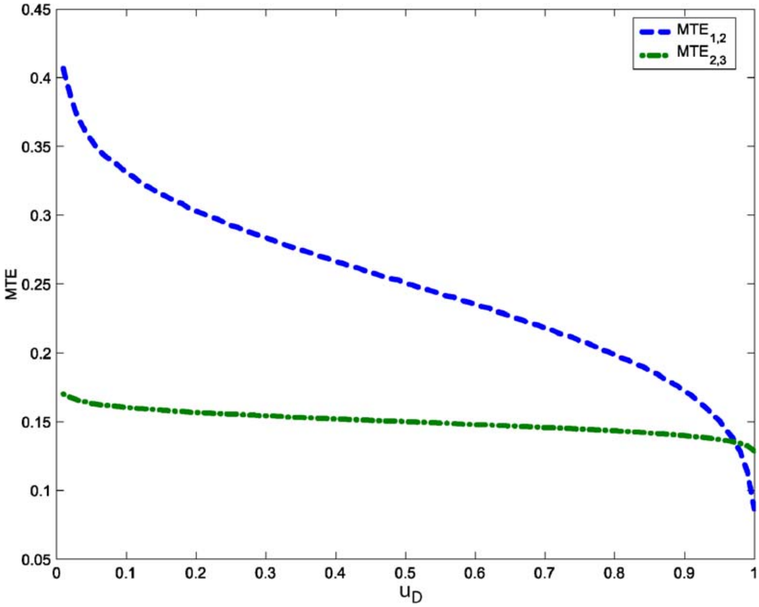
[2] See footnote below Figure 16 for details of the decompositions of IV^{Z} and IV^{W_1}.

Figure 19. (*Continued*)

at the base of the table. These are identical to the transition-specific MTE plotted in Figure 21A. Both of the Δ^{MTE} parameters have the typical shape of declining returns for people less likely to make the transition, i.e., those who have a higher $V = v$. Even though the levels are higher for outcomes 2 and 3, the marginal returns are higher for the transition 1 → 2. Figure 20B plots the policy weights for the two transitions for a policy that lowers W_2 ("reduces tuition").[108] It also plots the IV weights for the two Δ^{MTE} functions for the case where W_1 is the instrument. The correlation pattern for (W_1, W_2) is positive with specific values given below the figure. The policy studied in Figure 20B shifts 42.8% of the $D_1 = 1$ people into the category $D_3 = 1$ and 92.4% of D_2 people into D_3. In this simulation, the IV weights are positive. The IV weights and Δ^{PRTE} weights are distinctly different and the IV estimate is 0.201 vs. Δ^{PRTE} of 0.166.

When we change the correlation structure between W_1 and W_2 so that they are negatively correlated (Figure 20C), the IV weight for $\Delta_{2,3}^{\mathrm{MTE}}$ becomes *negative* while that for $\Delta_{1,2}^{\mathrm{MTE}}$ remains positive. The contrast in these figures between negative and positive IV weights depends on the correlation structure between W_1 and W_2. The stochastic order $(W_2 > W_1)$ is a force toward positive weights, which can be undone when the dependence induced by the density $(f(w_1, w_2))$ is sufficiently negative. The discord between the IV and Δ^{PRTE} weights is substantial and is reflected in the estimates ($\Delta^{\mathrm{PRTE}} = 0.159$ vs. $\Delta^{\mathrm{IV}} = 0.296$). As Figure 20D illustrates, the weights on Δ^{PRTE} are not guaranteed

[108] Notice that, for clarity, of exposition we change the notation for the weights in Figures 20 and 21 to distinguish IV from PRTE weights.



$$\begin{array}{llll}
Y_3 = \alpha + \beta_3 + U_3; D_3 = 1 \text{ if } W_2 < I < \infty; & U_3 = \sigma_3 \tau; & \sigma_3 = 0.02, \sigma_2 = 0.012, \sigma_1 = -0.05, \sigma_V = -1 \\
Y_2 = \alpha + \beta_2 + U_2; D_2 = 1 \text{ if } W_1 < I \leqslant W_2; & U_2 = \sigma_2 \tau; & \alpha = 0.67, \beta_2 = 0.25, \beta_3 = 0.4 \\
Y_1 = \alpha + U_1; \quad D_1 = 1 \text{ if } -\infty < I \leqslant W_1; & U_1 = \sigma_1 \tau; & Z \sim N(-0.0026, 0.27) \text{ and } Z \perp\!\!\!\perp V \\
I = Z - V & V = \sigma_V \tau; & \tau \sim N(0, 1) \\
 & U_D = F_V(V) &
\end{array}$$

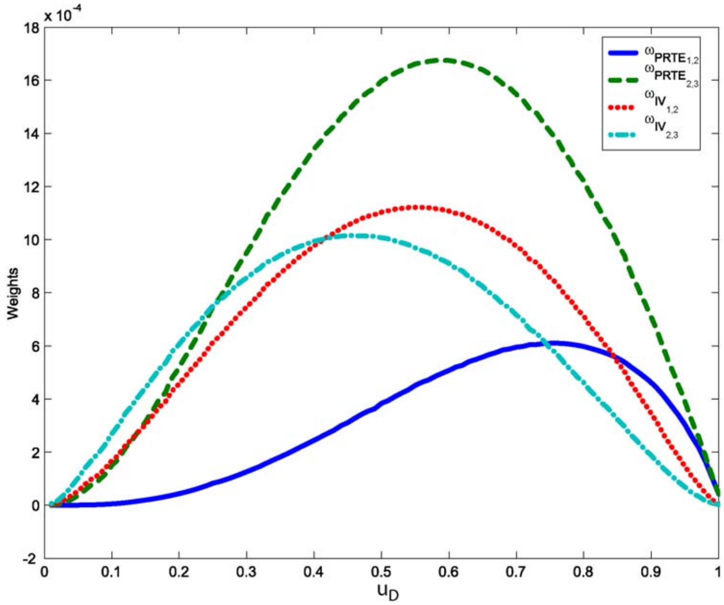
Sample size = 1500

Figure 20A. $W_2 - t$ where $t = 1.2$ and W_1 is the instrument: Marginal treatment effects by transition.

to be positive either. Thus neither the IV weights nor the weights on Δ^{PRTE} are guaranteed to be positive or negative and the relationship between the two sets of weights can be quite weak.

Figures 21A–21D present a parallel set of simulations when Z is used as an instrument. Changes in Z shift persons across all transitions whereas W_1 is a transition-specific shifter. Figure 21 reproduces the policy invariant Δ^{MTE} parameters from Figure 20A. Figure 21B shows that the IV weights for $\Delta^{\text{MTE}}_{1,2}$ assume both positive and negative values. The IV weights for $\Delta^{\text{MTE}}_{2,3}$ are positive but not monotonic. In Figure 21C, where there is negative dependence between W_1 and W_2, both sets of IV weights assume both positive and negative values. In the case where $f(w_1, w_2) = f_1(w_1) f_2(w_2)$, the weights on $\Delta^{\text{MTE}}_{1,2}$ for Δ^{PRTE} are negative.

These simulations show a rich variety of shapes and signs for the weights. They illustrate a main point of this chapter – that standard IV methods are not guaranteed to weight marginal treatment effects positively or to produce estimates close to policy rel-



$$(W_1, W_2) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$$

$$\Delta^{\text{PRTE}} = 0.166, \text{IV} = 0.201$$

Proportion induced to change from $D_1 = 1$ to $D_3 = 1 = 42.8\%$

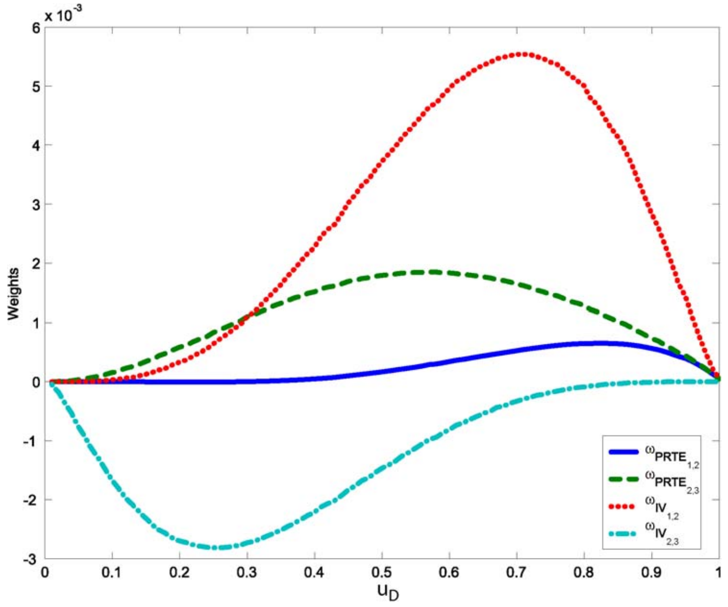
Proportion induced to change from $D_2 = 1$ to $D_3 = 1 = 92.4\%$

Figure 20B. $W_2 - t$ where $t = 1.2$ and W_1 is the instrument: Policy relevant treatment effect vs. instrumental variables weights by transition.

evant treatment effects or even to produce any gross treatment effect. Estimators based on LIV and its extension to the ordered model (7.3) identify Δ^{MTE} for each transition and answer policy relevant questions. We now turn to an analysis of a general unordered model.

7.3. Extension to multiple treatments that are unordered

The previous section analyzes a multiple treatment model where the treatment choice equation is an ordered choice model. In this section, we develop a framework for the analysis of multiple treatments when the choice equation is a nonparametric version of the classical multinomial choice model with no order imposed. Appendix B of Chapter 70, and Chapter 73 (Matzkin) analyze nonparametric and semiparametric identification of discrete choice models. With this framework, treatment effects can be defined as the difference in the counterfactual outcomes that would have been observed if the agent faced different general choice sets, i.e., the ef-



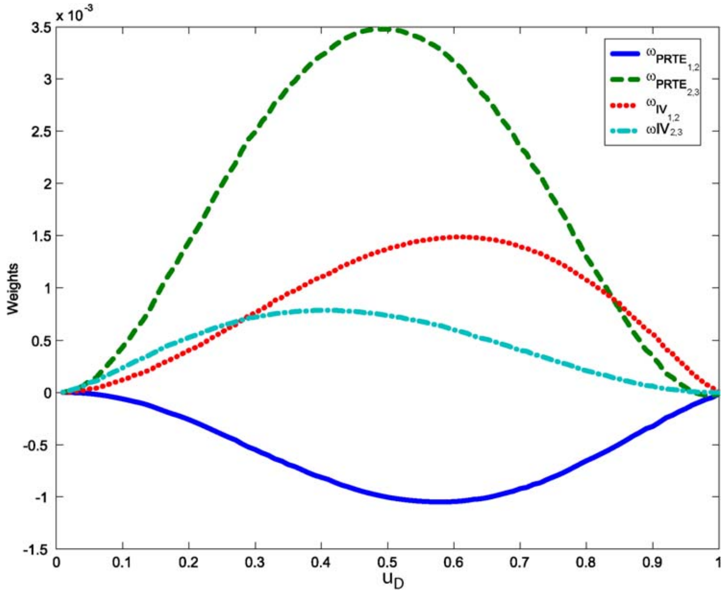
$$(W_1, W_2) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}\right)$$

$\Delta^{\text{PRTE}} = 0.159$, IV $= 0.296$
Proportion induced to change from $D_1 = 1$ to $D_3 = 1 = 32.1\%$
Proportion induced to change from $D_2 = 1$ to $D_3 = 1 = 64.7\%$

Figure 20C. $W_2 - t$ where $t = 1.2$ and W_1 is the instrument: Policy relevant treatment effect vs. instrumental variables weights by transition.

fect of the individual being forced to choose from one choice set instead of another. We define treatment parameters for a general multiple treatment problem and present conditions for the application of instrumental variables for identifying a variety of new treatment parameters. Our identification conditions are weaker than the ones used in Appendix B of Chapter 70, which establishes conditions under which it is possible to nonparametrically identify a full multinomial selection model.

Our use of choice theory is a unique aspect of our approach to the analysis of treatment effects. One particularly helpful result we draw on is the representation of the multinomial choices in terms of the choice between a particular choice and the best option among all other choices. This representation is crucial for understanding why LIV allows one to identify the MTE for the effect of one choice versus the best alternative option. The representation was introduced in Domencich and McFadden (1975), and has been used in the analysis of parametric multinomial selection models by Lee (1983)



$$(W_1, W_2) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$\Delta^{\text{PRTE}} = 0.110$, $\text{IV} = 0.210$

Proportion induced to change from $D_1 = 1$ to $D_3 = 1 = 27.5\%$

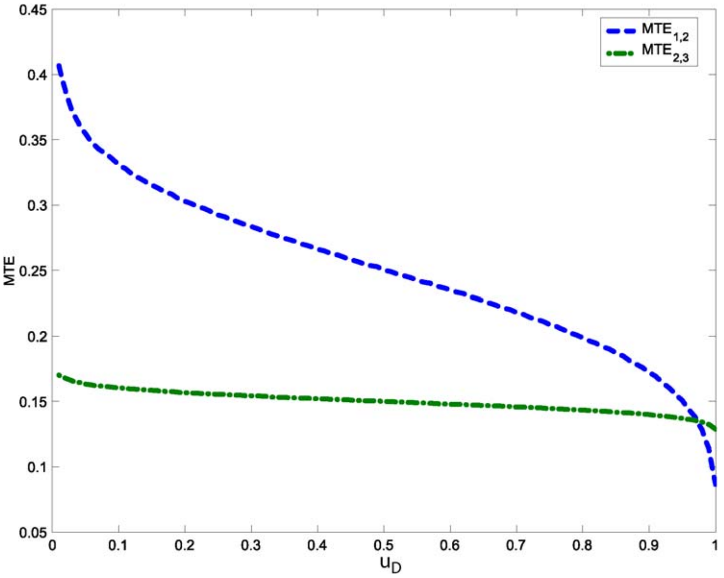
Proportion induced to change from $D_2 = 1$ to $D_3 = 1 = 76.8\%$

Figure 20D. $W_2 - t$ where $t = 1.2$ and W_1 is the instrument: Policy relevant treatment effect vs. instrumental variables weights by transition.

and Dahl (2002). Unlike those authors, we systematically explore treatment effect heterogeneity, consider nonparametric identification, and examine the application of the LIV methodology to such models.

Our analysis proceeds as follows. We first introduce our nonparametric, multinomial selection model and state our assumptions in Section 7.3.1. In Section 7.3.2, we define treatment effects in a general unordered model as the differences in the counterfactual outcomes that would have been observed if the agent faced different choice sets, i.e., the effects observed if individuals are forced to choose from one choice set instead of another. We also define the corresponding treatment parameters. Treatment effects in this context exhibit a form of treatment effect heterogeneity not present in the binary treatment case. The new form of heterogeneity arises from agents facing different choice sets, which we discuss in Section 7.3.3.

Section 7.3.4 establishes that LIV and the nonparametric Wald-IV estimand produce identification of the MTE/LATE versions of the effect of one choice versus the best alternative option without requiring knowledge of the latent index functions generat-

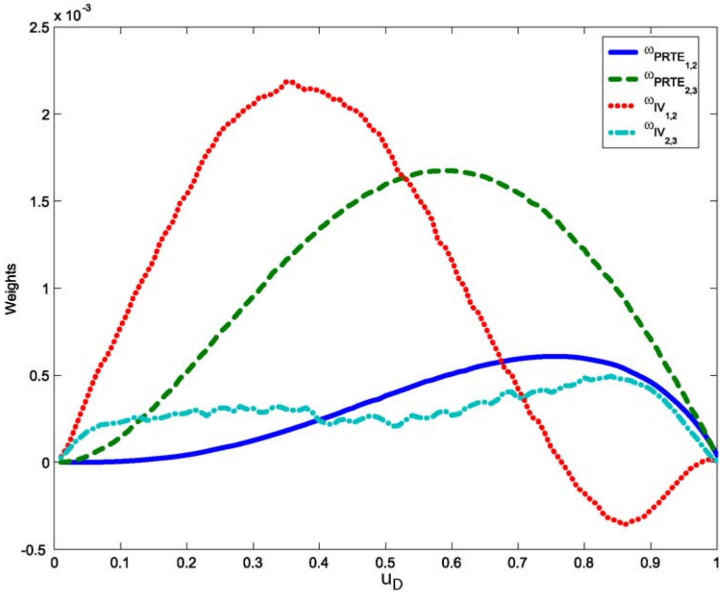


$$\begin{array}{llll}
Y_3 = \alpha + \beta_3 + U_3; & D_3 = 1 \text{ if } W_2 < I < \infty; & U_3 = \sigma_3 \tau; & \sigma_3 = 0.02, \sigma_2 = 0.012, \sigma_1 = -0.05, \sigma_V = -1 \\
Y_2 = \alpha + \beta_2 + U_2; & D_2 = 1 \text{ if } W_1 < I \leqslant W_2; & U_2 = \sigma_2 \tau; & \alpha = 0.67, \beta_2 = 0.25, \beta_3 = 0.4 \\
Y_1 = \alpha + U_1; & D_1 = 1 \text{ if } -\infty < I \leqslant W_1; & U_1 = \sigma_1 \tau; & Z \sim N(-0.0026, 0.27) \text{ and } Z \perp\!\!\!\perp V \\
I = Z - V & & V = \sigma_V \tau; & \tau \sim N(0, 1)
\end{array}$$

Sample size = 1500

Figure 21A. $W_2 - t$ where $t = 1.2$ and Z is the instrument: Marginal treatment effects by transition.

ing choices or large support assumptions. Mean treatment effects comparing one option versus the best alternative are the easiest treatment effects to study using instrumental variable methods because we effectively collapse a multiple outcome model to a series of two-outcome models, picking one outcome relative to the rest. In Section 7.3.5, we consider a more general case and state conditions for identifying the mean effect of the outcome associated with the best option in one choice set to the mean effect of the best option not in that choice set. We show that identification of the corresponding MTE/LATE parameters requires knowledge of the latent index functions of the multinomial choice model. Thus, to identify the parameters by using IV or LIV requires the formulation and estimation of an explicit choice model. In Section 7.3.6, we analyze the identification of treatment parameters corresponding to the mean effect of one specified choice versus another specified choice. Identification of marginal treatment parameters in this case requires the use of identification at infinity arguments relying on large support assumptions, but does not require knowledge of the latent index functions of the multinomial choice problem. This use of large support assumptions is closely related to



$$(W_1, W_2) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$$

$\Delta^{\text{PRTE}} = 0.166$, $\text{IV} = 0.247$

Proportion induced to change from $D_1 = 1$ to $D_3 = 1 = 42.8\%$

Proportion induced to change from $D_2 = 1$ to $D_3 = 1 = 92.4\%$

Figure 21B. $W_2 - t$ where $t = 1.2$ and Z is the instrument: Policy relevant treatment effect vs. instrumental variables weights by transition.

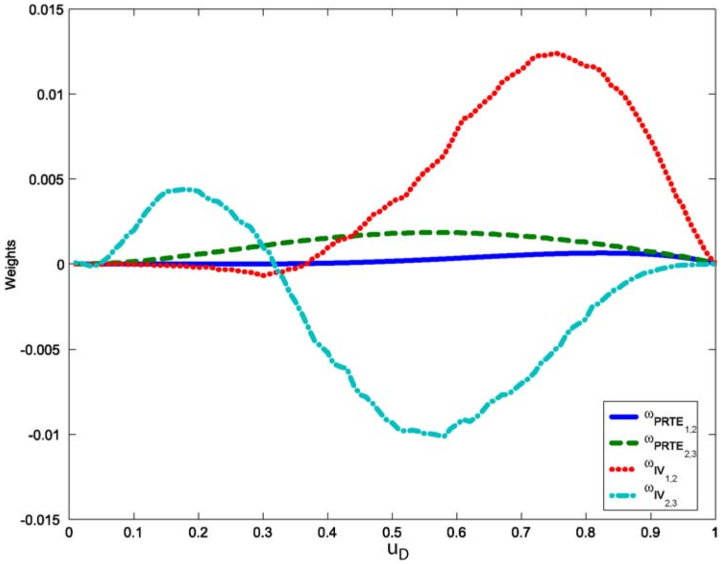
the need for large support assumptions to identify the full model developed in Appendix B of Chapter 70 of this Handbook. We summarize our analysis in Section 7.3.7.

7.3.1. Model and assumptions

Consider the following model with multiple choices and multiple outcome states for a general unordered model. Let $\mathcal{J}$ denote the agent's choice set, where $\mathcal{J}$ contains a finite number of elements. The value to the agent of choosing option $j \in \mathcal{J}$ is

$$R_j(Z_j) = \vartheta_j(Z_j) - V_j, \tag{7.10}$$

where Z_j are the agent's observed characteristics that affect the utility from choosing choice j, and V_j is the unobserved shock to the agent's utility from choice j. We will sometimes suppress the argument and write R_j for $R_j(Z_j)$. Let Z denote the random vector containing all unique elements of $\{Z_j\}_{j\in\mathcal{J}}$, i.e., $Z = \bigcup_{j\in\mathcal{J}}\{Z_j\}_{j\in\mathcal{J}}$. We will also sometimes write $R_j(Z)$ for $R_j(Z_j)$, leaving implicit that $R_j(\cdot)$ only depends on



$$(W_1, W_2) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}\right)$$

$\Delta^{\text{PRTE}} = 0.159$, $\text{IV} = 0.346$

Proportion induced to change from $D_1 = 1$ to $D_3 = 1 = 32.1\%$

Proportion induced to change from $D_2 = 1$ to $D_3 = 1 = 64.7\%$

Figure 21C. $W_2 - t$ where $t = 1.2$ and Z is the instrument: Policy relevant treatment effect vs. instrumental variables weights by transition.

those elements of Z that are contained in Z_j. Let $D_{\mathcal{J},j}$ be an indicator variable for whether the agent would choose option j if confronted with choice set $\mathcal{J}$[109]:

$$D_{\mathcal{J},j} = \begin{cases} 1 & \text{if } R_j \geqslant R_k, \ \forall k \in \mathcal{J}, \\ 0 & \text{otherwise.} \end{cases}$$

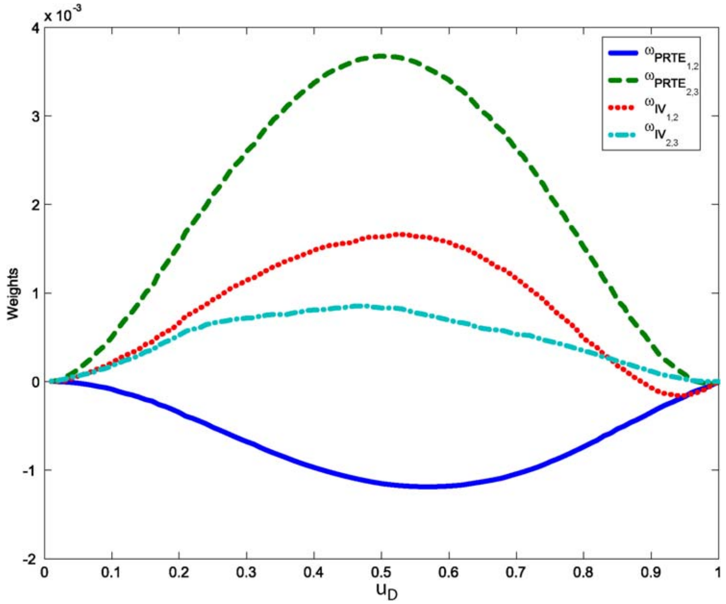
Let $I_{\mathcal{J}}$ denote the choice that would be made by the agent if confronted with choice set $\mathcal{J}$:

$$I_{\mathcal{J}} = j \quad \iff \quad D_{\mathcal{J},j} = 1.$$

Let $Y_{\mathcal{J}}$ be the outcome variable that would be observed if the agent faced choice set $\mathcal{J}$, determined by

$$Y_{\mathcal{J}} = \sum_{j \in \mathcal{J}} D_{\mathcal{J},j} Y_j,$$

[109] We will impose conditions such that ties, $R_j = R_k$ for $j \neq k$, occur with probability zero.



$$(W_1, W_2) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

$\Delta^{\text{PRTE}} = 0.104$, $\text{IV} = 0.215$

Proportion induced to change from $D_1 = 1$ to $D_3 = 1 = 27.3\%$

Proportion induced to change from $D_2 = 1$ to $D_3 = 1 = 69.3\%$

Figure 21D. $W_2 - t$ where $t = 1.2$ and Z is the instrument: Policy relevant treatment effect vs. instrumental variables weights by transition.

where Y_j is the potential outcome, observed only if option j is chosen. Y_j is determined by

$$Y_j = \mu_j(X_j, U_j),$$

where X_j is a vector of the agent's observed characteristics and U_j is an unobserved random vector. Let X denote the random vector containing all unique elements of $\{X_j\}_{j\in\mathcal{J}}$, i.e., $X = \bigcup_{j\in\mathcal{J}}\{X_j\}_{j\in\mathcal{J}}$. $(Z, X, I_{\mathcal{J}}, Y_{\mathcal{J}})$ is assumed to be observed. Define $R_{\mathcal{J}}$ as the maximum obtainable value given choice set $\mathcal{J}$:

$$R_{\mathcal{J}} = \max_{j\in\mathcal{J}}\{R_j\} = \sum_{j\in\mathcal{J}} D_{\mathcal{J},j} R_j.$$

We thus obtain the traditional representation of the decision process that choice j being optimal implies that choice j is better than the "next best" option:

$$I_{\mathcal{J}} = j \quad \iff \quad R_j \geqslant R_{\mathcal{J}\setminus j}.$$

More generally, a choice from $\mathcal{K}$ being optimal is equivalent to the highest value obtainable from choices in $\mathcal{K}$ being higher than the highest value that can be obtained from choices outside that set,

$$I_{\mathcal{J}} \in \mathcal{K} \quad \Longleftrightarrow \quad R_{\mathcal{K}} \geqslant R_{\mathcal{J}\setminus\mathcal{K}}.$$

As we will show, this simple representation is the key intuition for understanding how nonparametric instrumental variables estimate the effect of a given choice versus the "next best" alternative.

Analogous to our definition of $R_{\mathcal{J}}$, we define $R_{\mathcal{J}}(z)$ to be the maximum obtainable value given choice set $\mathcal{J}$ when instruments are fixed at $Z = z$,

$$R_{\mathcal{J}}(z) = \max_{j\in\mathcal{J}}\bigl\{R_j(z)\bigr\}.$$

Thus, for example, a choice from $\mathcal{K}$ is optimal when instruments are fixed at $Z = z$ if $R_{\mathcal{K}}(z) \geqslant R_{\mathcal{J}\setminus\mathcal{K}}(z)$.

We make the following assumptions, which generalize assumptions (A-1)–(A-5) invoked in Heckman and Vytlacil (2001b) and later used in Heckman and Vytlacil (2005), as developed in Section 2. We present the assumptions in a fashion parallel to (A-1)–(A-5) and (OC-1)–(OC-6). For that reason, we present the second assumption, which requires special attention, out of order.

(B-1) $\{(V_j, U_j)\}_{j\in\mathcal{J}}$ *is independent of* Z *conditional on* X.
(B-3) *The distribution of* $(\{V_j\}_{j\in\mathcal{J}})$ *is continuous.*[110]
(B-4) $E(|Y_j|) < \infty$ *for all* $j \in \mathcal{J}$.
(B-5) $\Pr(I_{\mathcal{J}} = j \mid X) > 0$ *for all* $j \in \mathcal{J}$.

Assumption (B-1) and (B-3) imply that $R_j \neq R_k$ w.p.1 for $j \neq k$, so that $\operatorname{argmax}\{R_j\}$ is unique w.p.1. Assumption (B-4) is required for the mean treatment parameters to be well defined. It allows us to integrate to the limit, which will be a crucial step for all identification analysis. Assumption (B-5) requires that at least some individuals participate in each program for all X.

Our definition and analysis of the treatment parameters only require assumptions (B-1) and (B-3)–(B-5). However, we will also impose an exclusion restriction for our identification analysis. Let $Z^{[j]}$ denote the jth components of Z that are in Z_j but not in Z_k, $k \neq j$. Let $Z^{[-j]}$ denote all elements of Z except for the components in $Z^{[j]}$. We work with two alternative assumptions for the exclusion restriction.[111] Consider

[110] Absolutely continuous with respect to Lebesgue measure on $\prod_{j\in\mathcal{J}}\mathbb{R}$.

[111] We work here with exclusion restrictions in part for ease of exposition. By adapting the analysis of Cameron and Heckman (1998) and Heckman and Navarro (2007), one can modify our analysis for the case of no exclusion restrictions if Z contains a sufficient number of continuous variables and there is sufficient variation in the ϑ_k function across k.

(B-2a) *for each $j \in \mathcal{J}$, there exists at least one element of Z, say $Z^{[j]}$, such that $Z^{[j]}$ is not an element of Z_k, $k \neq j$, and such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ is nondegenerate*,

or

(B-2b) *for each $j \in \mathcal{J}$, there exists at least one element of Z, say $Z^{[j]}$, such that $Z^{[j]}$ is not an element of Z_k, $k \neq j$, and such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ is continuous.*[112]

Assumption (B-2a) imposes the requirement that the analyst be able to independently vary the index for the given value function. This produces variation that affects only the value of the jth value function and causes people to enter or exit sector j. It imposes an exclusion restriction, that for any $j \in \mathcal{J}$, Z contains an element such that (i) it is contained in Z_j; (ii) it is not contained in any Z_k for $k \neq j$ and (iii) $\vartheta_j(\cdot)$ is a nontrivial function of that element conditional on all other regressors. Assumption (B-2b) strengthens (B-2a) by adding a smoothness assumption. A necessary condition for (B-2b) is for the excluded variable to have a density with respect to Lebesgue measure conditional on all other regressors and for $\vartheta_j(\cdot)$ to be a continuous and nontrivial function of the excluded variable.[113] Assumption (B-2a) will be used to identify a generalization of the LATE parameter. Assumption (B-2b) will be used to identify a generalization of the MTE parameter. For certain portions of the analysis, we strengthen (B-2b) to a large support condition, though the large support assumption will not be required for most of our analysis. Assumptions (B-2a) and (B-2b) mirror (A-2) for the binary choice model and are analogous to (OC-2) and (OC-6) in an ordered choice model.

7.3.2. Definition of treatment effects and treatment parameters

Treatment effects are defined as the difference in the counterfactual outcomes that would have been observed if the agent faced different choice sets. For any two choice sets, $\mathcal{K}, \mathcal{L} \subset \mathcal{J}$, define

$$\Delta_{\mathcal{K},\mathcal{L}} = Y_{\mathcal{K}} - Y_{\mathcal{L}}.$$

This is the effect of the individual being forced to choose from choice set $\mathcal{K}$ versus choice set $\mathcal{L}$. The conventional treatment effect is defined as the difference in potential outcomes between two specified states,

$$\Delta_{k,l} = Y_k - Y_l,$$

[112] Absolutely continuous with respect to Lebesgue measure.

[113] (B-2b) can be easily relaxed to the weaker assumption that the support of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ contains an open interval, or further weakened to the assumption that the conditional support contains at least one limit point. In these cases, the analysis of this section goes through without change for analysis for points within the open interval or more generally for any limit point.

which is nested within this framework by taking $\mathcal{K} = \{k\}$, $\mathcal{L} = \{l\}$. It is the effect for the individual of having no choice except to choose state l.

$\Delta_{\mathcal{K},\mathcal{L}}$ will be zero for agents who make the same choice when confronted with choice set $\mathcal{K}$ and choice set $\mathcal{L}$. Thus, $I_{\mathcal{K}} = I_{\mathcal{L}}$ implies $\Delta_{\mathcal{K},\mathcal{L}} = 0$, and we have

$$\Delta_{\mathcal{K},\mathcal{L}} = \mathbf{1}(I_{\mathcal{L}} \neq I_{\mathcal{K}})\Delta_{\mathcal{K}\setminus\mathcal{L},\mathcal{L}} \tag{7.11}$$

$$= \mathbf{1}(I_{\mathcal{L}} \neq I_{\mathcal{K}})\left(\sum_{j\in\mathcal{K}\setminus\mathcal{L}} D_{\mathcal{K},j}\Delta_{j,\mathcal{L}}\right). \tag{7.12}$$

Two examples will be of particular importance for our analysis. First, consider choice set $\mathcal{K} = \{k\}$ versus choice set $\mathcal{L} = \mathcal{J} \setminus \{k\}$. In this case, $\Delta_{k,\mathcal{J}\setminus k}$ is the difference between the agent's potential outcome in state k versus the outcome that would have been observed if he or she had not been allowed to choose state k. If $I_{\mathcal{J}} = k$, then $\Delta_{k,\mathcal{J}\setminus k}$ is the difference between the outcome in the agent's preferred state and the outcome in the agent's "next-best" state. Second, consider the set $\mathcal{K} = \mathcal{J}$ versus choice set $\mathcal{L} = \mathcal{J} \setminus \{k\}$. In this case, $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$ is the difference between the agent's best outcome and what his or her outcome would have been if state k had not been available. Note that

$$\Delta_{\mathcal{J},\mathcal{J}\setminus k} = D_{\mathcal{J},k}\Delta_{k,\mathcal{J}\setminus k}.$$

Thus, there is a trivial connection between the two parameters, $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$ and $\Delta_{k,\mathcal{J}\setminus k}$. We will focus on $\Delta_{k,\mathcal{J}\setminus k}$, the effect of being forced to choose option k versus being denied option k. However, one can use Equation (7.11) to use the results for $\Delta_{k,\mathcal{J}\setminus k}$ to obtain results for $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$.

To fix ideas regarding these alternative definitions of treatment effects, consider the following example concerning GED certification. The GED is an exam that certifies that high school dropouts who pass the test are the equivalents of high school graduates.

Example (*GED certification*). Consider studying the effect of GED certification on later wages. Consider the case where $\mathcal{J} = \{\{\text{GED}\}, \{\text{HS Degree}\}, \{\text{Permanent Dropout}\}\}$. Let $j = \{\text{GED}\}$, $k = \{\text{HS Degree}\}$, and $l = \{\text{Permanent Dropout}\}$. Suppose one wishes to study the effect of the GED. Then possible definitions of the effect of the GED include:

- $\Delta_{j,k}$ is the individual's outcome if he or she received the GED versus if he or she had graduated from high school;
- $\Delta_{j,l}$ is the individual's outcome if he or she received the GED versus if he or she had been a permanent dropout;
- $\Delta_{j,\mathcal{J}\setminus j}$ is the individual's outcome if he or she had received the GED versus what the outcome would have been if he or she had not had the option of receiving the GED;
- $\Delta_{\mathcal{J},\mathcal{J}\setminus j}$ is the individual's outcome if he or she had the option of receiving the GED versus the outcome if he or she did not have the option of receiving the GED. Notice that $\Delta_{\mathcal{J},\mathcal{J}\setminus j}$ is a version of an option value treatment effect.

We now define treatment parameters for a general unordered model.

Treatment parameters The conventional definition of the average treatment effect (ATE) is

$$\Delta_{k,l}^{\text{ATE}}(x,z) = E(\Delta_{k,l} \mid X = x, Z = z),$$

which immediately generalizes to the class of parameters discussed in this section as

$$\Delta_{\mathcal{K},\mathcal{L}}^{\text{ATE}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}} \mid X = x, Z = z).$$

Notice that the treatment parameters now depend on the value of Z. We explain the source of this dependence below. The conventional definition of the treatment on the treated (TT) parameter is

$$\Delta_{k,l}^{\text{TT}}(x,z) = E(\Delta_{k,l} \mid X = x, Z = z, I_{\mathcal{J}} = k),$$

which we generalize to

$$\Delta_{\mathcal{K},\mathcal{L}}^{\text{TT}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}} \mid X = x, Z = z, I_{\mathcal{J}} \in \mathcal{K}).$$

We also generalize the marginal treatment effect (MTE) and local average treatment effect (LATE) parameters considered in Heckman and Vytlacil (2001b). We generalize the MTE parameter to be the average effect conditional on being indifferent between the best option among choice set $\mathcal{K}$ versus the best option among choice set $\mathcal{L}$ at some fixed value of the instruments, $Z = z$:

$$\Delta_{\mathcal{K},\mathcal{L}}^{\text{MTE}}(x,z) = E\bigl(\Delta_{\mathcal{K},\mathcal{L}} \bigm| X = x, Z = z, R_{\mathcal{K}}(z) = R_{\mathcal{L}}(z)\bigr). \tag{7.13}$$

We generalize the LATE parameter to be the average effect for someone for whom the optimal choice in choice set $\mathcal{K}$ is preferred to the optimal choice in choice set $\mathcal{L}$ at $Z = \tilde{z}$, but who prefers the optimal choice in choice set $\mathcal{L}$ to the optimal choice in choice set $\mathcal{K}$ at $Z = z$:

$$\Delta_{\mathcal{K},\mathcal{L}}^{\text{LATE}}(x,z,\tilde{z}) = E\bigl(\Delta_{\mathcal{K},\mathcal{L}} \bigm| X = x, Z = z, R_{\mathcal{K}}(\tilde{z}) \geqslant R_{\mathcal{L}}(\tilde{z}), R_{\mathcal{L}}(z) \geqslant R_{\mathcal{K}}(z)\bigr). \tag{7.14}$$

An important special case of this parameter arises when $z = \tilde{z}$ except for elements that enter the index functions only for choices in $\mathcal{K}$ and not for any choice in $\mathcal{L}$. In that special case, Equation (7.14) simplifies to

$$\Delta_{\mathcal{K},\mathcal{L}}^{\text{LATE}}(x,z,\tilde{z}) = E\bigl(\Delta_{\mathcal{K},\mathcal{L}} \bigm| X = x, Z = z, R_{\mathcal{K}}(\tilde{z}) \geqslant R_{\mathcal{L}}(z) \geqslant R_{\mathcal{K}}(z)\bigr),$$

since $R_{\mathcal{L}}(z) = R_{\mathcal{L}}(\tilde{z})$ in this special case.

We have defined each of these parameters as conditional not only on X but also on the "instruments" Z. In general, the parameters depend on the Z evaluation point. For example, $\Delta_{\mathcal{K},\mathcal{L}}^{\text{ATE}}(x,z)$ generally depends on the z evaluation point. To see this, note that $Y_{\mathcal{K}} = \sum_{k \in \mathcal{K}} D_{\mathcal{K},k} Y_k$, and $Y_{\mathcal{L}} = \sum_{l \in \mathcal{L}} D_{\mathcal{L},l} Y_l$. By conditional independence assumption (B-1), $Z \perp\!\!\!\perp \{Y_j\}_{j \in \mathcal{J}} \mid X$, but $D_{\mathcal{K},k}$ and $D_{\mathcal{L},l}$ depend on Z conditional on X and thus $Y_{\mathcal{K}} - Y_{\mathcal{L}}$, in general, is dependent on Z conditional on X.[114] In other words, even though Z is conditionally independent of each individual potential outcome, it is correlated with the indicator for the choice that is optimal within the sets $\mathcal{K}$ and $\mathcal{L}$ and thus is related to $Y_{\mathcal{K}} - Y_{\mathcal{L}}$.

7.3.3. Heterogeneity in treatment effects

Consider heterogeneity in the pairwise treatment effect $\Delta_{j,k}$ (with $(j,k) \in \mathcal{J}$) defined as

$$\Delta_{j,k} = Y_j - Y_k = \mu_j(X_j, U_j) - \mu_k(X_k, U_k),$$

which in general will vary with both observables (X) and unobservables (U_j, U_k). Since we have not assumed that the error terms are additively separable, the treatment effect will in general vary with unobservables even if $U_j = U_k$.

The mean treatment parameters for $\Delta_{j,k}$ will differ if the effect of treatment is heterogeneous and agents base participation decisions, in part, on their idiosyncratic treatment effect. In general, the ATE, TT, and the marginal treatment parameters for $\Delta_{j,k}$ will differ as long as there is dependence between (U_j, U_k) and the decision rule, i.e., if there is dependence between (U_j, U_k) and $(\{V_l\}_{l\in\mathcal{J}})$. If we impose that $(\{V_l\}_{l\in\mathcal{J}})$ is independent of (U_j, U_k), then the treatment effect will still be heterogeneous, but the average treatment effect, average effect of treatment on the treated, and the marginal average treatment effects will all coincide.

The literature on treatment effects often imposes additive separability in outcomes between observables and unobservables. In particular, it is commonly assumed that U_j and U_k are scalar random variables and that $Y_j = \mu_j(X_j) + U_j$, $Y_k = \mu_k(X_k) + U_k$. In that case, a common treatment effect model is produced if the additive error term does not vary with the treatment state: $U_j = U_k$.[115] Thus, in the special case of additive separability, the treatment parameters for $\Delta_{j,k}$ will be the same even if there is dependence between $\{V_l\}_{l\in\mathcal{J}}$ and (U_j, U_k) as long as $U_j = U_k$.[116]

There is an additional source of treatment heterogeneity in the more general case of $\Delta_{\mathcal{K},\mathcal{L}}$ arising from heterogeneity in which states are being compared. Consider, for

[114] An exception is if $\mathcal{K} = \{k\}$, $\mathcal{L} = \{l\}$, i.e., both sets are singletons.

[115] More generally, if U_j, U_k are vector-valued, then additive separability is $Y_j = \mu_{1j}(X_j) + \mu_{2j}(U_j)$, $Y_k = \mu_{1k}(X_k)+\mu_{2k}(U_k)$, and the standard result is that a common treatment effect is produced if $\mu_{2j}(U_j) = \mu_{2k}(U_k)$.

[116] Because the literature often assumes additive separability in outcome equations, questions about the existence of a common treatment effect hinge on whether the additively separable error terms differ by treatment state. If the error terms differ by treatment state, there will be differences in the treatment parameters according to whether the differences in the error terms are stochastically dependent on the participation decision. Aakvik, Heckman and Vytlacil (1999) examine the case where the outcome variable is binary so that an additive separability assumption is not appropriate and Heckman and Vytlacil (2001b, 2005) consider cases without additive separability. Vytlacil, Santos and Shaikh (2005) and Vytlacil and Yildiz (2006) develop the case where $U_j = U_k$, but the model is not additively separable.

example, $\Delta_{j,\mathcal{J}\setminus j}$. We have that

$$\Delta_{j,\mathcal{J}\setminus j} = \sum_{k\in\mathcal{J}\setminus j} D_{\mathcal{J}\setminus j,k}\Delta_{j,k},$$

which will vary over individuals even if each individual has the same $\Delta_{j,k}$ treatment effect. Consider the corresponding ATE and TT parameters:

$$\begin{aligned}
&\Delta^{\text{ATE}}_{j,\mathcal{J}\setminus j}(x,z)\\
&\quad= E(\Delta_{j,\mathcal{J}\setminus j} \mid X=x, Z=z)\\
&\quad= \sum_{k\in\mathcal{J}\setminus j} \Pr(I_{\mathcal{J}\setminus j}=k \mid X=x, Z=z)E(\Delta_{j,k} \mid X=x, Z=z, I_{\mathcal{J}\setminus j}=k)
\end{aligned}$$

and

$$\begin{aligned}
&\Delta^{\text{TT}}_{j,\mathcal{J}\setminus j}(x,z)\\
&\quad= E(\Delta_{j,\mathcal{J}\setminus j} \mid X=x, Z=z, I_{\mathcal{J}}=j)\\
&\quad= \sum_{k\in\mathcal{J}\setminus j} \Pr(I_{\mathcal{J}\setminus j}=k \mid X=x, Z=z, I_{\mathcal{J}}=j)\\
&\qquad\qquad \times E(\Delta_{j,k} \mid X=x, Z=z, I_{\mathcal{J}}=j, I_{\mathcal{J}\setminus j}=k).
\end{aligned}$$

Even in the case where $\{U_j\}_{j\in\mathcal{J}}$ is independent of $\{V_j\}_{j\in\mathcal{J}}$, so that $E(\Delta_{j,k} \mid X=x, Z=z, I_{\mathcal{J}\setminus j}=k) = E(\Delta_{j,k} \mid X=x, Z=z, I_{\mathcal{J}}=j, I_{\mathcal{J}\setminus j}=k)$, it will still in general be the case that $\Delta^{\text{ATE}}_{j,\mathcal{J}\setminus j}(x,z) \neq \Delta^{\text{TT}}_{j,\mathcal{J}\setminus j}(x,z)$ since in general $\Pr(I_{\mathcal{J}\setminus j}=k \mid X=x, Z=z) \neq \Pr(I_{\mathcal{J}\setminus j}=k \mid X=x, Z=z, I_{\mathcal{J}}=j)$. Thus, the ATE and TT parameters will differ in part because they place different weights on the alternative pairwise treatment effects, and thus will differ even in the case where the pairwise (j versus k) treatment effects are common across all individuals.

In summary, $\Delta_{j,k}$ will be heterogeneous depending on the functional form of the $\mu_j(\cdot)$ and $\mu_k(\cdot)$ equations and on the pairwise dependence between the U_j and U_k terms. The $\Delta_{j,k}$ mean treatment parameters will also vary depending on the dependence between $\{V_l\}_{l\in\mathcal{J}}$ and (U_j, U_k). For $\Delta_{j,\mathcal{J}\setminus j}$, there is an additional source of heterogeneity arising from variability in the optimal option in the set $\mathcal{J}\setminus j$. Even if there is no heterogeneity in the pairwise $\Delta_{j,k}$ terms, there will still be heterogeneity in $\Delta_{j,\mathcal{J}\setminus j}$, and heterogeneity in the corresponding mean treatment parameters.

7.3.4. LIV and nonparametric Wald estimands for one choice vs. the best alternative

We first consider identification of treatment parameters corresponding to averages of $\Delta_{j,\mathcal{J}\setminus j}$, the effect of choosing option j versus the preferred option in $\mathcal{J}$ if j is not available. We analyze both a discrete change (Wald form for the instrumental variables estimand) and the local instrumental variables (LIV) estimand.[117] Using a concise notation, define $Z^{[j]}$ as the vector of elements in Z_j that do not enter any other choice index, and that $Z^{[-j]}$ is a vector of elements of Z not in $Z^{[j]}$. The $Z^{[j]}$ thus act as shifters attracting people into or out of state j but not affecting the valuations in the arguments of the other choice functions. For this case, we can develop an analysis of IV parallel to that given for the binary case or the ordered choice case if we condition on $Z^{[-j]}$. We obtain monotonicity or uniformity in this model if the movements among states induced by $Z^{[j]}$ are the same for all persons conditional on $Z^{[-j]} = z^{[-j]}$ and $X = x$. For example, *ceteris paribus* if $Z^{[j]} = z^{[j]}$ increases, $R_j(Z_j)$ increases but the $R_k(Z_k)$ are not affected, so the flow is toward state j.

Let $D_{\mathcal{J},j}$ be an indicator variable denoting whether option j is selected:

$$\begin{aligned}
D_{\mathcal{J},j} &= \mathbf{1}\Big(R_j(Z_j) \geqslant \max_{\ell \neq j}\big\{R_\ell(Z_\ell)\big\}\Big) \\
&= \mathbf{1}\Big(\vartheta_j(Z_j) \geqslant V_j + \max_{\ell \neq j}\big\{R_\ell(Z_\ell)\big\}\Big) \\
&= \mathbf{1}\big(\vartheta_j(Z_j) \geqslant \tilde{V}_j\big),
\end{aligned} \tag{7.15}$$

where $\tilde{V}_j = V_j + \max_{\ell \neq j}\{R_\ell(Z_\ell)\}$. Thus we obtain $D_{\mathcal{J},j} = \mathbf{1}(P_j(Z_j) \geqslant U_{D_j})$, where $U_{D_j} = F_{\tilde{V}_J | Z^{[-j]}}(V_j + \max_{\ell \neq j}\{R_\ell(Z_\ell)\} \mid Z^{[-j]} = z^{[-j]})$, where $F_{\tilde{V}_J | Z^{[-j]}}$ is the cdf of $\tilde{V}_j$ given $Z^{[-j]} = z^{[-j]}$. In a format parallel to the binary model, we write

$$Y = D_{\mathcal{J},j} Y_j + (1 - D_{\mathcal{J},j}) Y_{\mathcal{J} \setminus j}, \tag{7.16}$$

where $Y_{\mathcal{J} \setminus j}$ is the outcome that would be observed if option j were not available. This case is just a version of the binary case developed in previous sections of the paper. There is one crucial difference, however, and that is that the distributions of the $\tilde{V}_j$ now depend on the excluded $Z = z$. Thus instruments and parameters have to be defined conditionally on $Z = z$. We can define MTE as

$$E\big(Y_j - Y_{\mathcal{J} \setminus j} \bigm| X = x, Z = z, \vartheta_j(z_j) - V_j = R_{\mathcal{J} \setminus j}(z)\big).$$

We have to condition on $Z = z$ because the choice sets are defined over the max of elements in $\mathcal{J} \setminus j$ (see Equation (7.15)).

We now show that our identification strategies presented in the preceding part of this paper extend naturally to the identification of treatment parameters for $\Delta_{j,\mathcal{J} \setminus j}$. In particular, it is possible to recover LATE and MTE parameters for $\Delta_{j,\mathcal{J} \setminus j}$ by use of discrete change IV methods and local instrumental variable methods, respectively. Averages of the effect of option j versus the next best alternative are the easiest effects to study using instrumental variable methods and are natural generalizations of our two-outcome analysis.

[117] An estimand is the population version of the estimator.

The discrete change instrumental variables estimand will allow us to recover a version of the local average treatment effect (LATE) parameter.[118] Invoke assumption (B-2a). Assume only one excluded variable $Z^{[j]}$ in Z_j. If there are more, pick any one that satisfies (B-2a). Let $Z^{[-j]}$ denote the excluded variable for option j with properties assumed in (B-2a). We let $Z = [Z^{[-j]}, Z^{[j]}]$ and $\tilde{Z} = [\tilde{Z}^{[-j]}, \tilde{Z}^{[j]}]$ be two values where we only manipulate scalar $Z^{[j]}$.

$$\Delta_j^{\text{Wald}}\big(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}\big) = \frac{E(Y \mid X = x, Z = \tilde{z}) - E(Y \mid X = x, Z = z)}{\Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = \tilde{z}) - \Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = z)},$$

where for notational convenience we are assuming that $Z^{[j]}$ is the last element of Z. Note that all components of z and $\tilde{z}$ are the same except for the jth component. Without loss of generality, we assume that $\vartheta_j(\tilde{z}) > \vartheta_j(z)$.

If there were no X regressors, and if Z were a scalar, binary random variable, then $\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ would be the probability limit of the Wald form of two-stage least squares regression (2SLS). With X regressors, and with Z a vector possibly including continuous components, it no longer corresponds to a Wald/2SLS, but rather to a nonparametric version of the Wald estimator where the analyst nonparametrically conditions on X and on Z taking one of two specified values.

The local instrumental variables estimator (LIV) estimand introduced in Heckman (1997), and developed further in Heckman and Vytlacil (1999, 2000, 2005) and Florens et al. (2002), will allow us to recover a version of the marginal treatment effect (MTE) parameter. Impose (B-2b), and let $Z^{[j]}$ denote the excluded variable for option j with properties assumed in (B-2b). Because of the index structure, the LIV estimand will be invariant to which particular variable in $Z^{[j]}$ satisfying (B-2b) is used if there is more than one variable with the property assumed in (B-2b). The effects are *not* invariant to variables in $Z^{[-j]}$. Define

$$\Delta_j^{\text{LIV}}(x, z) \equiv \frac{\partial}{\partial z^{[j]}} E(Y \mid X = x, Z = z) \Big/ \frac{\partial}{\partial z^{[j]}} \Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = z).$$

$\Delta_j^{\text{LIV}}(x, z)$ is thus the limit form of $\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ as $\tilde{z}^{[j]}$ approaches $z^{[j]}$. Given our previous assumptions, one can easily show that this limit exists w.p.1. LIV corresponds to a nonparametric, local version of indirect least squares. It is a function of the distribution of the observable data, and it can be consistently estimated using any nonparametric estimator of the derivative of a conditional expectation.

Given these definitions, we have the following identification theorem.

[118] We are using the Z directly in the following manipulations instead of directly manipulating the $\{\vartheta_l(Z_l)\}_{l\in\mathcal{J}}$ indices. One can modify the following analysis to directly use $\{\vartheta_l(Z_l)\}_{l\in\mathcal{J}}$, with the disadvantage of requiring identification of $\{\vartheta_l(Z_l)\}_{l\in\mathcal{J}}$ (e.g., by an identification at infinity argument) but with the advantage of being able to follow the analysis of Heckman and Navarro (2007) in not requiring an exclusion restriction if Z contains a sufficient number of continuous variables and there is sufficient variation in the ϑ_k functions across k.

THEOREM 6.

1. *Assume* (B-1), (B-3)–(B-5), *and* (B-2a). *Then*
$$\Delta_j^{\text{Wald}}\big(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}\big) = \Delta_{j,\mathcal{J}\backslash j}^{\text{LATE}}(x, z, \tilde{z}),$$
where $\tilde{z} = (z^{[-j]}, \tilde{z}^{[j]})$.
2. *Assume* (B-1), (B-3)–(B-5), *and* (B-2b). *Then*
$$\Delta_j^{\text{LIV}}(x, z) = \Delta_{j,\mathcal{J}\backslash j}^{\text{MTE}}(x, z).$$

PROOF. See Appendix J. □

The intuition underlying the proof is simple. Under (B-1), (B-3)–(B-5), and (B-2a), we can convert the problem of comparing the outcome under j with the outcome under the next best option. This is an IV version of the selection modeling of Dahl (2002).

$\Delta_{j,\mathcal{J}\backslash j}^{\text{LATE}}(x, z, \tilde{z})$ is the average effect of switching to state j from state $I_{\mathcal{J}\backslash j}$ for individuals who would choose $I_{\mathcal{J}\backslash j}$ at $Z = z$ but would choose j at $Z = \tilde{z}$. $\Delta_{j,\mathcal{J}\backslash j}^{\text{MTE}}(x, z)$ is the average effect of switching to state j from state $I_{\mathcal{J}\backslash j}$ (the best option besides state j) for individuals who are indifferent between state j and $I_{\mathcal{J}\backslash j}$ at the given values of the selection indices (at $Z = z$, i.e., at $\{\vartheta_k(Z_k) = \vartheta_k(z_k)\}_{k\in\mathcal{J}}$).

The mean effect of state j versus state $I_{\mathcal{J}\backslash j}$ (the next best option) is a weighted average over $k \in \mathcal{J} \setminus j$ of the effect of state j versus state k, conditional on k being the next best option, weighted by the probability that k is the next best option. For example, for the LATE parameter,

$$\begin{aligned}
&\Delta_{j,\mathcal{J}\backslash j}^{\text{LATE}}(x, z, \tilde{z}) \\
&\quad= E\big(\Delta_{j,\mathcal{J}\backslash j} \bigm| X = x, Z = z, R_j(\tilde{z}) \geqslant R_{\mathcal{J}\backslash j}(z) \geqslant R_j(z)\big) \\
&\quad= \sum_{k\in\mathcal{J}\backslash j} \big[\Pr\big(I_{\mathcal{J}\backslash j} = k \bigm| Z \in \{z, \tilde{z}\}, X = x, R_j(\tilde{z}) \geqslant R_{\mathcal{J}\backslash j}(z) \geqslant R_j(z)\big) \\
&\qquad\times E\big(\Delta_{j,k} \bigm| X = x, Z \in \{z, \tilde{z}\}, R_j(\tilde{z}) \geqslant R_{\mathcal{J}\backslash j}(z) \geqslant R_j(z), I_{\mathcal{J}\backslash j} = k\big)\big],
\end{aligned}$$

where we use the result that $R_{\mathcal{J}\backslash j}(z) = R_{\mathcal{J}\backslash j}(\tilde{z})$ since $z = \tilde{z}$ except for one component that only enters the index for the jth option. The higher $\vartheta_k(z_k)$, holding the other indices constant, the larger the weight given to k as the base state. Thus, how heavily each option is weighted in this average depends on the switching probability $\Pr(I_{\mathcal{J}\backslash j} = k \mid Z = z, X = x, R_j(\tilde{z}_j) \geqslant R_k(z_k) \geqslant R_j(z_j))$, which in turn depends on $\{\vartheta_k(z_k)\}_{k\in\mathcal{J}\backslash j}$.

The LIV and Wald estimands depend on the z evaluation point. Alternatively, one can define averaged versions of the LIV and Wald estimands that will recover averaged versions of the MTE and LATE parameters,

$$\begin{aligned}
&\int \Delta_j^{\text{Wald}}\big(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}\big)\, dF_{Z^{[-j]}}\big(z^{[-j]}\big) \\
&\quad= \int \Delta_{j,\mathcal{J}\backslash j}^{\text{LATE}}(x, z, \tilde{z})\, dF_{Z^{[-j]}}\big(z^{[-j]}\big)
\end{aligned}$$

$$= E\big(\Delta_{j,\mathcal{J}\setminus j} \bigm| X = x, R_j\big(Z^{[-j]}, \tilde{z}^{[j]}\big) \geqslant R_{\mathcal{J}\setminus j}\big(Z^{[-j]}\big) \geqslant R_j\big(Z^{[-j]}, z^{[j]}\big)\big)$$

and

$$\int \Delta_j^{\mathrm{LIV}}(x,z)\,dF_Z(z) = \int \Delta_{j,\mathcal{J}\setminus j}^{\mathrm{MTE}}(x,z)\,dF_Z(z)$$
$$= E\big(\Delta_{j,\mathcal{J}\setminus j} \bigm| X = x, R_j(Z) = R_{\mathcal{J}\setminus j}(Z)\big).[119]$$

Thus far we have only considered identification of marginal treatment effect parameters, LATE and MTE, and not of the more standard treatment parameters like ATE and TT. However, following Heckman and Vytlacil (1999, 2001b), LATE can approximate ATE or TT arbitrarily well given the appropriate support conditions. Theorem 6 shows that we can use Wald estimands to identify LATE for $\Delta_{j,\mathcal{J}\setminus j}$, and we can thus adapt the analysis of Heckman and Vytlacil (2001b, 2005), as reviewed in Section 4, to identify ATE or TT for $\Delta_{j,\mathcal{J}\setminus j}$. Suppose that $Z^{[j]}$ denotes the excluded variable for option j with properties assumed in (B-2a), and suppose that: (i) the support of the distribution of $Z^{[j]}$ conditional on all other elements of Z is the full real line; (ii) $\vartheta_j(z_j) \to \infty$ as $z^{[j]} \to \infty$, and $\vartheta_j(z_j) \to -\infty$ as $z^{[j]} \to -\infty$. Then $\Delta_{j,\mathcal{J}\setminus j}^{\mathrm{ATE}}(x,z)$ and $\Delta_j^{\mathrm{LATE}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ are arbitrarily close when evaluated at a sufficiently large value of $\tilde{z}^{[j]}$ and a sufficiently small value of $z^{[j]}$. Following Heckman and Vytlacil (1999), $\Delta_{j,\mathcal{J}\setminus j}^{\mathrm{TT}}(x,z)$ and $\Delta_j^{\mathrm{LATE}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ are arbitrarily close for sufficiently small $z^{[j]}$. Using Theorem 6, we can use Wald estimands to identify the LATE parameters, and thus can use the Wald estimand to identify the ATE and TT parameters provided that there is sufficient support for the Z. While this discussion has used the Wald estimands, alternatively we could also follow Heckman and Vytlacil (1999), as summarized in Section 3, in expressing ATE and TT as integrated versions of MTE. By Theorem 6, we can use LIV to identify MTE and can thus express ATE and TT as integrated versions of the LIV estimand.

For a general instrument $J(Z^{[j]}, Z^{[-j]})$ constructed from $(Z^{[j]}, Z^{[-j]})$, which we denote as $J^{[j]}$, we can obtain a parallel construction to the characterization of standard IV given in Section 4.3:

$$\Delta_{J^{[j]}}^{\mathrm{IV}} = \int_0^1 \Delta^{\mathrm{MTE}}(x, z, u_{D_j})\,\omega_{\mathrm{IV}}^{J^{[j]}}(u_{D_j})\,du_{D_j}, \tag{7.17}$$

where

$$\omega_{\mathrm{IV}}^{J^{[j]}}(u_{D_j}) = \frac{E[J^{[j]} - E(J^{[j]}) \mid P_j(Z) \geqslant u_{D_j}]\Pr(P_j(Z) \geqslant u_{D_j} \mid Z^{[-j]} = z^{[-j]})}{\mathrm{Cov}(Z^{[j]}, D_{\mathcal{J},j})}, \tag{7.18}$$

[119] We assume that the support of $Z^{[-j]}$ conditional on $Z^{[j]}$ is the same as the support of $Z^{[-j]}$ conditional on $\tilde{Z}^{[j]}$.

where u_{D_j} is defined at the beginning of this subsection and where we keep the conditioning on $X = x$ implicit.

Note that from Theorem 6, we obtain that

$$\begin{aligned}
&\frac{\frac{\partial}{\partial z^{[j]}} E[Y \mid X = x, Z = z]}{\frac{\partial P_j(z)}{\partial z^{[j]}}} \\
&\quad = \frac{\partial E[Y \mid X = x, Z = z]}{\partial P_j(z)} \\
&\quad = E\big[Y_j - Y_{\mathcal{J}\setminus j} \bigm| X = x, Z = z, \vartheta_j(Z_j) - V_j = R_{\mathcal{J}\setminus j}(Z)\big]
\end{aligned}$$

so LIV identifies MTE and linear IV is a weighted average of LIV with the weights summing to one. These results mirror the results established in the binary case.

In the literature on the effects of schooling ($S = \sum_{j \in \mathcal{J}} j D_{\mathcal{J},j}$) on earnings ($Y_{\mathcal{J}}$), it is conventional to instrument S. The website of Heckman, Urzua and Vytlacil (2006) presents an analysis of this case. For the general unordered case,

$$\Delta^{\text{IV}}_{J^{[j]}} = \frac{\operatorname{Cov}(J^{[j]}, Y_{\mathcal{J}})}{\operatorname{Cov}(J^{[j]}, S)}$$

can be decomposed into economically interpretable components where the weights can be identified but the objects being weighted cannot be identified using local instrumental variables or LATE without making large support assumptions. However, the components can be identified using a structural model.

The trick we have used in this subsection comparing outcomes in j to the next best option converts a general unordered multiple outcome model into a two-outcome setup. This effectively partitions $Y_{\mathcal{J}}$ into two components, as in (7.16). Thus we write

$$Y_{\mathcal{J}} = D_{\mathcal{J},j} Y_j + (1 - D_{\mathcal{J},j}) Y_{\mathcal{J}\setminus j},$$

where

$$Y_{\mathcal{J}\setminus j} = \sum_{\substack{\ell \neq j \\ \ell \in \mathcal{J}}} \frac{D_{\mathcal{J},\ell}}{1 - D_{\mathcal{J},j}} Y_\ell \cdot \mathbf{1}(D_{\mathcal{J},j} \neq 1).$$

In the more general unordered case with three or more choices, to analyze IV estimates of the effect of S on $Y_{\mathcal{J}}$, we must work with $Y_{\mathcal{J}} = \sum_{k \in \mathcal{J}} D_{\mathcal{J},k} Y_k$ and make multiple comparisons across potential outcomes. This requires us to move outside of the LATE/LIV framework, which is inherently based on binary comparisons. We turn to that analysis next.

7.3.5. Identification: Effect of best option in $\mathcal{K}$ versus best option not in $\mathcal{K}$

We just presented an analysis of identification for treatment parameters defined as averages of $\Delta_{j,\mathcal{J}\setminus j}$, the effect of choosing option j versus the preferred option in $\mathcal{J}$ if j were not available. We now consider identification of $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$, the effect of choosing the preferred choice among set $\mathcal{K}$ versus the preferred choice among $\mathcal{J}$ if no option in $\mathcal{K}$ were available. This is an effect where we compare sets of options, and not just a single option compared to the rest.

We first start with an analysis that varies the $\{\vartheta_k(\cdot)\}_{k\in\mathcal{J}}$ indices directly. This analysis would be useful if one first identifies the index function, e.g., through an identification at infinity argument using the analysis in Matzkin (1993), as in Appendix B of Chapter 70 or Chapter 73 (Matzkin) in this Handbook. We then perform an analysis shifting Z directly. We show that it is possible to identify MTE and LATE averages of the $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$ effect if one has knowledge of the $\{\vartheta_k(\cdot)\}_{k\in\mathcal{J}}$ index functions but is not possible using shifts in Z without knowledge of the index functions. The one exception to this result is the special case already considered, when $\mathcal{K}=k$, i.e., the set only contains one element, in which case it is possible to identify the marginal parameters using shifts in Z directly without knowledge of the index functions.

Let $\vartheta_{\mathcal{J}}(Z)$ denote a random vector stacking the indices,

$$\vartheta_{\mathcal{J}}(Z)=\bigcup_{k\in\mathcal{J}}\bigl\{\vartheta_k(Z)\colon\ k\in\mathcal{J}\bigr\}.$$

Let $\vartheta_{\mathcal{J}}$ be a vector denoting a potential evaluation point of $\vartheta_{\mathcal{J}}(Z)$, $\vartheta_{\mathcal{J}}=\{\vartheta_k\colon\ k\in\mathcal{J}\}$, so that $\vartheta_{\mathcal{J}}(Z)=\vartheta_{\mathcal{J}}$ denotes the event $\{\vartheta_k(Z)=\vartheta_k\colon\ k\in\mathcal{J}\}$.[120] Let $\vartheta_{\mathcal{J}}+h$ denote $\{\vartheta_k+h\colon\ k\in\mathcal{J}\}$, where $h\in\mathbb{R}$. We now define a version of the Wald estimand that uses the indices directly as instruments instead of using Z as instruments,

$$\begin{aligned}
&\tilde{\Delta}^{\text{Wald}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}},h)\\
&\quad\equiv\bigl[E\bigl(Y\bigm| X=x,\vartheta_{\mathcal{K}}(Z)=\vartheta_{\mathcal{K}}+h,\vartheta_{\mathcal{J}\setminus\mathcal{K}}(Z)=\vartheta_{\mathcal{J}\setminus\mathcal{K}}\bigr)\\
&\qquad-E\bigl(Y\bigm| X=x,\vartheta_{\mathcal{J}}(Z)=\vartheta_{\mathcal{J}}\bigr)\bigr]\\
&\qquad\times\bigl[\Pr\bigl(I_{\mathcal{J}}\in\mathcal{K}\bigm| X=x,\vartheta_{\mathcal{K}}(Z)=\vartheta_{\mathcal{K}}+h,\vartheta_{\mathcal{J}\setminus\mathcal{K}}(Z)=\vartheta_{\mathcal{J}\setminus\mathcal{K}}\bigr)\\
&\qquad-\Pr\bigl(I_{\mathcal{J}}\in\mathcal{K}\bigm| X=x,\vartheta_{\mathcal{J}}(Z)=\vartheta_{\mathcal{J}}\bigr)\bigr]^{-1}.
\end{aligned}$$

$\tilde{\Delta}^{\text{Wald}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}},h)$ corresponds to the effect of a shift in each index in $\mathcal{K}$ upward by h while holding each index in $\mathcal{J}\setminus\mathcal{K}$ constant. Using indices, we define a version of the LIV estimand using indices $\tilde{\Delta}^{\text{LIV}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}})$ through a limit expression

$$\tilde{\Delta}^{\text{LIV}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}})=\lim_{h\to 0}\tilde{\Delta}^{\text{Wald}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}},h).$$

Likewise, we define versions of the LATE and MTE parameters that are functions of the ϑ indices instead of functions of z evaluation points,

$$\begin{aligned}
&\tilde{\Delta}^{\text{LATE}}_{\mathcal{K},\mathcal{L}}(x,\vartheta_{\mathcal{J}},h)\\
&\quad=E\bigl(\Delta_{\mathcal{K},\mathcal{L}}\bigm| X=x,\vartheta_{\mathcal{J}}(Z)=\vartheta_{\mathcal{J}},R_{\mathcal{K}}(Z)+h\geqslant R_{\mathcal{L}}(Z)\geqslant R_{\mathcal{K}}(Z)\bigr),
\end{aligned}$$

[120] Note that in our notation, $R_{\mathcal{J}}=\max\{R_k\}_{k\in\mathcal{J}}$ is a scalar, while $\vartheta_{\mathcal{J}}(Z)=\{\vartheta_k(Z)\colon\ k\in\mathcal{J}\}$ is a vector.

$$\tilde{\Delta}^{\mathrm{MTE}}_{\mathcal{K},\mathcal{L}}(x,\vartheta_{\mathcal{J}}) = E\big(\Delta_{\mathcal{K},\mathcal{L}} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_{\mathcal{K}}(Z) = R_{\mathcal{L}}(Z)\big).$$

We state the following identification theorem:

THEOREM 7.

1. *Assume* (B-1), (B-3)–(B-5), *and* (B-2a). *Then*

$$\tilde{\Delta}^{\mathrm{Wald}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}},h) = \tilde{\Delta}^{\mathrm{LATE}}_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}(x,\vartheta_{\mathcal{J}},h).$$

2. *Assume* (B-1), (B-3)–(B-5), *and* (B-2b). *Then*

$$\tilde{\Delta}^{\mathrm{LIV}}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}}) = \tilde{\Delta}^{\mathrm{MTE}}_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}(x,\vartheta_{\mathcal{J}}).$$

PROOF. Follows with trivial modifications from the proof of Theorem 6. □

Now consider the same analysis shifting Z directly instead of shifting the indices. First consider LATE. If one knew what shifts in Z corresponded to shifting each index in $\mathcal{K}$ upward by the same amount while holding each index in $\mathcal{J} \setminus \mathcal{K}$ constant, then one could immediately follow the preceding analysis to recover $E(\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_{\mathcal{K}}(Z) + h \geqslant R_{\mathcal{J}\setminus\mathcal{K}}(Z) \geqslant R_{\mathcal{K}}(Z))$. However, unless $\mathcal{K}$ is a singleton, without knowledge of the index functions one does not know what shifts in Z will have this property. One possible approach would be to only shift elements of Z that are elements of Z_j for $j \in \mathcal{K}$ but are excluded from Z_j for $j \in \mathcal{J} \setminus \mathcal{K}$. However, unless the shifts move the indices for choices in $\mathcal{K}$ all by the same amount, the shift in Z will result in movement not only from the set $\mathcal{J} \setminus \mathcal{K}$ to the set $\mathcal{K}$ but also cause movement between choices within $\mathcal{K}$. Thus, one can use shifts in Z to recover a LATE-type parameter for $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$ only if either (i) the index functions are known, or (ii) $\mathcal{K} = \{k\}$, i.e., the set $\mathcal{K}$ contains only one element. Our analysis establishes a fundamental role for choice theory in recovering the indices needed to perform IV analysis.

Thus far, we have only considered identification of marginal treatment effect parameters for $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$ and not of the more standard treatment parameters ATE and TT for $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$. As in the immediately preceding section, we can follow Heckman and Vytlacil (1999) in expressing ATE and TT as integrated versions of MTE or show that ATE and TT can be approximated arbitrarily well by LATE parameters. Given appropriate support conditions, we can again identify MTE over the appropriate range or identify the appropriate LATE parameters and thus identify ATE and TT given the required support conditions.

7.3.6. Identification: Effect of one fixed choice versus another

Consider evaluating the effect of fixed option j versus fixed option k, $\Delta_{j,k}$, i.e., the effect for the individual of having no choice except to choose state j versus no choice except to choose state k. We show that it is possible to identify averages of $\Delta_{j,k}$ if one has sufficient support conditions. These conditions supplement the standard IV conditions developed for the binary case [Heckman, Urzua and Vytlacil (2006)] with the conditions more commonly used in semiparametric estimation. We start by considering the analysis if one knows the ϑ index functions, say from a semiparametric analysis of discrete choice, and then show that knowledge of the ϑ index functions is not necessary.

For notational purposes, for any $j, k \in \mathcal{J}$, define $U_{j,k} = U_j - U_k$, and let $\vartheta_{j,k}(Z) = \vartheta_j(Z_j) - \vartheta_k(Z_k)$. One might try to follow our previous strategy to identify treatment parameters for $\Delta_{j,k}$ if one could shift $\vartheta_j - \vartheta_k = \vartheta_{j,k}$ while holding constant $\{\vartheta_{l,m}\}_{(l,m)\in\mathcal{J}\times\mathcal{J}\setminus\{j,k\}}$, i.e., while holding all other utility contrasts fixed.[121] However, given the structure of the latent variable model determining choices, these are incompatible conditions. To see this, note that $\vartheta_{j,k} = \vartheta_{l,k} - \vartheta_{l,j}$ for any l, and thus $\vartheta_{j,k}$ cannot be shifted while holding $\vartheta_{l,j}$ and $\vartheta_{l,k}$ constant.[122]

To bypass this problem, we develop a limit strategy to make the consequences of shifting $\vartheta_{j,k}$ negligible. Our strategy relies on an identification at infinity argument. For example, consider the case where $\mathcal{J} = \{1, 2, 3\}$, and consider identification of the MTE parameter for option 3 versus option 1. Recall that $D_{\mathcal{J}\setminus 3,l}$ is an indicator variable for whether option l would be chosen if option 3 were not available, so that $D_{\mathcal{J}\setminus 3,l}\Delta_{3,\mathcal{J}\setminus 3} = D_{\mathcal{J}\setminus 3,l}\Delta_{3,l}$. Since 1 and 2 are the only options if 3 is not available, it follows that $\Delta_{3,\mathcal{J}\setminus 3} = D_{\mathcal{J}\setminus 3,1}\Delta_{3,1} + D_{\mathcal{J}\setminus 3,2}\Delta_{3,2}$, and we have that

$$\begin{aligned}
&E\bigl(\Delta_{3,\mathcal{J}\setminus 3} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\setminus 3}(Z)\bigr) \\
&\quad = E\bigl(D_{\mathcal{J}\setminus 3,1}\Delta_{3,1} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\setminus 3}(Z)\bigr) \\
&\qquad + E\bigl(D_{\mathcal{J}\setminus 3,2}\Delta_{3,2} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\setminus 3}(Z)\bigr).
\end{aligned}$$

The smaller ϑ_2 is (holding ϑ_1 and ϑ_3 fixed), the larger the probability that the "next best option" is 1 and not 2. Note that $E(\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_1(Z))$ does not depend on the ϑ_2 evaluation point given independence assumption (B-1), so that

$$\begin{aligned}
&E\bigl(\Delta_{3,1} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_1(Z)\bigr) \\
&\quad = E\bigl(\Delta_{3,1} \bigm| X = x, \vartheta_{\mathcal{J}\setminus 2}(Z) = \vartheta_{\mathcal{J}\setminus 2}, R_3(Z) = R_1(Z)\bigr).
\end{aligned}$$

[121] Alternatively, one can allow $\vartheta_{l,m}(z) \neq \vartheta_{l,m}(z')$ if $\Pr(U_{l,m} \in [\vartheta_{l,m}(z), \vartheta_{l,m}(z')]) = 0$. Such a possibility would be ruled out except "at the limit" by the standard assumption that the support of $U_{l,m}$ is connected. (We discuss this below.) Even without such an assumption, such a possibility occurring simultaneously for all $(l, m) \in \mathcal{J} \times \mathcal{J} \setminus \{j, k\}$ for a particular (z, z') seems extremely implausible, and we will therefore not consider this possibility further.

[122] This suggests a nonparametric test of the latent variable model. If there exists (z, z') such that $\Pr(I_{\mathcal{J}} = j \mid Z = z) \neq \Pr(I_{\mathcal{J}} = j \mid Z = z')$, and $\Pr(I_{\mathcal{J}} = k \mid Z = z) \neq \Pr(I_{\mathcal{J}} = k \mid Z = z')$, but $\Pr(I_{\mathcal{J}} = l \mid Z = z) = \Pr(I_{\mathcal{J}} = l \mid Z = z')$ for all $l \in \mathcal{J} \setminus \{j, k\}$, then the latent variable model is rejected. However, shifts in only two indices are possible for sequential models since unexpected innovations in agent information sets will act to shift the current decision without affecting previous decisions. Consider the following sequential model of GED certification. In the first period, the agent chooses to graduate from high school or to dropout of high school. If the agent drops out of high school in the first period, he or she has the option in the second period of attaining GED certification or staying a permanent dropout. An unexpected shock in the second period to the relative value of GED certification versus permanent dropout status will shift the GED/permanent dropout choice without changing the probability of high school graduation.

Thus, by assumptions (B-1) and (B-3) and the Dominated Convergence Theorem, we have that

$$\begin{aligned}&\lim_{\vartheta_2 \to -\infty} E\bigl(D_{\mathcal{J}\backslash 3,1}\Delta_{3,1} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\bigr)\\ &\quad = E\bigl(\Delta_{3,1} \bigm| X = x, \vartheta_{\mathcal{J}\backslash 2}(Z) = \vartheta_{\mathcal{J}\backslash 2}, R_3(Z) = R_1(Z)\bigr)\end{aligned}$$

while

$$\lim_{\vartheta_2 \to -\infty} E\bigl(D_{\mathcal{J}\backslash 3,2}\Delta_{3,2} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\bigr) = 0,$$

so that

$$\begin{aligned}&\lim_{\vartheta_2 \to -\infty} E\bigl(\Delta_{3,\mathcal{J}\backslash 3} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\bigr)\\ &\quad = E\bigl(\Delta_{3,1} \bigm| X = x, \vartheta_{\mathcal{J}\backslash 2}(Z) = \vartheta_{\mathcal{J}\backslash 2}, R_3(Z) = R_1(Z)\bigr).\end{aligned}$$

In other words, as the value of option 2 becomes arbitrarily small, the probability of the "next best option" being 1 becomes arbitrarily close to one. Thus the MTE parameter for option 3 versus the next best option becomes arbitrarily close to the MTE parameter for option 3 versus option 1.

We can identify the MTE parameter for option 3 versus the next best option using the LIV estimand as in Theorem 6, and thus conditioning on ϑ_2 arbitrarily small we have that the LIV estimand is arbitrarily close to the MTE parameter for option 3 versus option 1. This analysis requires the appropriate support conditions in order for the limit operations to be well defined. The following theorem formalizes this idea, and is for the more general case where $\mathcal{J}$ is a general finite set.

THEOREM 8. *Assume* (B-1)*,* (B-3)–(B-5)*, and* (B-2b)*. Assume that, for any* $t \in \mathbb{R}$*,*

$$\Pr\bigl(\vartheta_l(Z_l) \leqslant t \bigm| \vartheta_j(Z_j), \vartheta_k(Z_k)\bigr) \geqslant 0 \quad \forall l \in \mathcal{J} \setminus \{j, k\}.$$

Then

$$\begin{aligned}&\lim_{\max_{l \in \mathcal{J}\backslash\{j,k\}}\{\vartheta_l\} \to -\infty} \tilde{\Delta}_j^{\mathrm{LIV}}(x, \vartheta_{\mathcal{J}})\\ &\quad = E\bigl(\Delta_{j,k} \bigm| X = x, \vartheta_{j,k}(Z) = \vartheta_{j,k}, R_j(Z) = R_k(Z)\bigr)\end{aligned}$$

for any

$$x \in \lim_{t \to -\infty} \mathrm{Supp}\Bigl(X \Bigm| \vartheta_j(Z_j) = \vartheta_j, \vartheta_k(Z_k) = \vartheta_k, \max_{l \in \mathcal{J}\backslash\{j,k\}}\bigl\{\vartheta_l(Z)\bigr\} \leqslant t\Bigr).$$

PROOF. By a trivial modification to the proof of Theorem 6, we have that

$$\tilde{\Delta}_j^{\mathrm{LIV}}(x, \vartheta_{\mathcal{J}}) = E\bigl(\Delta_{j,\mathcal{J}\backslash j} \bigm| X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_j(Z) = R_{\mathcal{J}\backslash j}(Z)\bigr).$$

The remainder of the proof follows from an immediate extension of the 3-option case just analyzed. □

Thus, for x values in the appropriate limit support, we can approximate $E(\Delta_{j,k} \mid X = x, \vartheta_{\{j,k\}}(Z) = \vartheta_{\{j,k\}}, R_j(z) = R_k(z))$ arbitrarily well by $\Delta_j^{\mathrm{LIV}}(x, \vartheta_{\mathcal{J}})$ for an arbitrarily small $\max_{l \in \mathcal{J} \setminus \{j,k\}} \{\vartheta_l\}$.

This analysis uses the ϑ index functions directly, but the results can be restated without using the ϑ functions directly. Again consider the three-choice example. The central aspect of the identification strategy is to "zero-out" the second choice by making ϑ_2 arbitrarily small, allowing one to then use the LIV estimand to identify the MTE parameter for the first option versus the third as if the second choice were not an option. If we do not know the ϑ_2 function, we cannot condition on it. However, if we know that ϑ_2 is decreasing in a particular element of Z, say $Z^{[j']}$, where $Z^{[j']}$ does not enter the index function for choices 1 and 3 and where $\vartheta_2(z_2) \to 0$ as $z^{[j']} \to -\infty$, then we can follow the same strategy as if we knew the ϑ_2 index except we condition on $Z^{[j']}$ being small instead of conditioning on ϑ_2 being small. The idea naturally extends to the case of more than three options.

We can follow Heckman and Vytlacil (1999) in following a two-step identification strategy for ATE and TT parameters of $\Delta_{j,k}$. We first identify the appropriate MTE or LATE parameters and then use them to identify ATE and TT given the appropriate support conditions. Notice that the required support conditions are now stronger than those required for the ATE and TT parameters of $\Delta_{j,\mathcal{J}\setminus j}$. For identification of the ATE and TT parameters of $\Delta_{j,\mathcal{J}\setminus j}$, we require a large support assumption only on the jth index. In particular, we require that it be possible to condition on Z values that make ϑ_j arbitrarily small or arbitrarily large while holding the remaining indices fixed. In contrast, for identification of the ATE and TT parameters of $\Delta_{j,k}$, we require a large support assumption on each index. We require that for each index we can condition on Z values that make the index arbitrarily small or arbitrarily large while holding the remaining indices fixed. The reason for this stronger condition is that for identification of $\Delta_{j,k}$ we need to use an identification at infinity strategy on all but the j and k indices to even obtain the marginal parameters. We then need an additional identification at infinity step to use the marginal parameters to recover the ATE and TT parameters.

7.3.7. Summarizing the results for the unordered model

We have obtained the following results on the unordered choice model in this section:

- $E(\Delta_{j,\mathcal{J}\setminus j} \mid X = x, Z = z, R_j(z) = R_{\mathcal{J}\setminus j}(z))$ and $E(\Delta_{j,\mathcal{J}\setminus j} \mid X = x, Z = z, R_j(\tilde{z}) \geqslant R_{\mathcal{J}\setminus j}(\tilde{z}) \geqslant R_j(z))$ can be identified without a limit argument.
- $E(\Delta_{j,k} \mid X = x, \{\vartheta_k\}_{k\in\mathcal{J}}, R_j(z) = R_k(z))$ and $E(\Delta_{j,k} \mid X = x, \{\vartheta_k\}_{k\in\mathcal{J}}, R_j(\tilde{z}) \geqslant R_k(\tilde{z}) \geqslant R_j(z))$ can be identified with a limit argument on each index in $\mathcal{J} \setminus \{j,k\}$.
- $\Delta_{j,\mathcal{J}\setminus j}^{\mathrm{ATE}}(x, z)$ and $\Delta_{j,\mathcal{J}\setminus j}^{\mathrm{TT}}(x, z)$ can be identified with a limit argument using the ϑ_j index.
- $\Delta_{j,k}^{\mathrm{ATE}}(x, z)$ and $\Delta_{j,k}^{\mathrm{TT}}(x, z)$ can be identified with a limit argument using each index.

These results establish the central role of choice theory (via $\{\vartheta_k\}_{k\in\mathcal{J}}$) and identification at infinity in using an IV strategy to identify a variety of treatment parameters and their extensions to a general multiple choice model. Our analysis extends the analysis of ordered outcome models developed in the preceding section to a general unordered case. Local instrumental variables identify the marginal treatment effect corresponding to the effect of one option versus the best alternative option without requiring large support assumptions or knowledge of the parameters of the choice model. This result preserves the spirit of the Imbens and Angrist (1994) LATE analysis and the analysis of Heckman and Vytlacil (1999, 2001b, 2005). More generally, LIV can provide identification of the marginal treatment effect corresponding to the effect of choosing between one choice set versus not having that choice set available. However, identification of the more general parameters requires knowledge (identification) of the structural, latent index functions of the multinomial choice model. LIV can also provide identification of the effect of one specified choice versus another, requiring large support assumptions but not knowledge of the latent index functions. In order to identify some treatment parameters, we require identification of the latent index functions generating the multinomial choice model or else having large support assumptions. This connects the LIV analysis in this paper to the more ambitious but demanding identification conditions for the full multinomial selection model developed in Heckman and Navarro (2007), Chapter 73 (Matzkin) of this Handbook, and Appendix B of Chapter 70. We next develop the case of the continuum of outcomes.

7.4. Continuous treatment

Thus far we have considered the case of a treatment variable taking a finite number of values. Now consider the case where the treatment variable D can take a continuum of values. Suppose that

$$Y = \mu(D, X, U),$$
$$D = \vartheta(Z, V),$$

with D a continuous random variable. We do not in general need to restrict U or V to be scalar random variables. We can rewrite this model in potential outcome notation by defining

$$Y_d \equiv \mu_d(X, U) \equiv \mu(d, X, U).$$

For ease of exposition, we will assume that X is exogenous in addition to Z being exogenous, so that $(X, Z) \perp\!\!\!\perp (U, V)$.

We assume that $\mu(d, x, u)$ is continuous in its first argument. Equivalently, we assume that $\{Y_d\}$ is continuous in d for any realization. Implicit in the continuity assumption is an ordering, that two treatments that are close to one another have associated outcomes that are close to one another. The restriction is qualitatively different from any restriction we have considered thus far. In the previous sections, there are no restrictions connecting Y_d to $Y_{d'}$. Equivalently, there are no restrictions connecting $\mu_d(X, U)$ and $\mu_{d'}(X, U)$. In the case of a continuum of treatments, we now tightly link counterfactual values that correspond to treatments that are close to one another.

The literature analyzing continuous endogenous regressors often defines the object of interest not as a treatment effect but instead as the "average structural function" (ASF). Following Blundell and Powell (2004), the ASF is defined as

$$\mu(d, x) = E(Y_d \mid X = x) = \int \mu(d, x, u)\, dF_U(u).$$

In other words, the ASF is defined as the average value of Y that would result from assigning treatment d to all individuals with $X = x$. If D is endogenous, the ASF does not in general equal the conditional expected value of Y in the data, $E(Y_d \mid X = x) \neq E(Y \mid D = d, X = x)$, since $\int \mu(d, x, u)\, dF_U(u) \neq \int \mu(d, x, u)\, dF_{U|X,D}(u \mid x, d)$. This is just a version of the distinction between fixing and conditioning introduced in Haavelmo (1943) and discussed in Chapter 70.

Instead of working with the ASF, we can follow the lead of Florens et al. (2002) and define treatment effect parameters for a continuous treatment. Suppose that $\mu(d, x, u)$ is differentiable in d for any (x, u). We can define the average treatment effect as

$$\Delta_d^{\mathrm{ATE}}(x) = E\left(\frac{\partial}{\partial d} Y_d \;\middle|\; X = x\right) = \int \frac{\partial}{\partial d} \mu(d, x, u)\, dF_U(u),$$

which is the average effect of a marginal increase in the treatment if individuals were randomly assigned treatment level d. Note that in this expression the average treatment effect depends on the base treatment level, d, and for any of the continuum of possible base treatment levels we have a different average treatment effect. The average treatment effect is the derivative of the Blundell and Powell ASF:

$$\Delta_d^{\mathrm{ATE}}(x) = \frac{\partial}{\partial d} \mu(d, x).$$

Florens et al. (2002) define treatment on the treated as

$$\begin{aligned}\Delta_d^{\mathrm{TT}}(x) &= E\left(\frac{\partial}{\partial d_1} Y_{d_1} \;\middle|\; D = d_2, X = x\right)\Bigg|_{d=d_1=d_2} \\ &= \int \left[\frac{\partial}{\partial d_1} \mu(d_1, x, u)\Bigg|_{d=d_1}\right] dF_{U|X,D}(u \mid x, d),\end{aligned}$$

which is the average effect among those currently choosing treatment level d of an incremental increase in the treatment while leaving their unobservables fixed. Likewise, define the marginal treatment effect as

$$\Delta_d^{\mathrm{MTE}}(x, v) = E\left(\frac{\partial}{\partial d} Y_d \;\middle|\; V = v, X = x\right) = \int \frac{\partial}{\partial d} \mu(d, x, u)\, dF_{U|V}(u \mid v).$$

To illustrate these definitions, suppose D is schooling level measured as a continuous variable, and suppose Y is wages. Then, e.g., Y_{12} would be the potential wage corresponding to receiving exactly 12 years of schooling and $\mu_{12} = E(Y_{12})$ is the average wage if individuals were exogenously assigned exactly 12 years of schooling. Δ_{12}^{ATE} is the average effect on wages of being assigned marginally more than 12 years of schooling versus being assigned exactly 12 years of schooling, and Δ_{12}^{TT} would be the average effect of obtaining marginally more schooling for those who self-select to obtain exactly 12 years of schooling.

One approach to identification of the treatment parameters is to impose more structure on the outcome equation while allowing the treatment selection equation to be unspecified. The nonparametric instrumental variable approach of Darolles, Florens and Renault (2002), Hall and Horowitz (2005), and Newey and Powell (2003) requires that the unobservables in the outcome equation (U) be a scalar random variable and that the outcome be an additive function of the unobservables – Chapter 73 (Matzkin) of this Handbook surveys this literature. Their additivity assumption imposes the restriction of no treatment effect heterogeneity (conditional on X), so that all treatment effect parameters coincide. In exchange for this restriction on the outcome equation, they do not require any structure on the first stage equation so that D does not need to be increasing in V and V is not required to be a scalar random variable. Furthermore, they only require that U be mean independent of (X, Z), not that (U, V) be fully independent of (X, Z).

The additive error term assumption is relaxed by Chernozhukov, Imbens, and Newey (2007), who impose the stronger requirement that the outcome is a strictly increasing function of the error term (i.e., $\mu(x, d, u)$ strictly increasing in u),[123] while strengthening the required independence property to be $(Z, X) \perp\!\!\!\perp U$. The restriction of a scalar error term with the outcome strictly increasing in this error term is again a strong restriction on the forms of treatment effect heterogeneity that are possible in the model.[124] Suppress X for ease of exposition. Under their restriction, if $\mu(d, u) > \mu(d, u')$ at some treatment level d, then $\mu(\tilde{d}, u) > \mu(\tilde{d}, u')$ for all treatment levels $\tilde{d}$. In other words, if individual one has a higher potential outcome at some value of the treatment than a second individual, than that first individual has a higher potential outcome for any value of the treatment than the second individual. Under this restriction, treatment cannot change the rank ordering of outcomes across individuals. These restrictions are in contrast with the Roy model and generalized Roy model, where one individual may have a higher with-treatment potential outcome but a lower without-treatment potential outcome compared to a second individual.

[123] More generally, that μ is a weakly separable function of U, so that μ can be rewritten as a function of a scalar aggregator of U.

[124] See also Chernozhukov and Hansen (2005), who allow for richer treatment effect heterogeneity but impose a "rank similarity" restriction that requires agents not to act upon their own individual effects. This can be shown to eliminate the general form of heterogeneous responses analyzed by the generalized Roy model. For a discussion of the analysis of Chernozhukov and Hansen (2005), see Chapter 73 (Matzkin) of this Handbook.

In contrast to these approaches, control variate approaches impose more structure on the selection equation, imposing that the unobservables in the treatment selection equation (V) be a scalar random variable,[125] and that the treatment is an additive function of the unobservables or more generally a strictly increasing function of the unobservables. Such approaches thus impose strong restrictions on the heterogeneity in the treatment selection equation. In exchange for these restrictions, such approaches do not require Y to be increasing in U and do not require U to be a scalar random variable. Imbens and Newey (2002) consider identification and estimation of the average structural function in a nonparametric model using the control variate approach, building on the work of Blundell and Powell (2004) and Altonji and Matzkin (2005). Their approach does not impose any further restrictions on the outcome equation, but does require a large support assumption. Another recent contribution to the control function literature is Florens et al. (2006), who restrict Y to be determined by a stochastic polynomial in D but do not require a large support assumption. We now further discuss both approaches.

Imbens and Newey (2002) proceed as follows. They assume that $\vartheta(z, v)$ is strictly monotonic in v. Suppose that $(U, V) \perp\!\!\!\perp (X, Z)$, and without loss of generality normalize V to be unit uniform. Then V is immediately identified (up to the normalization) from $V = F(Y \mid X, Z)$. Given identification of V, they can identify $E(Y \mid D, X, V)$. Their independence assumptions imply that $U \perp\!\!\!\perp D \mid (X, V)$, so that

$$E(Y \mid D = d, X = x, V = v) = E(Y_d \mid X = x, V = v).$$

$E(Y_d \mid X = x, V = v)$ corresponds to the marginal treatment effect except that it is the conditional expectation in level instead of the derivative of the conditional expectation. Then, in parallel to the way Heckman and Vytlacil (1999) integrate up the MTE to recover the ATE, Imbens and Newey integrate up $E(Y_d \mid X = x, V = v)$ to obtain the ASF:

$$\begin{aligned} E(Y_d \mid X = x) &= \int E(Y_d \mid X = x, V = v)\, dF_V(v) \\ &= \int E(Y \mid D = d, X = x, V = v)\, dF_V(v). \end{aligned}$$

Imbens and Newey do not explicitly consider the ATE, TT, or MTE, but we can adapt the Heckman and Vytlacil (1999) weighting analysis summarized in Section 3 to obtain these parameters as a slight modification of the Imbens and Newey analysis. First consider the MTE. We have that

$$\frac{\partial}{\partial d} E(Y \mid D = d, X = x, V = v) = E\left(\frac{\partial}{\partial d} Y_d \,\middle|\, X = x, V = v\right),$$

[125] More generally, that ϑ is a weakly separable function of V, so that ϑ can be rewritten as a function of a scalar aggregator of V.

so that the MTE is identified. Integrating up the MTE we obtain ATE

$$E\left(\frac{\partial}{\partial d}Y_d \,\Big|\, X=x\right) = \int E\left(\frac{\partial}{\partial d}Y_d \,\Big|\, X=x, V=v\right) dF_V(v)$$
$$= \int \frac{\partial}{\partial d}E(Y \mid D=d, X=x, V=v)\, dF_V(v)$$

and TT

$$E\left(\frac{\partial}{\partial d_1}Y_{d_1} \,\Big|\, D=d_2, X=x\right)\bigg|_{d=d_1=d_2}$$
$$= \int E\left(\frac{\partial}{\partial d}Y_d \,\Big|\, X=x, V=v\right) dF_{V|D=d_2,X}(v \mid x)$$
$$= \int \frac{\partial}{\partial d}E(Y \mid D=d, X=x, V=v)\, dF_{V|D=d_2,X}(v \mid x).$$

Note the strong connection between the control variate approach and the LIV/MTE approach of Heckman and Vytlacil (1999). They both proceed by identifying an expectation conditional on the first stage error term, and then integrating that expectation up to obtain the parameter of interest. The primary distinction is that, in the control variate approach with a continuous endogenous treatment, it is possible to assume that the treatment is a strictly increasing function of an error term that is independent of the instruments, to identify this error term, and then to explicitly include the identified first-stage error term as a regressor in the second stage regression for the outcome. In contrast, with a discrete endogenous treatment, it is not possible to characterize the treatment as a strictly increasing function of an error term that is independent of the instruments. It is thus not possible to identify the first-stage error term, and thus not possible to explicitly include an identified first-stage error term in the second stage. The LIV strategy is the approach in the discrete case that by-passes the need to explicitly identify the first stage error term.

In order to be able to integrate $E(Y \mid D=d, X=x, V=v) = E(Y_d \mid X=x, V=v)$ up to obtain the ASF (or to integrate MTE to obtain ATE), it is necessary to evaluate $E(Y \mid D=d, X=x, V=v)$ at all values of v in the support of the distribution of V conditional on X. This is a nontrivial requirement. To show this, suppress X for ease of exposition. One can only evaluate $E(Y \mid D=d, V=v)$ at values of v in the support of the distribution of V conditional on $D=d$, so that the requirement is that the support of the distribution of V conditional on $D=d$ equal the support of the unconditional distribution. This requires, in turn, a large support assumption on an element of Z. For example, suppose that $\vartheta(Z,V) = P(Z)+V$, so that $D=P(Z)+V$. Let $\mathcal{P}$ denote the support of the distribution of $P(Z)$. Then

$$\operatorname{Supp}(V \mid D=d) = \operatorname{Supp}\big(V \,\big|\, P(Z)+V=d\big)$$
$$= \operatorname{Supp}\big(V \,\big|\, V=d-P(Z)\big) = \{d-p\colon\ p\in\mathcal{P}\},$$

where the last equality uses $Z \perp\!\!\!\perp V$. For example, if $\mathcal{P} = [a, b]$, then $\{d - p\colon\ p \in [a, b]\} = [d - b, d - a]$ which does not depend on d if and only if $a = -\infty$ and $b = \infty$, i.e., if and only if $\mathcal{P} = \mathbb{R}$. For standard models, this requirement in turn necessitates a regressor with unbounded support, analogous to the identification at infinity requirement in selection models shown by Heckman (1990). We have noted the central role played by identification at infinity assumptions in many different settings throughout this Handbook.

Next consider the analysis of Florens et al. (2002). They assume that $(U, V) \perp\!\!\!\perp (X, Z)$. They impose additional structure on the outcome equation, in particular that the outcome equation can be expressed by a finite order stochastic polynomial in the treatment variable:

$$Y = \mu(D, X) + \sum_{j=0}^{K} D^j U_j$$

so that

$$Y_d = \mu_d(X) + \sum_{j=0}^{K} d^j U_j.$$

This specification can be seen as a nonparametric extension of the random coefficient models of Heckman and Vytlacil (1998) and Wooldridge (1997, 2003). As a consequence of the structure on the outcome equation, Florens et al. (2006) are able to identify the ATE without requiring the large support assumption of Imbens and Newey (2002). Instead of a large support assumption, they require measurable separability of D and V conditional on X.

Measurable separability is the requirement that any function of D and X that almost surely equals a function of V and X must be a function of X only. This assumption can be shown to be equivalent to requiring that D not lie in a subset of its support if and only if V lies in a subset of its support (conditional on X). As shown by Florens et al. (2006), measurable separability between D and V follows from the independence assumption $(U, V) \perp\!\!\!\perp (X, Z)$ along with mild regularity conditions. Thus the Florens, Heckman, Meghir, and Vytlacil approach allows for identification of the average treatment effect with continuous endogenous regressors without requiring large support assumptions in exchange for requiring a finite-order, stochastic polynomial assumption on the outcome equation. We next consider the method of matching, which is based on the assumption of conditional independence that is assumed to characterize data structures.

8. Matching

The method of matching assumes selection of treatment based on potential outcomes

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

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:

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

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

(Q-2) $0 < \Pr(D = 1 \mid Q) < 1$.

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. In principle, condition (Q-1) can be satisfied using a set of variables Q distinct from all or some of the components of X and Z. The conditioning variables do not have to be exogenous.

From condition (Q-1), we recover the distributions of Y_0 and Y_1 given Q, $\Pr(Y_0 \leqslant y_0 \mid Q = q) = F_0(y_0 \mid Q = q)$ and $\Pr(Y_1 \leqslant y_1 \mid Q = q) = F_1(y_1 \mid Q = q)$ – but not the joint distribution $F(y_0, y_1 \mid Q = q)$, because we do 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 we discuss in this section.

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

(Q-3) $Y = Q\alpha + D\beta + U$,

where α is the treatment effect and

(Q-4) $E(U \mid Q, D) = 0$.

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

(Q-5) $\mathrm{Var}(Q, D)$ *full rank.*

Under these conditions, we can identify β even though D and U are dependent: $D \not\!\perp\!\!\!\perp U$. Controlling for the observable Q eliminates any spurious mean dependence between D and U: $E(U \mid D) \neq 0$ but $E(U \mid D, Q) = 0$. (Q-4) is the linear regression counterpart to (Q-1). (Q-5) is the linear regression counterpart to (Q-2). Failure of (Q-5) would mean that using a nonparametric estimator, we might perfectly predict D given Q, and that $\Pr(D = 1 \mid Q = q) = 1$ or 0.[126]

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

(Q-5)$'$ If the goal of the analysis is only to identify β, in place of (Q-4) we can get by with

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

Assuming $\text{Var}(D \mid Q) > 0$, we can identify β even if we cannot separate αQ from $E(U \mid Q)$.

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. To link our notation in this section to that in the rest of the chapter, we assume that $Q = (X, Z)$ and that X and Z are the same except where otherwise noted. Thus we invoke assumptions (M-1) and (M-2) presented in Section 2, even though in principle we can use a more general conditioning set.

Assumptions (M-1) and (M-2) introduced in Section 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 we weaken (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 Section 3, so again all mean treatment parameters are the same. (3) We show that IV and matching make distinct identifying assumptions even though they both invoke conditional independence assumptions. (4) We 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) We 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) We show that standard methods for selecting the conditioning variables used in matching assume exogeneity. This is a property shared with many econometric estimators, as noted in Chapter 70, Section 5.2. Violations of the exogeneity assumption 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 22 shows the support problem that can arise in linear least squares

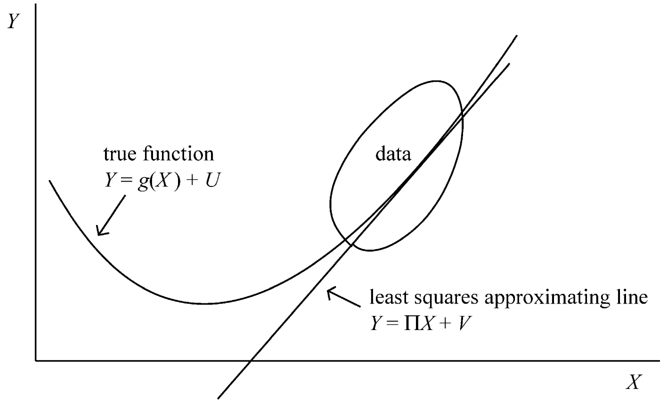


Figure 22. The least squares extrapolation problem avoided by using nonparametric regression or matching.

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 et al. (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).

We 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.

8.1. 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

$$\Delta^{\text{MTE}}(x, u_D) = E(Y_1 - Y_0 \mid X = x, U_D = u_D) = E(Y_1 - Y_0 \mid X = x). \tag{8.1}$$

This, in turn, implies that Δ^{MTE} conditional on X is flat in u_D, so that matching invokes assumption (C-1) invoked in Section 4.2.1. Under our assumptions for the generalized Roy model, it assumes that $E(Y \mid P(Z) = p)$ 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 substitutions do not contradict any observed data.

While the claim is true, it ignores 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 more general than the linear regression model (Q-3).

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

$$\begin{aligned}\Delta^{\text{OLS}}(X) &= E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 0)\\ &= E(Y_1 - Y_0 \mid X, D = 1)\\ &\quad + \bigl\{E(Y_0 \mid X, D = 1) - E(Y_0 \mid X, D = 0)\bigr\}.\end{aligned} \tag{8.2}$$

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 (8.2)) 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, $U_1 = Y_1 - E(Y_1 \mid X)$, $U_0 = Y_0 - E(Y_0 \mid X)$. Unlike the weights for Δ^{TT} and Δ^{ATE}, the OLS weights do not necessarily integrate to one and they are not necessarily nonnegative. Application of IV eliminates the contribution of the second term of Equation (8.2). 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 our 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 the 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 ($E(Y_0 \mid X, D = 1) - E(Y_0 \mid X, D = 0) < 0$). The differences among the policy relevant treatment effects, the conventional treatment effects and the OLS estimand are illustrated in Figure 4A and Table 3 for the generalized Roy model example. As is evident from Table 3, it is not at all clear that the instrumental variable estimator, with instruments that satisfy classical properties, performs better than nonparametric OLS in identifying the policy relevant treatment effect in this example. While IV eliminates the term in braces in (8.2), 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, $U_D \perp\!\!\!\perp (Y_0, Y_1) \mid X$, then $E(U_1 \mid X, U_D) = E(U_1 \mid X) = 0$ and $E(U_0 \mid X, U_D) = E(U_0 \mid X) = 0$ 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 (A-1)–(A.2). If there is such a Z, the conditional independence condition implies under (A-1)–(A-5) that $E(Y \mid X, P(Z) = p)$ is linear in p. 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 we 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 (8.2) that we can weaken (M-1) to

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

This is possible because $E(Y_1 \mid X, D = 1)$ is known 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. We now consider more generally the implications of assuming mean independence of the errors rather than full independence.

8.2. Matching and MTE using mean independence conditions

To identify all mean treatment parameters, one can weaken the assumption (M-1) to the condition that Y_0 and Y_1 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. Such a balancing is not generic in the Roy model and in the generalized Roy model.

In particular, assume that $Y_j = \mu_j(X) + U_j$ for $j = 0, 1$ and further assume that $D = \mathbf{1}[Y_1 - Y_0 \geqslant C(Z) + U_C]$. Let $V = U_C - (U_1 - U_0)$. Assume $(U_0, U_1, V) \perp\!\!\!\perp (X, Z)$. Then if $V \perp\!\!\!\perp (U_1 - U_0)$, and U_C has a log concave density, then $E(Y_1 - Y_0 \mid X, V = v)$ is decreasing in v, $\Delta^{\text{TT}}(x) > \Delta^{\text{ATE}}(x)$, and the matching conditions do not hold. If $V \perp\!\!\!\perp (U_1 - U_0)$ but V does not have a log concave density, then it is still the case that $(U_1 - U_0, V)$ is negative quadrant dependent. One can show that $(U_1 - U_0, V)$ being negative quadrant dependent implies that $\Delta^{\text{TT}}(x) > \Delta^{\text{ATE}}(x)$, and thus again that the matching conditions cannot hold. We now develop a more general analysis.

Suppose that we assume selection model (3.3) so that $D = \mathbf{1}[P(Z) \geqslant U_D]$, where Z is independent of (Y_0, Y_1) conditional on X, where $U_D = F_{V|X}(V)$ and

$P(Z) = F_{V|X}(\mu_D(Z))$. Consider the weaker mean independence assumptions in place of assumption (M-1):

(M-3) $E(Y_1 \mid X, D) = E(Y_1 \mid X)$, $E(Y_0 \mid X, D) = E(Y_0 \mid X)$.

This assumption is all that is needed to identify the mean treatment parameters because under it

$$E(Y \mid X = x, Z = z, D = 1) = E(Y_1 \mid X = x, Z = z, D = 1) = E(Y_1 \mid X = x)$$

and

$$E(Y \mid X = x, Z = z, D = 0) = E(Y_0 \mid X = x, Z = z, D = 0) = E(Y_0 \mid X = x).$$

Thus we can identify all the mean treatment parameters over the support that satisfies (M-2).

Recalling that $\Delta = Y_1 - Y_0$, (M-3) implies in terms of U_D that

$$\begin{aligned}&E\big(\Delta \,\big|\, X = x, Z = z, U_D \leqslant P(z)\big) = E(\Delta \mid X = x)\\ &\quad\Longleftrightarrow\quad E\big(\Delta^{\mathrm{MTE}}(X, U_D) \,\big|\, X = x, U_D \leqslant P(z)\big) = E(\Delta \mid X = x),\end{aligned}$$

and hence

$$\begin{aligned}&E\big(\Delta^{\mathrm{MTE}}(X, U_D) \,\big|\, X = x, U_D \leqslant P(z)\big)\\ &\quad = E\big(\Delta^{\mathrm{MTE}}(X, U_D) \,\big|\, X = x, U_D > P(z)\big).\end{aligned}$$

If the support of $P(Z)$ is the full unit interval conditional on $X = x$, then $\Delta^{\mathrm{MTE}}(X, U_D) = E(\Delta \mid X = x)$ for all U_D. If the support of $P(Z)$ is a proper subset of the full unit interval, then generically (M-3) will hold only if $\Delta^{\mathrm{MTE}}(X, U_D) = E(\Delta \mid X = x)$ for all U_D, though positive and negative parts could balance out for any particular value of X.

To see this, note that

$$\begin{aligned}&E_Z\big(E\big(\Delta^{\mathrm{MTE}}(X, U_D) \,\big|\, X = x, U_D \leqslant P(z)\big) \,\big|\, X = x, D = 1\big)\\ &\quad = E_Z\big(E\big(\Delta^{\mathrm{MTE}}(X, U_D) \,\big|\, X = x, U_D > P(z)\big) \,\big|\, X = x, D = 0\big).\end{aligned}$$

Working with $V = F_{V|X}^{-1}(U_D)$, suppose that $D = \mathbf{1}[\mu_D(Z, V) \geqslant 0]$. Let $\Omega(z) = \{v\colon \mu_D(z, v) \geqslant 0\}$. Then (M-3) implies that

$$E\big(\Delta^{\mathrm{MTE}}(X, V) \,\big|\, X = x, V \in \Omega(z)\big) = E\big(\Delta^{\mathrm{MTE}}(X, V) \,\big|\, X = x, V \in \big(\Omega(z)\big)^c\big)$$

so we expect that generically under assumption (M-3) we obtain a flat MTE in terms of $V = F_{V|X}^{-1}(U_D)$. We conduct a parallel analysis for the nonseparable choice model in Appendix K and obtain similar conditions. Matching assumes a flat MTE, i.e., that marginal returns conditional on X and V do not depend on V (alternatively, that marginal returns do not depend on U_D given X).

We already noted in Section 2 that 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.

We explore the relationship between matching with exclusion and IV in Appendix L, 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 $U_0 = Y_0 - E(Y_0 \mid X)$, we establish that 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. We next consider how to implement matching.

8.3. Implementing the method of matching

We draw on Heckman et al. (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, we 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, we 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, we 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. We denote outcomes for person i in the treatment group by Y_i^t and we match these outcomes to the outcomes of a subsample of persons in the comparison group to estimate a treatment effect. In principle, we can use a different subsample as a comparison group for each person.

In practice, we 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 comparison 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 $\mathcal{A}_i = \{j \mid X_j \in \xi(X_i)\}$. 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 $0 \leqslant W(i, j) \leqslant 1$. Form a weighted comparison group mean for person i, given by

$$\bar{Y}_i^c = \sum_{j=1}^{N_c} W(i, j) Y_j^c. \tag{8.3}$$

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. In this chapter, we briefly consider two widely-used methods. The nearest neighbor matching estimator defines $\mathcal{A}_i$ such that only one j is selected so that it is closest to X_i in some metric:

$$\mathcal{A}_i = \left\{ j \mid \min_{j \in \{1,\ldots,N_c\}} \|X_i - X_j\| \right\},$$

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

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

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

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

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 $\mathcal{A}_i = \{1, \ldots, N_c\}$, 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.[127] 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:

$$\hat{\Delta}^{\mathrm{TT}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i^t - \bar{Y}_i^c \right) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i^t - \sum_{j=1}^{N_c} W(i,j) Y_j^c \right). \tag{8.4}$$

[127] See, e.g., Härdle (1990) or Ichimura and Todd (2007) (Chapter 74 of this Handbook) for a discussion of kernels and choices of bandwidths.

We 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)].[128]

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 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

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

and

$$E(U_0 \mid X, D = 0) = E(U_0 \mid X)$$

is linear in X ($E(U \mid X) = \varphi X$). Under these assumptions, controlling for X via linear regression allows one to identify $E(Y_0 \mid X, D = 1)$ from the data on nonparticipants. Under assumption (Q-4)$'$, setting $X = Q$, this approach justifies OLS equation (Q-3) for identifying treatment effects.[129] 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", $P(X) = \Pr(D = 1 \mid X)$. Rosenbaum and Rubin (1983) demonstrate that under assumptions (M-1) and (M-2),

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid P(X) \quad \text{for } X \in \chi_c, \tag{8.5}$$

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.[130]

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

[129] In Equation (Q-3), this approach shows that α combines the effect of Q on U_0 with the causal effect of Q on Y.

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

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 nonparametrically.

Conditioning on P identifies all treatment parameters but as we have seen, it imposes the assumption of a flat MTE. Marginal returns and average returns are the same. A consequence of (8.5) is that

$$E\big(Y_1 \mid D=0, P(X)\big) = E\big(Y_1 \mid D=1, P(X)\big) = E\big(Y_1 \mid P(X)\big),$$
$$E\big(Y_0 \mid D=1, P(X)\big) = E\big(Y_0 \mid D=0, P(X)\big) = E\big(Y_0 \mid P(X)\big).$$

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 justifies 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, as we discuss 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 as we have demonstrated in Section 4, recent work in semiparametric estimation relaxes these assumptions. As noted in Section 8.2, 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 et al. (1998) and are discussed extensively in Heckman, LaLonde and Smith (1999). See Todd (1999, 2007, 2008) for software and extensive discussion of the mechanics of matching. We now contrast the identifying assumptions used in the method of control functions with those used in matching.

8.3.1. Comparing matching and control functions approaches

The method of matching eliminates the dependence between (Y_0, Y_1) and D, $(Y_0, Y_1) \not\!\perp\!\!\!\perp D$, by assuming access to conditioning variables X such that (M-1) is satisfied: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$. 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, $(Y_0, Y_1) \perp\!\!\!\perp Z \mid X$, as well as a rank condition ($\Pr(D = 1 \mid X, Z)$ 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). We 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. Chapter 73 (Matzkin) of this Handbook provides a comprehensive review of these methods. In Section 11, we present 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. All of these methods attempt to eliminate the θ in (U-1) that produces the dependence captured in (U-2).

In this section, we relate matching to the form of the control function introduced in Heckman (1980) and Heckman and Robb (1985a, 1986a). This version was used in our analysis of local instrumental variables (LIV) in Section 4, where we compare LIV with control function approaches and show that LIV and LATE estimate derivatives of the control functions. We analyze conditional means because of their familiarity. Using the fact that $E(\mathbf{1}(Y \leqslant y) \mid X) = F(y \mid X)$, the analysis applies to marginal distributions as well.

Thus we 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.[131] 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) defined in Section 2.

For the additively separable case (2.2), the control function for mean outcomes models the conditional expectations of Y_1 and Y_0 given X, Z, and D as

$$E(Y_1 \mid Z, X, D = 1) = \mu_1(X) + E(U_1 \mid Z, X, D = 1),$$
$$E(Y_0 \mid Z, X, D = 0) = \mu_0(X) + E(U_0 \mid Z, X, D = 0).$$

[131] Examples of nonseparable selection models are found in Cameron and Heckman (1998). See also Altonji and Matzkin (2005) and Chapter 73 (Matzkin) of this Handbook.

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.[132] It is not required that X or Z be stochastically independent of U_1 or U_0, although conventional methods often assume this.

Assume that $(U_0, U_1, V) \perp\!\!\!\perp (X, Z)$ and adopt Equation (3.3) 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: $D = \mathbf{1}(\mu_D(Z) \geqslant 0)$. Let $U_D = F_{V|X}(V)$ and $P(Z) = F_{V|X}(\mu_D(Z))$. In this notation, the control functions are

$$\begin{aligned} E(U_1 \mid Z, D = 1) &= E\big(U_1 \bigm| \mu_D(Z) \geqslant V\big) \\ &= E\big(U_1 \bigm| P(Z) \geqslant U_D\big) = K_1\big(P(Z)\big) \quad \text{and} \\ E(U_0 \mid Z, D = 0) &= E\big(U_0 \bigm| \mu_D(Z) < V\big) \\ &= E\big(U_0 \bigm| P(Z) < U_D\big) = K_0\big(P(Z)\big), \end{aligned}$$

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

(CF-1) $(U_0, U_1, V) \perp\!\!\!\perp X, Z$.

This assumption is not strictly required but it is traditional and useful in relating LIV and selection models (as in Section 4) and selection models and matching (this section). Under this condition

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

with $\lim_{P \to 1} K_1(P) = 0$ and $\lim_{P \to 0} K_0(P) = 0$. It is assumed that Z can be independently varied for all X, and the limits are obtained by changing Z while holding X fixed.[133] 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.[134,135]

[132] Heckman and Robb (1985a, 1986a) introduce this general formulation of control functions. The identifiability requires that the members of the pairs $(\mu_1(X), E(U_1 \mid X, Z, D = 1))$ and $(\mu_0(X), E(U_0 \mid X, Z, D = 0))$ be variation-free so that they can be independently varied against each other.

[133] More precisely, we assume that $\operatorname{Supp}(Z \mid X) = \operatorname{Supp}(Z)$ and that limit sets of Z, $\mathbb{Z}_0$, and $\mathbb{Z}_1$ exist so that as $Z \to \mathbb{Z}_0$, $P(Z, X) \to 0$, and as $Z \to \mathbb{Z}_1$, $P(Z, X) \to 1$.

[134] As noted in our discussion in Section 4, we need identification at infinity to obtain ATE and TT. This is a feature of any evaluation model with general heterogeneity.

[135] One 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 differencing.

As noted in Section 4, 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 $K_0(P)$ 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.[136] Without any functional form assumptions one cannot rule out a worst case analysis where, for example, if $X = Z$, then $K_1(P(X)) = \tau\mu(X)$ where τ 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.[137] Even though this case is not generic, it is possible. The method of matching does not require an exclusion restriction, but 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.[138] 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 $D = 0$: $\operatorname{Supp}(X \mid D = 1) \cap \operatorname{Supp}(X \mid D = 0)$. This is the requirement for any estimator that seeks to identify treatment effects by comparing samples of treated persons with samples of untreated persons.

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 dependence 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.[139] This assumption plays no role in matching if the correct conditioning set is known.[140] However, as noted below, exogeneity plays a key role in devising algorithms to select the conditioning variables. In addition, as noted in Section 6, exogeneity is helpful in making out-of-sample forecasts. The method of control functions does not require that $(U_0, U_1) \perp\!\!\!\perp D \mid (X, Z)$, which is a central requirement of matching. Equivalently, the method of control functions does not require

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

[136] No exclusion is required for many common functional forms for the distributions of unobservables.

[137] Clearly $K_1(P(X))$ and $\mu(X)$ cannot be independently varied in this case.

[138] Symmetry 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).

[139] Relaxing 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).]

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

whereas matching does and typically equates X and Z. Thus matching assumes access to a richer set of conditioning variables than is assumed in the method of control functions.

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 functions[141] in which under assumptions (M-1) and (M-2),

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

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

$$\begin{aligned} E(Y \mid X, Z, D) &= E(Y_1 \mid X, Z, D = 1)D + E(Y_0 \mid X, Z, D = 0)(1 - D)\\ &= \mu_0(X) + \big[\mu_1(X) - \mu_0(X)\big]D\\ &\quad + E\big(U_1 \bigm| P(Z), D = 1\big)D + E\big(U_0 \bigm| P(Z), D = 0\big)(1 - D)\\ &= \mu_0(X) + K_0\big(P(Z)\big)\\ &\quad + \big[\mu_1(X) - \mu_0(X) + K_1\big(P(Z)\big) - K_0\big(P(Z)\big)\big]D. \end{aligned} \tag{8.6}$$

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

$$\begin{aligned} E\big(Y \bigm| P(X), D\big) &= \mu_0\big(P(X)\big) + E\big(U_0 \bigm| P(X)\big)\\ &\quad + \big[\big(\mu_1\big(P(X)\big) - \mu_0\big(P(X)\big)\big)\\ &\quad + E\big(U_1 \bigm| P(X)\big) - E\big(U_0 \bigm| P(X)\big)\big]D. \end{aligned} \tag{8.7}$$

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

$$E\big(Y \bigm| P(X), D\big) = \mu_0\big(P(X)\big) + \big[\mu_1\big(P(X)\big) - \mu_0\big(P(X)\big)\big]D, \tag{8.8}$$

[141] See Aakvik, Heckman and Vytlacil (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 (Chapter 72).

since $E(U_1 \mid P(X)) = E(U_0 \mid P(X)) = 0$. 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 (8.7) or (8.8). 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 (8.6) for the more general control function model does not correspond to any treatment parameter, whereas the coefficients on D in Equations (8.7) and (8.8) correspond to treatment parameters under the assumptions of the matching model. Under assumption (CF-1), $\mu_1(P(X)) - \mu_0(P(X)) =$ ATE and ATE $=$ TT $=$ MTE, so the method of matching identifies all of the (conditional on $P(X)$) mean treatment parameters.[142]

Under the assumptions justifying matching, when 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_0, U_1) 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.[143] Aakvik, Heckman and Vytlacil (2005), Carneiro, Hansen and 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 discuss this work in Chapter 72.

[142] This 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

$$E\big(U_1 \mid P(X)\big) - E\big(U_0 \mid P(X)\big).$$

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

8.4. Comparing matching and classical control function methods for a generalized Roy model

Figure 10, developed in connection with our discussion of instrumental variables, shows the contrast between the shape of the MTE and the OLS matching estimand as a function of p for the extended Roy model developed in Section 4. 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, we perform sensitivity analyses under different assumptions about the parameters of the underlying selection model. In particular, we assume that the data are generated by the model of Equations (3.1) and (3.2), where $\mu_D(Z) = Z\gamma$, $\mu_0(X) = \mu_0$, $\mu_1(X) = \mu_1$, and

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

We assume in this section that $D = \mathbf{1}[\mu_D(Z) + V \geqslant 0]$, in conformity with the examples presented in Heckman and Navarro (2004), from which we draw. This reformulation of choice model (3.3) simply entails a change in the sign of V. We assume that $Z \perp\!\!\!\perp (U_0, U_1, V)$. Using the selection formulae derived in Appendix M, we can write the biases conditional on $P(Z) = p$ using propensity score matching in a generalized Roy model as

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

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 $P(z) = p$. We assume that $\mu_1 = \mu_0$ so that the true average treatment effect is zero.

We simulate the mean bias for TT (Table 10) and ATE (Table 11) for different values of the ρ_{jV} and σ_j. The results in the tables show that, as we let 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 do we avoid bias by matching.

Table 10
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 $= \rho_{0V} * \sigma_0 * M(p)$.

$M(p) = \frac{\phi(\Phi^{-1}(1-p))}{p(1-p)}$.

Source: Heckman and Navarro (2004).

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.

8.5. *The informational requirements of matching and the bias when they are not satisfied*

In this section, we present some examples of when matching "works" and when it breaks down. This section is based on Heckman and Navarro (2004). In particular, we show how matching on some of the relevant information but not all can make the bias using matching worse for standard treatment parameters. These examples also introduce factor models that play a key role in the analysis of Abbring and Heckman in Chapter 72.

Section 2 of this chapter 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

$$D^* = Z\gamma + V, \quad \text{where}$$
$$Z \perp\!\!\!\perp V \quad \text{and}$$
$$V = \alpha_{V1} f_1 + \alpha_{V2} f_2 + \varepsilon_V,$$

Table 11
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 $= \rho_{1V} * \sigma_1 * M_1(p) - \rho_{0V} * \sigma_0 * M_0(p)$.
$M_1(p) = \frac{\phi(\Phi^{-1}(1-p))}{p}$.
$M_0(p) = \frac{-\phi(\Phi^{-1}(1-p))}{(1-p)}$.
Source: Heckman and Navarro (2004).

$$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

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

We remind the reader that contrary to the analysis throughout the rest of this chapter we add V and do not subtract it in the decision equation. This is the familiar representation. By a change in sign in V, we can go back and forth between the specification used in this section and the specification used in other sections of the chapter.

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 Vytlacil (2005), Carneiro, Hansen and Heckman (2001, 2003), and Hansen, Heckman and Mullen (2004), among others. We 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,

$$I_R = \{f_1, f_2\}.$$[144]

Recall that we 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 we assume that ε_0, ε_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 2, 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, we study what happens when the matching assumption that $\sigma(I_E) \supseteq \sigma(I_R)$ does not hold. That is, we 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, we 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 $(\sigma^2_{f_1}, \sigma^2_{f_2}, \sigma^2_{\varepsilon_V}, \sigma^2_{\varepsilon_1}, \sigma^2_{\varepsilon_0})$ on the diagonal and zero in all the nondiagonal elements. This assumption links matching models to conventional normal selection models of the sort developed in Chapter 70 and further analyzed in Section 2 of this chapter. However, the examples based on this specification illustrate more general principles. We now analyze various commonly encountered cases.

[144] Notice 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_0, Y_1).

8.5.1. The economist uses the minimal relevant information: $\sigma(I_R) \subseteq \sigma(I_E)$

We 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

$$\begin{aligned} &E(Y_1 \mid D = 1, I_E) - E(Y_0 \mid D = 0, I_E) \\ &\quad = \mu_1 - \mu_0 + (\alpha_{11} - \alpha_{01}) f_1 + (\alpha_{12} - \alpha_{02}) f_2, \end{aligned}$$

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 $(\varepsilon_0, \varepsilon_1) \perp\!\!\!\perp V$. Recall that for arbitrary choices of $\alpha_{11}, \alpha_{01}, \alpha_{12}$, and α_{02}, $I_R = \{f_1, f_2\}$ and the economist needs less information to achieve (M-1) than is contained in I_E.

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 ($D^* \lesseqgtr 0$, given $f = \tilde{f}$ and $Z = z$) can be written as $\frac{\varepsilon_V}{\sigma_{\varepsilon_V}} \lesseqgtr \Phi^{-1}(1 - P(z, \tilde{f}))$, where Φ is the cdf of a standard normal random variable and $f = (f_1, f_2)$. We 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

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

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

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

and

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

where

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

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

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

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. We can estimate the treatment parameters with no bias since, as a consequence of our assumptions, $(U_0, U_1) \perp\!\!\!\perp D \mid (f, Z)$, which is exactly what matching requires. The minimal relevant information set is even smaller. For arbitrary factor loadings, we only need to know (f_1, f_2) to secure conditional independence. We can define the propensity score solely in terms of f_1 and f_2, and the Rosenbaum–Rubin result still goes through. Our 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. We 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 $Z \not\!\perp\!\!\!\perp (U_0, U_1, V)$, 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. We present an example of this point below. We first consider a case where $Z \perp\!\!\!\perp (U_0, U_1, V)$ 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.

8.5.2. 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 nonzero. Recall that we assume that Z and the f are independent. In this case, the event $(D^* \lesseqgtr 0, Z = z)$ is characterized by

$$\frac{\alpha_{V1} f_1 + \alpha_{V2} f_2 + \varepsilon_V}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} \lesseqgtr \Phi^{-1}\bigl(1 - P(z)\bigr).$$

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

$$\operatorname{Bias}\mathrm{TT}(Z = z) = \operatorname{Bias}\mathrm{TT}\bigl(P(Z) = P(z)\bigr) = \eta_0 M\bigl(P(z)\bigr), \tag{8.9}$$

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

$$\begin{aligned} \operatorname{Bias}\mathrm{ATE}(Z = z) &= \operatorname{Bias}\mathrm{ATE}\bigl(P(Z) = P(z)\bigr) \\ &= M\bigl(P(z)\bigr)\bigl\{\eta_1\bigl[1 - P(z)\bigr] + \eta_0 P(z)\bigr\}, \end{aligned} \tag{8.10}$$

where

$$\eta_1 = \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2 + \alpha_{V2}\alpha_{12}\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}},$$

$$\eta_0 = \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2 + \alpha_{V2}\alpha_{02}\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}}.$$

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 us 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. We next use expressions (8.9) and (8.10) as benchmarks against which to compare the relative size of the bias when we enlarge the econometrician's information set beyond Z.

8.5.3. 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 $P = p$ are reduced in absolute value relative to their values in Section 8.5.2 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 our model, we define expressions comparable to η_1 and η_0 for this case:

$$\eta'_1 = \frac{\alpha_{V1}\alpha_{11}\sigma^2_{f_1}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \sigma^2_{\varepsilon_V}}},$$

$$\eta'_0 = \frac{\alpha_{V1}\alpha_{01}\sigma^2_{f_1}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \sigma^2_{\varepsilon_V}}}.$$

We compare the biases under the two cases using formulae (8.9)–(8.10), 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*

$$|\eta_1(1-p) + \eta_0 p| > |\eta'_1(1-p) + \eta'_0 p|.$$

PROOF OF CONDITIONS 1 AND 2. These conditions are a direct consequence of formulae (8.9) and (8.10), 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 we condition on the same value of p in deriving these expressions although the variables in P are different across different specifications of the model. Propensity-score matching defines them conditional on $P = p$, so we are being faithful to that method.

These conditions 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.

Consider whether Condition 1 is satisfied in general. Assume $\eta_0 > 0$ for all α_{02}, α_{V2}. Then $\eta_0 > \eta_0'$ if

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

When $\frac{\alpha_{02}}{\alpha_{V2}} = 0$, clearly $\eta_0 < \eta_0'$. Adding information to the conditioning set increases bias. We can vary $(\frac{\alpha_{02}}{\alpha_{V2}})$ holding all of the other parameters constant and hence can make the left-hand side arbitrarily large.[145] As α_{02} increases, there is some critical value α_{02}^* beyond which $\eta_0 > \eta_0'$. If we 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 23A and 23B 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.[146] Values of the parameters of the model are presented at the base of the figures. In these figures, we compare conditioning on $P(z)$, which in general is not guaranteed to eliminate bias, with conditioning on $P(z)$ and f_2 but not f_1. Adding f_2 to the conditioning increases bias.

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 we change the control function from

$$K_1\big(P(Z) = P(z)\big) = \eta_1 M_1\big(P(z)\big),$$
$$K_0\big(P(Z) = P(z)\big) = \eta_0 M_0\big(P(z)\big)$$

to

$$K_1'\big(P(Z, f_2) = P(z, \tilde{f}_2)\big) = \eta_1' M_1\big(P(z, \tilde{f}_2)\big),$$

[145] A direct computation shows that

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

[146] Heckman and Navarro (2004) show comparable plots for MTE.

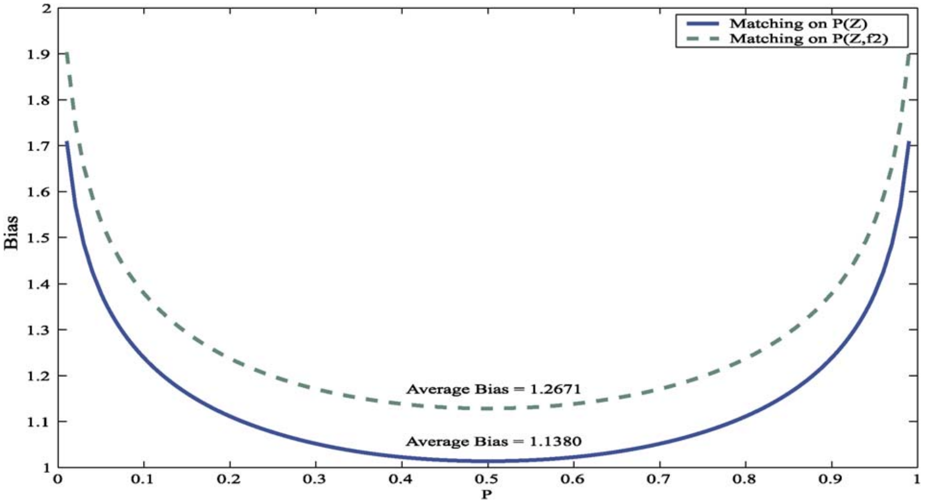
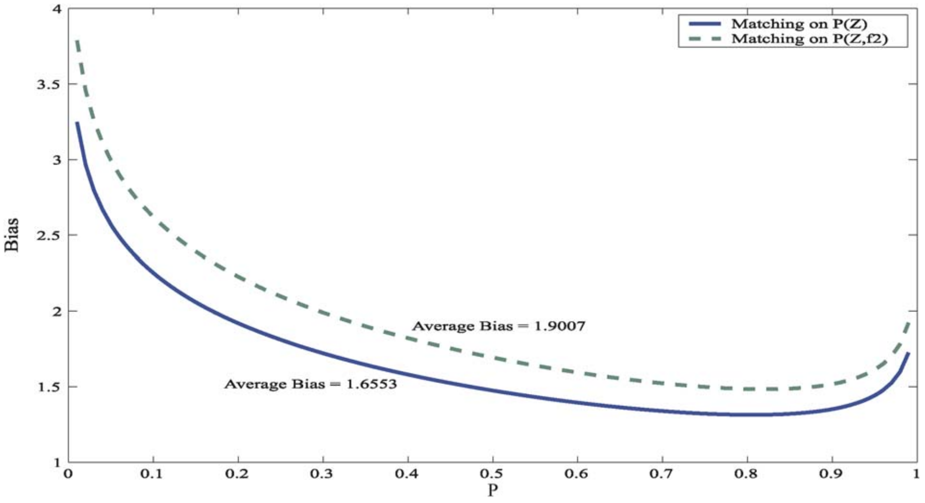


Figure 23A. Bias for treatment on the treated. *Source*: Heckman and Navarro (2004).



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

Model:

$V = Z + f_1 + f_2 + \varepsilon_V$;	$Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$;	$Y_0 = f_1 + 0.1f_2 + \varepsilon_0$
$\varepsilon_V \sim N(0, 1)$;	$\varepsilon_1 \sim N(0, 1)$;	$\varepsilon_0 \sim N(0, 1)$
$f_1 \sim N(0, 1)$;	$f_2 \sim N(0, 1)$	

Figure 23B. Bias for average treatment effect. *Source*: Heckman and Navarro (2004).

$$K_0'\big(P(Z,f_2)=P(z,\tilde{f}_2)\big)=\eta_0' M_0\big(P(z,\tilde{f}_2)\big)$$

but do not generate any bias in 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 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.

8.5.4. *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.[147] 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 we assume that

$$\tilde{Z}\sim N\big(0,\sigma_{\tilde{Z}}^2\big),$$
$$\operatorname{corr}(\tilde{Z},f_2)=\rho,\quad \text{and}\quad \tilde{Z}\perp\!\!\!\perp(\varepsilon_0,\varepsilon_1,\varepsilon_V,f_1).$$

We define expressions comparable to η and η':

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

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 8.5.3, we can obtain results for the bias in this case. 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 8.5.3.

[147] For example, the returns-to-schooling literature often uses different test scores, like AFQT or IQ, to proxy for missing ability variables. We discuss these proxy, replacement function, methods in Section 11. See also Abbring and Heckman (Chapter 72).

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 $\tilde{Z}$ is a perfect proxy for f_2. If $\rho = 0$, we are essentially back to the case analyzed in Section 8.5.3. 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 8.5.3, 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. Returning to 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

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

Figures 24A and 24B, 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 25A and 25B 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 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 we note in Section 8.5.6. Having good proxies in the standard usage of that term can create substantial bias in estimating treatment effects. Half a loaf may be worse than none.

8.5.5. *The case of a discrete outcome variable*

Heckman and Navarro (2004) construct parallel examples for cases including discrete dependent variables. In particular, they consider nonnormal, nonseparable equations for odds ratios and probabilities. The proposition that matching identifies the correct treatment parameter 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.

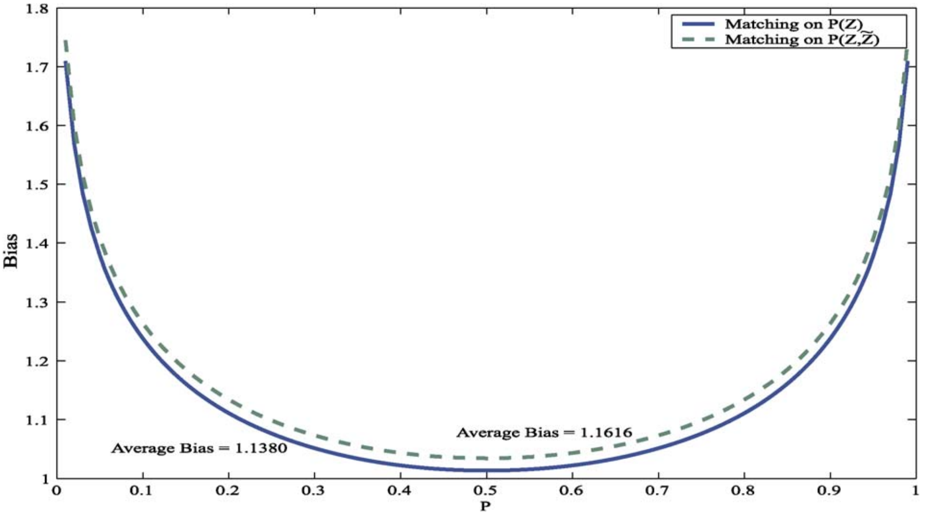
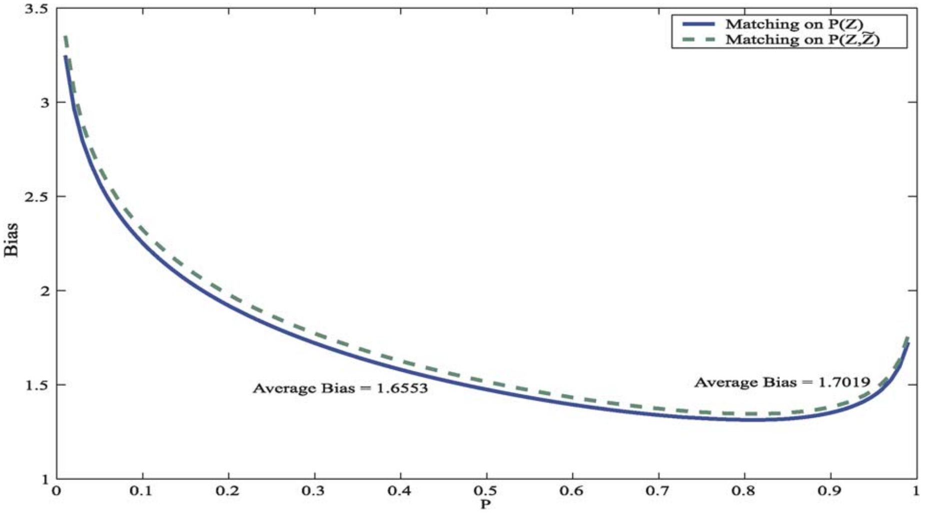


Figure 24A. Bias for treatment on the treated. *Source*: Heckman and Navarro (2004).



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

Model:

$V = Z + f_1 + f_2 + \varepsilon_V$;	$Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$;	$Y_0 = f_1 + 0.1f_2 + \varepsilon_0$
$\varepsilon_V \sim N(0, 1)$;	$\varepsilon_1 \sim N(0, 1)$;	$\varepsilon_0 \sim N(0, 1)$
$f_1 \sim N(0, 1)$;	$f_2 \sim N(0, 1)$	

Figure 24B. Bias for average treatment effect. *Source*: Heckman and Navarro (2004).

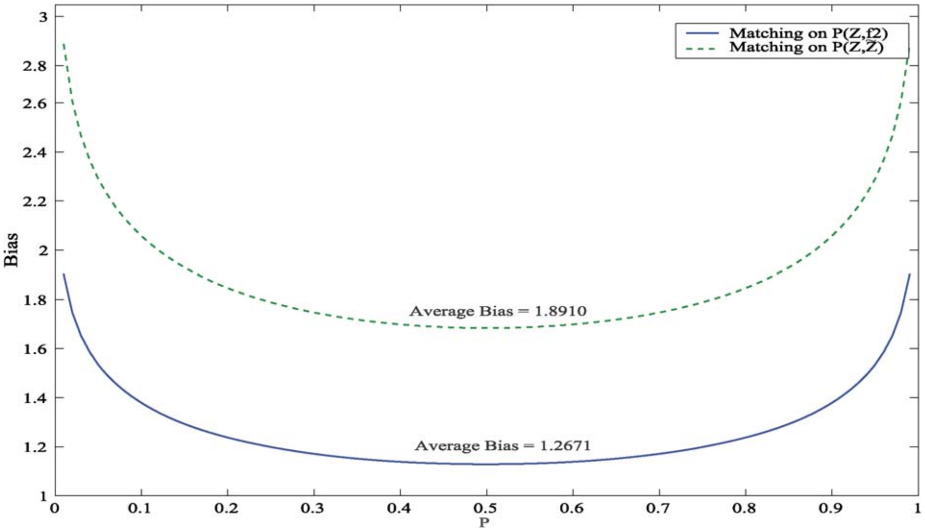
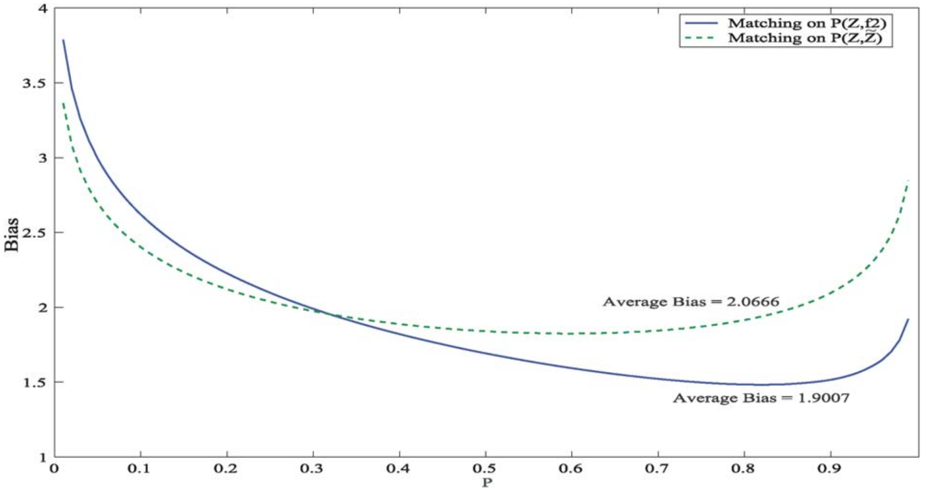


Figure 25A. Bias for treatment on the treated. *Source*: Heckman and Navarro (2004).



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

Model:

$V = Z + f_1 + f_2 + \varepsilon_V$; $Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$; $Y_0 = f_1 + f_2 + \varepsilon_0$
$\varepsilon_V \sim N(0, 1)$; $\varepsilon_1 \sim N(0, 1)$; $\varepsilon_0 \sim N(0, 1)$
$f_1 \sim N(0, 1)$; $f_2 \sim N(0, 1)$

Figure 25B. Bias for average treatment effect. *Source*: Heckman and Navarro (2004).

8.5.6. On the use of model selection criteria to choose matching variables

We have already shown by way of example that adding more variables from a 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 we have used our prototypical Roy model as our point of departure, 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 Section 8.5.3 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 et al. (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 Section 8.5.3, using f_2 improves goodness of fit of the model for D, but increases bias for the parameters. 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:

(E-1) $(Y_0, Y_1) \perp\!\!\!\perp D \mid I_{\text{int}}, \mathcal{X}$,

where I_{int} 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 $(I_{\text{int}}, \mathcal{X})$ 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 trumpeted as one of the virtues of the method. In this section, we show some examples that 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 better goodness of fit in predicting D (a higher λ in the notation of Table 12) are alleged to be better than sets of variables with lower λ in the sense that they generate lower biases. However, we have already shown that this is not true. We know that enlarging the analyst's information from $I_{\text{int}} = \{Z\}$ to $I'_{\text{int}} = \{Z, f_2\}$ will improve fit since f_2 is also in I_A and I_R. But, going from I_{int} to I'_{int} 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 12 illustrates this point using our normal example. Going from row 1 to

Table 12

Variables in probit	Goodness of fit statistics $\boldsymbol{\lambda}$		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

Model: $V = Z + f_1 + f_2 + \varepsilon_V$; $\varepsilon_V \sim N(0,1)$; $Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$; $\varepsilon_1 \sim N(0,1)$; $Y_0 = f_1 + 0.1f_2 + \varepsilon_0$; $\varepsilon_0 \sim N(0,1)$; $S_1 = V + U_1$; $U_1 \sim N(0,4)$; $S_2 = V + U_2$; $U_2 \sim N(0,0.25)$; $f_1 \sim N(0,1)$; $f_2 \sim N(0,1)$.

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 $\lambda = \lambda''$ larger than that of any other I. In our examples, conditioning on $\{Z, f_1, f_2\}$ generates zero bias. We 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 satisfy (E-1) where $I_{\text{int}} = (f_1, f_2)$. 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 we include 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 12) 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 12.

We next turn to the method of randomization, which is frequently held up to be an ideal approach for evaluating social programs. Randomization attempts to use a random assignment to achieve the conditional independence assumed in matching.

9. Randomized evaluations

This section analyzes randomized social experiments as tools for evaluating social programs. In the introduction to this chapter, we discussed an ideal randomization where treatment status is randomly assigned. In this section, we discuss actual social experiments, where self-selection decisions often intrude on the randomization decisions of experimenters.

Two cases have been made for the application of social experimentation. One case is a classical argument in experimental design. Inducing variation in regressors increases precision of estimates and the power of tests. The other case focuses on solving endogeneity and self-selection problems. Randomization is an instrumental variable.[148] The two cases are mutually compatible, but involve different emphases.

Both cases can be motivated within a linear regression model for outcome Y with treatment indicator D and covariates X:

$$Y = X\alpha + D\beta + U, \tag{9.1}$$

where U is an unobservable. β may be the same for all observations (conditional on X) as in the common coefficient setup, or it may be a variable coefficient of the type extensively discussed in this chapter. D (and the X) may be statistically dependent on U. We also entertain the possibility that when β is random it is dependent on D, as in the generalized Roy model.

Both cases for social experimentation seek to secure identification of some parameters of (9.1) or parameters that can be generated from (9.1). Analysts advocating the first case for experimentation typically assume a common coefficient model for α and β. They address the problem that variation in (X, D) may be insufficient to identify or precisely estimate (α, β). Manipulating (X, D) through randomization, or more generally, through controlled variation, can secure identification. It is typically assumed that (X, D) is independent of U or at least mean independent. This is the traditional case analyzed in a large literature on experimental design in statistics.[149]

Good examples in economics of experimentation designed to increase the variation in the regressors are studies by Conlisk (1973), Conlisk and Watts (1969), and Aigner (1979a, 1979b, 1985). The papers by Conlisk show how experimental manipulation can solve a multicollinearity problem. In analyzing the effects of taxes on labor supply, it is necessary to isolate the effect of wages (the substitution effect) from the effect of pure asset income (the income effect) on labor supply. In observational data, empirical measures of wages and asset income are highly intercorrelated. In addition, asset income is often poorly measured. By experimentally assigning these variables as in the negative income tax experiments, it is possible to identify both income and substitution effects in labor supply equations [see Cain and Watts (1973)]. Aigner (1979b) shows how variation in the prices paid for electricity across the day can identify price effects that cannot be identified in regimes with uniform prices across all hours of the day.[150]

Random assignment is not essential to this approach. Any regressor assignment rule based on variables Q that are stochastically independent of U will suffice, although the

[148] See Heckman (1996).

[149] See, e.g., Silvey (1970).

[150] Zellner and Rossi (1987) present a comprehensive discussion of this literature.

efficiency of the estimates will depend on the choice of Q and care must be taken to avoid inducing multicollinearity by the choice of an assignment rule.

The second case for social experiments and the one that receives the most attention in applied work in economics and in this chapter focuses on the dependence between (X, D) and U that invalidates least squares as an estimator of the causal effect of X and D on Y. This is the problem of least squares bias raised by Haavelmo (1943) and extensively discussed in Chapter 70. In the second case, experimental variation in (X, D) is sought to make it "exogenous" or "external" to U. A popular argument in favor of experiments is that they produce simple, transparent estimates of the effects of the programs being evaluated in the presence of such biases. A quotation from Banerjee (2006) is apt:

> *The beauty of randomized evaluations is that the results are what they are*: *we compare the outcome in the treatment* [*group*] *with the outcome in the control group, see whether they are different, and if so by how much. Interpreting quasi-experiments sometimes requires statistical legerdemain, which makes them less attractive …*

This argument assumes that interesting evaluation questions can be answered by the marginal distributions produced from experiments. It also assumes that no economic model is needed to interpret evidence, contrary to a main theme of this chapter.

Randomization is an instrument. As such, it shares all of the assets and liabilities of IV already discussed. In particular, randomization applied to a correlated random coefficient (or a model of essential heterogeneity) raises the same issues about the multiplicity of parameters identified by different randomizations as were discussed there in connection with the multiplicity of parameters identified by different instruments.

The two popular arguments for social experimentation are closely related. Exogenous variation in (X, D) can, if judiciously administered, solve collinearity, precision, and endogeneity problems. Applying the terminology of Chapter 70 to the analysis of model (9.1), randomization can identify a model that can solve all three policy evaluation problems: P-1, the problem of internal validity; P-2, the problem of extrapolation to new environments (by virtue of the linearity of (9.1)); and P-3, the problem of forecasting new policies that can be described by identifiable functions of (X, D) and any external variables.

As noted in the concluding section of Chapter 70, the modern literature tends to reject functional form assumptions such as those embodied in Equation (9.1). It has evolved towards a more focused attempt to solve problem P-1 to protect against endogeneity of D with respect to U. Sometimes the parameter being identified is not clearly specified. When it is, this focus implements Marschak's Maxim of doing one thing well, as discussed in Chapter 70.

Common to the literature on IV estimation, proponents of randomization often ignore the consequences of heterogeneity in β and dependence of β on D – the problem of essential heterogeneity. Our discussion in the previous sections applies with full force to randomization as an instrument. Only if the randomization (instrument) corresponds exactly to the policy that is sought to be evaluated will the IV (randomization) identify the parameters of economic interest.[151] This section considers the case for randomization as an instrumental variable to solve endogeneity problems.

9.1. Randomization as an instrumental variable

The argument justifying randomization as an instrument assumes that randomization (or more generally the treatment assignment rule) does not alter subjective or objective potential outcomes. This is covered by assumption (PI-3) presented in Chapter 70. We also maintain absence of general equilibrium effects (PI-4) throughout this section. We discuss violations of (PI-3) when we discuss randomization bias.[152,153]

To be explicit about particular randomization mechanisms, we return to our touchstone generalized Roy model. Potential outcomes are (Y_0, Y_1) and cost of participation is C. Assume perfect certainty in the absence of randomization. Under self-selection, the treatment choice is governed by

$$D = \mathbf{1}(Y_1 - Y_0 - C \geqslant 0).$$

This model of program participation abstracts from the important practical feature of many social programs that multiple agents contribute to decisions about program participation. We consider a more general framework in Section 9.5. We assume additive separability between the observables (X, W) and the unobservables (U_0, U_1, U_C) for convenience:

$$\begin{aligned}
&Y_1 = \mu_1(X) + U_1, \qquad Y_0 = \mu_0(X) + U_0,\\
&C = \mu_C(W) + U_C, \qquad V = U_1 - U_0 - U_C,\\
&\mu_I(X, W) = \mu_1(X) - \mu_0(X) - \mu_C(W), \qquad Z = (X, W).
\end{aligned}$$

Only some components of X and/or W may be randomized. Randomization can be performed unconditionally or conditional on strata, Q, where the strata may or may not include components of (X, W) that are not randomized. Specifically, it can be performed conditional on X, just as in our analysis of IV. Parameters can be defined conditional on X.[154] Examples of treatments randomly assigned include the tax/benefit plans of the negative income tax programs; the price of electricity over the course of the day; variable tolls and bonuses; textbooks to pupils; reducing class size. Under invariance condition (PI-3), the functions $\mu_0(X)$, $\mu_1(X)$, $\mu_C(W)$ (and hence $\mu_I(X, W)$) are

[151] The exchange between Banerjee (2006) and Deaton (2006) raises this point.

[152] We maintain the absence of general equilibrium or spill over effects, assumption (PI-2). Such effects are discussed in Abbring and Heckman (Chapter 72).

[153] For evaluation of distributional and mean parameters, assumption (PI-3) can be weakened as in our invocation of policy invariance for the MTE to say that randomization does not alter the distributions of outcomes or certain means or conditional means (recall assumption (A-7)).

[154] In Equation (9.1), if X is endogenous and we randomize treatment D conditional on X with respect to U, we cannot identify α, but we can identify β.

invariant to such modifications. The intervention is assumed to change the arguments of functions without shifting the functions themselves. Thus for the intervention of randomization, the functions are assumed to be structural in the sense of Hurwicz (1962). The distributions of (U_0, U_1, U_C) conditional on X, and hence the distribution of V conditional on X, are also invariant. Under full compliance, the manipulated Z are the same as the Z facing the agent. We formalize this assumption:

(R-4) *The Z assigned agent ω conditional on X are the Z realized and acted on by the agent conditional on X.*[155]

In terms of the generalized Roy model, this assumption states that the Z assigned ω given X is the W that appears in the cost function and the derived decision rule.

Some randomizations alter the environments facing agents in a more fundamental way by introducing new random variables into the model instead of modifying the variables that would be present in a pre-experimental environment. Comparisons of these randomizations involve an implicit dynamics, better exposited using the dynamic models presented in Abbring and Heckman (Chapter 72). For simplicity and to present some main ideas, we initially invoke an implicit dynamics suitable to the generalized Roy model. We develop a more explicit dynamic model of randomized evaluation in Section 9.5.

The most commonly used randomizations restrict eligibility either in advance of agent decisions about participation in a program or after agent decisions are made, but before actual participation begins. Unlike statistical discussions of randomization, we build agent choice front and center into our analysis. Agents choose and experimenters can only manipulate choice sets.

Let $\xi = 1$ if an agent is eligible to participate in the program; $\xi = 0$ otherwise. $\tilde{\xi} = \{0, 1\}$ is the set of possible values of ξ. Let D indicate participation under ordinary conditions. In the absence of randomization, D is an indicator of whether the agent actually participates in the program. Let actual participation be A. By construction, under invariance condition (PI-3) presented in Chapter 70,

$$A = D\xi. \tag{9.2}$$

This assumes that eligibility is strictly enforced.

There is a distinction between desired participation by the agent (D) and actual participation (A). This distinction is conceptually distinct from the *ex-ante*, *ex-post* distinction. At all stages of the application and enrollment process, agents may be perfectly informed about their value of ξ and desire to participate (D), but may not be allowed to participate. On the other hand, the agent may be surprised by ξ after applying to the program. In this case, there is revelation of information and there is a distinction between *ex ante* expectations and *ex post* realizations. Our analysis covers both cases.

We consider two types of randomization of eligibility.

[155] Assumptions (R-1)–(R-3) are presented in Section 2.

RANDOMIZATION OF TYPE 1. *A random mechanism (possibly conditional on* (X, Z)*) is used to determine* ξ. *The probability of eligibility is* $\Pr(\xi = 1 \mid X, Z)$.

For this type of randomization, in the context of the generalized Roy model, it is assumed that

(e-1a) $\xi \perp\!\!\!\perp (U_0, U_1, U_C) \mid X, Z$ (*Randomization of eligibility*)

and

(e-1b) $\Pr(A = 1 \mid X, Z, \xi)$ *depends on* ξ.

This randomization affects the eligibility of the agent for the program but because agents still self-select, there is no assurance that eligible agents will participate in the program. This condition does not impose exogeneity on X, Z.[156] Thus Z can fail as an instrument but ξ remains a valid instrument. Alternatively, (e-1a) and (e-1b) may be formulated according to the notation of Imbens and Angrist (1994). Define $A(z, e)$ to be the value of A when we set $Z = z$ and $\xi = e$. Define $\mathcal{Z}$ as the set of admissible Z and $\widetilde{\xi}$ as the set of admissible ξ. In this notation, we may rewrite assumptions (e-1a) and (e-1b) as

(e-1a)$'$ $\xi \perp\!\!\!\perp (Y_0, Y_1, \{A(z, e)\}_{(z,e)\in\mathcal{Z}\times\tilde{\xi}}) \mid X, Z$

and

(e-1b)$'$ $\Pr(A = 1 \mid X, Z, \xi)$ *depends on* ξ.[157]

A second type of randomization conditions on individuals manifesting a desire to participate through their decision to apply to the program. This type of randomization is widely used.

RANDOMIZATION OF TYPE 2. *Eligibility may be a function of D (conditionally on some or all components of* X, Z, Q *or unconditionally). It is common to deny entry into programs among people who applied and were accepted into the program* $(D = 1)$ *so the probability of eligibility is* $\Pr(\xi = 1 \mid X, Z, Q, D = 1)$. *This assumes* (PI-3) *stated in Chapter* 70.

For this type of randomization of eligibility, it is assumed that

(e-2a) $\xi \perp\!\!\!\perp (U_0, U_1) \mid X, Z, Q, D = 1$

and

[156] In place of the randomization, one might assign treatment on the basis of external variables Q including variables in addition to X and Z. Care must be taken to avoid inducing collinearity problems. Random assignment is simpler. It produces through randomization the independent variation assumed in matching.

[157] When ξ is deterministic, (e-1a)$'$ is trivially satisfied.

(e-2b) $\Pr(A = 1 \mid X, Z, D = 1, \xi = 1) = 1; \Pr(A = 1 \mid X, Z, D = 1, \xi = 0) = 0.$

Agent failure to comply with the eligibility rules or protocols of experiments can lead to violations of (e-1) and/or (e-2).

An equivalent way to formulate (e-2a) and (e-2b) uses the Imbens–Angrist notation for IV:

(e-2a)′ $\xi \perp\!\!\!\perp (Y_0, Y_1) \mid X, Z, Q, D = 1$

and

(e-2b)′ $\Pr(A = 1 \mid X, Z, D = 1, \xi = 1) = 1; \Pr(A = 1 \mid X, Z, D = 1, \xi = 0) = 0.$

Both randomizations are instruments as defined in Section 4. Under the stated conditions, both satisfy (IV-1) and (IV-2), suitably redefined for eligibility randomizations, replacing D by A.

A variety of conditioning variables is permitted by these definitions. Thus, (e-1) and (e-2) allow for the possibility that the conventional instruments Z fail (IV-1) and (IV-2), but nonetheless the randomization generates a valid instrument ξ. The simplest randomizations do not condition on any variables.[158] We next consider what these instruments identify.

9.2. *What does randomization identify?*[159]

Under invariance assumption (PI-3) and under one set of randomization assumptions just presented, IV is an instrument that identifies some treatment effect for an ongoing program. The question is: which treatment effect? Following our discussion of IV with essential heterogeneity presented in Section 4, different randomizations (or instruments) identify different parameters unless there is a common coefficient model ($Y_1 - Y_0 = \beta(X)$ is the same for everyone given X) or unless there is no dependence between the treatment effect ($Y_1 - Y_0$) and the indicator D of the agents' desire to participate in the treatment. In these two special cases, all mean treatment parameters are the same. Using IV, we can identify the marginal distributions $F_0(y_0 \mid X)$ and $F_1(y_1 \mid X)$.[160]

In a model with essential heterogeneity, the instruments generated by randomization can identify parameters that are far from the parameters of economic interest. Randomization of components of W (or Z given X) under (R-4) and conditions (IV-1) and (IV-2) from Section 2 produces instruments with the same problems and possibilities as analyzed in our discussion of instrumental variables. Using W as an instrument may lead to negative weights on the underlying LATEs or MTEs.[161] Thus, unless we condi-

[158] We do not discuss optimal randomized experiments and the best choice of a randomization mechanism.

[159] This subsection is based on Heckman (1992).

[160] We can also identify $F_0(y_0 \mid X, Z)$ and $F_1(y_1 \mid X, Z)$ if Z does not satisfy the conditions required for it to be an instrument but experimental variation provides new instruments.

[161] In the special case where randomization of some components of W makes them fully independent of the other components of W, under monotonicity for the randomized component irrespective of the values of the other components, the IV weights must be nonnegative.

tion on the other instruments, the IV defined by randomization can be negative even if all of the underlying treatment effects or LATEs and MTEs generating choice behavior are positive. The weighted average of the MTE generated by the instrument may be far from the policy relevant treatment effect.

Under (PI-3) and (e-1), or equivalently (e-1)′, the first type of eligibility randomization identifies $\Pr(D = 1 \mid X, Z)$ (the choice probability) and hence relative subjective evaluations, and the marginal outcome distributions $F_0(y_0 \mid X, D = 0)$ and $F_1(y_1 \mid X, D = 1)$ for the eligible population ($\xi = 1$). Agents made eligible for the program self-select as usual. For those deemed ineligible ($\xi = 0$), under our assumptions, we would identify the distribution of Y_0, which can be partitioned into components for those who would have participated in the program had it not been for the randomization and components for those who would not have participated if offered the opportunity to do so:

$$\begin{aligned}F_0(y_0 \mid X) = {}& F_0(y_0 \mid X, D = 0)\Pr(D = 0 \mid X)\\ &+ F_0(y_0 \mid X, D = 1)\Pr(D = 1 \mid X).\end{aligned}$$

Since we know $F_0(y_0 \mid X, D = 0)$ and $\Pr(D = 1 \mid X)$ from the eligible population, we can identify $F_0(y_0 \mid X, D = 1)$. This is the new piece of information produced by the randomization compared to what can be obtained from standard observational data. In particular, we can identify the parameter TT, $E(Y_1 - Y_0 \mid X, D = 1)$, but without further assumptions, we cannot identify the other treatment parameters ATE ($= E(Y_1 - Y_0 \mid X)$) or the joint distributions $F(y_0, y_1 \mid X)$ or $F(y_0, y_1 \mid X, D = 1)$.

To show that ξ is a valid instrument for A, form the Wald estimand,

$$\text{IV}_{(\text{e-1})} = \frac{E(Y \mid \xi = 1, Z, X) - E(Y \mid \xi = 0, Z, X)}{\Pr(A = 1 \mid \xi = 1, Z, X) - \Pr(A = 1 \mid \xi = 0, Z, X)}. \tag{9.3}$$

Under invariance assumption (PI-3), $\Pr(D = 1 \mid Z, X)$ is the same in the presence or absence of randomization.[162] Assuming full compliance so that agents randomized to ineligibility do not show up in the program,

$$\Pr(A = 1 \mid \xi = 0, Z, X) = 0,$$

and

$$\begin{aligned}E(Y \mid \xi = 0, Z, X) &= E(Y_0 \mid Z, X)\\ &= E(Y_0 \mid D = 1, X, Z)\Pr(D = 1 \mid X, Z)\\ &\quad + E(Y_0 \mid D = 0, X, Z)\Pr(D = 0 \mid X, Z).\end{aligned}$$

If Z also satisfies the requirement (IV-1) that it is an instrument, then $E(Y_0 \mid Z, X) = E(Y_0 \mid X)$. Under (e-1) or (e-1)′ we do not have to assume that Z is a valid instrument.[163] Using (e-1) and assumption (PI-3), the first term in the numerator of (9.3) can

[162] $\Pr(D = 1 \mid Z, X, \xi = 0) = \Pr(D = 1 \mid Z, X, \xi = 1)$.

[163] If Z fails to be an instrument, absorb Z into X.

be written as

$$\begin{aligned}E(Y \mid \xi = 1, Z, X) = E(Y_1 \mid D = 1, Z, X)\Pr(D = 1 \mid Z, X) \\ + E(Y_0 \mid D = 0, Z, X)\Pr(D = 0 \mid Z, X).\end{aligned}$$

Substituting this expression into the numerator of Equation (9.3) and collecting terms, $\mathrm{IV}_{(\text{e-1})}$ identifies the parameter treatment on the treated:

$$\mathrm{IV}_{(\text{e-1})} = E(Y_1 - Y_0 \mid D = 1, Z, X).$$

It does not identify the other mean treatment effects, such as LATE or the average treatment effect ATE, unless the common coefficient model governs the data or $(Y_1 - Y_0)$ is mean independent of D. Using the result that $F(y \mid X) = E(\mathbf{1}(Y \leqslant y) \mid X)$, $\mathrm{IV}_{(\text{e-1})}$ also identifies $F_0(y_0 \mid X, D = 1)$, since we can compute conditional means of $\mathbf{1}(Y \leqslant y)$ for all y. The distribution $F_1(y_1 \mid X, D = 1)$ can be identified from observational data. Thus we can identify the outcome distributions for Y_0 and for Y_1 separately, conditional on $D = 1, X, Z$, but without additional assumptions we cannot identify the joint distribution of outcomes or the other treatment parameters.

Randomization not conditional on (X, Z) $(\xi \perp\!\!\!\perp (X, Z))$ creates an instrument ξ that satisfies the monotonicity or uniformity conditions. If the randomization is performed on (X, Z) strata, then the IV must be used conditional on the strata variables to ensure monotonicity is satisfied.

The second type of eligibility randomization proceeds conditionally on $D = 1$. Accordingly, data generated from such experiments do not identify choice probabilities $(\Pr(D = 1 \mid X, Z))$ and hence do not identify the subjective evaluations of agents [Heckman (1992), Moffitt (1992)]. Under (PI-3) and (e-2) (or equivalent conditions (e-2)$'$) randomization identifies $F_0(y_0 \mid D = 1, X, Z)$ from the data on the randomized-out participants. This conditional distribution cannot be constructed from ordinary observational data unless additional assumptions are invoked. From the data for the eligible $(\xi = 1)$ population, we identify $F_1(y_1 \mid D = 1, X, Z)$.

The Wald estimator for mean outcomes in this case is

$$\mathrm{IV}_{(\text{e-2})} = \frac{E(Y \mid D = 1, \xi = 1, X, Z) - E(Y \mid D = 1, \xi = 0, X, Z)}{\Pr(A = 1 \mid D = 1, \xi = 1, X, Z) - \Pr(A = 1 \mid D = 1, \xi = 0, X, Z)}.$$

Under (e-2)/(e-2)$'$,

$$\begin{aligned}&\Pr(A = 1 \mid D = 1, \xi = 1, X, Z) = 1,\\ &\Pr(A = 1 \mid D = 1, \xi = 0, X, Z) = 0,\\ &E(Y \mid A = 0, D = 1, \xi = 0, X, Z) = E(Y_0 \mid D = 1, X, Z) \quad \text{and}\\ &E(Y \mid A = 1, D = 1, \xi = 1, X, Z) = E(Y_1 \mid D = 1, X, Z).\end{aligned}$$

Thus,

$$\mathrm{IV}_{(\text{e-2})} = E(Y_1 - Y_0 \mid D = 1, X, Z).$$

In the general model with essential heterogeneity, randomized trials with full compliance that do not disturb the activity being evaluated answer a limited set of questions, and do not in general identify the policy relevant treatment effect (PRTE). Randomizations have to be carefully chosen to make sure that they answer interesting economic questions. Their analysis has to be supplemented with the methods previously analyzed to answer the full range of policy questions addressed there.

Thus far we have assumed that the randomizations do not violate the invariance assumption (PI-3). Yet many randomizations alter the environment they are studying and inject what may be unwelcome sources of uncertainty into agent decision making. We now examine the consequences of violations of invariance.

9.3. *Randomization bias*

If randomization alters the program being evaluated, the outcomes of a randomized trial may bear little resemblance to the outcomes generated by an ongoing version of the program that has not been subject to randomization. In this case, assumption (PI-3) is violated. Such violations are termed "Hawthorne effects" and are called "randomization bias" in the economics literature.[164] The process of randomization may affect objective outcomes, subjective outcomes or both.

Even if (PI-3) is violated, randomization may still be a valid instrument for the altered program. Although the program studied may be changed, under the assumptions made in Section 9.2, randomization can produce "internally valid" treatment effects for the altered program. Thus randomization can answer policy question P-1 for a program changed by randomization, but not for the program as it would operate in the absence of randomization.

As noted repeatedly, a distinctive feature of the econometric approach to social program evaluation is its emphasis on choice and agent subjective evaluations of programs. This feature accounts for the distinction between the statistician's invariance assumption (PI-1) and the economist's invariance assumption (PI-3). (These are presented in Chapter 70.) It is instructive to consider the case where assumption (PI-1) is valid but assumption (PI-3) is not. This case might arise when randomization alters risk-averse agent decision behavior but has no effects on potential outcomes. Thus the $R(s, \omega)$ are affected, but not the $Y(s, \omega)$.

In this case, the parameter $\text{ATE}(X) = E(Y_1 - Y_0 \mid X)$ is the same in the ongoing program as in the population generated by the randomized trial. However, treatment parameters conditional on choices such as

$$\text{TT}(X) = E(Y_1 - Y_0 \mid X, D = 1),$$
$$\text{TUT}(X) = E(Y_1 - Y_0 \mid X, D = 0)$$

[164] See Campbell and Stanley (1963) for a discussion of Hawthorne effects and evidence of their prevalence in educational interventions. See Heckman (1992) for a discussion of randomization bias in economics.

are not, in general, invariant. If the subjective valuations are altered, so are the parameters based on choices produced by the subjective valuations. Different random variables generate the conditioning sets in the randomized and nonrandomized regimes and, in general, they will have a different dependence structure with the outcomes $Y(s, \omega)$. This arises because randomization alters the composition of participants in the conditioning set that defines the treatment parameter.

This analysis applies with full force to LATE. LATE based on $P(Z)$ for two distinct values of Z ($Z = z$ and $Z = z'$) is $E(Y_1 - Y_0 \mid X, P(z') \leqslant U_D \leqslant P(z))$. In the randomized trial, violation of (PI-3) because of lack of invariance of $R(s, \omega)$ changes U_D and the values of $P(Z)$ for the same $Z = z$. In general, this alters LATE.[165]

The case where (PI-1) holds, but (PI-3) does not, generates invariant conditional (on choice) parameters if there is no treatment effect heterogeneity or if there is such heterogeneity that is independent of D. These are the familiar conditions: (a) $Y_1 - Y_0$ is the same for all people with the same $X = x$ or (b) $Y_1 - Y_0$ is (mean) independent of D given $X = x$. In these cases, the MTE is flat in U_D.

In general, in a model with essential heterogeneity, even if the Rubin invariance conditions (PI-1) and (PI-2) are satisfied, but conditions (PI-3) and (PI-4) are not, treatment parameters defined conditional on choices are not invariant to the choice of randomization.[166] This insight shows the gain in clarity in interpreting what experiments identify from adopting a choice-theoretic, econometric approach to the evaluation of social programs, as opposed to the conventional approach adopted by statisticians. We now show another advantage of the economic approach in an analysis of noncompliance and its implications for interpreting experimental evidence.

9.4. Compliance

The statistical treatment effect literature extensively analyzes the problem of noncompliance.[167] Persons assigned to a treatment may not accept it. In the notation of Equation (9.3), let $\xi = 1$ if a person is assigned to treatment, $\xi = 0$ otherwise. Compliance is said to be perfect when $\xi = 1 \Rightarrow A = 1$ and $\xi = 0 \Rightarrow A = 0$. In the presence of self-selection by agents, these conditions do not, in general, hold. People assigned to treatment may not comply ($\xi = 1$ but $D = 0$). This is also called the "dropout" problem [Mallar, Kerachsky and Thorton (1980), Bloom (1984)]. In its formulation of this problem, the literature assumes that outcomes are measured for each participant but that outcomes realized are not always those intended by the randomizers.[168] In addition,

[165] Technically, for identifying MTE or LATE, we can get by with weaker conditions than (PI-3) and (PI-4). All we need is invariance of the conditional mean of $Y_1 - Y_0$ with respect to U_D. Recall our discussion of policy invariance surrounding our discussion of assumption (A-7).

[166] Rubin combines (PI-1) and (PI-2) in his "SUTVA" condition.

[167] See, e.g., Bloom (1984), Manski (1996), and Hotz, Mullin and Sanders (1997).

[168] The problem of missing data is called the attrition problem. Thus we assume no attrition from the database, but we allow for the possibility that people assigned to a treatment do not receive it.

people denied treatment may find substitutes for the treatment outside of the program. This is the problem of substitution bias. Since self-selection is an integral part of choice models, noncompliance, as the term is used by the statisticians, is a feature of most social experiments.

The econometric approach builds in the possibility of self-selection as an integral part of model specification. As emphasized in the econometric literature since the work of Gronau (1974), Heckman (1974a, 1974b, 1976b), and McFadden (1974), agent decisions to participate are informative about their subjective evaluations of the program. In the dynamic setting discussed in Section 3 of Chapter 72 of this Handbook, agent decisions to attrite from a program are informative about their update of information about the program [Heckman and Smith (1998), Chan and Hamilton (2006), Smith, Whalley and Wilcox (2006) and Heckman and Navarro (2007)]. Noncompliance is a source of information about subjective evaluations of programs.

Noncompliance is a problem if the goal of the social experiment is to estimate $\text{ATE}(X) = E(Y_1 - Y_0 \mid X)$ without using the econometric methods previously discussed. We established in Section 9.3 that in the presence of self-selection, in a general case with essential heterogeneity, experiments under assumptions (PI-3) and (PI-4) and (e-1) or (e-2) identify $E(Y_1 - Y_0 \mid X, D = 1)$ instead of $\text{ATE}(X)$.

Concerns about noncompliance often arise from adoption of the Neyman–Cox–Rubin "causal model" discussed in Chapter 70, Section 4.4. Experiments are conceived as tools for direct allocation of agricultural treatments. For that reason, that literature elevates ATE to pre-eminence as the parameter of interest because it is thought that this parameter can be produced by experiments. In social experiments, it is rare that the experimenter can force anyone to do anything. As the old adage goes, "you can lead a horse to water but you cannot make it drink". Agent choice behavior intervenes. Thus it is no accident that if they are not compromised, the two randomizations most commonly implemented directly identify parameters conditional on choices.[169]

There is a more general version of the noncompliance problem which requires a dynamic formulation. Agents are assigned to treatment ($\xi = 1$) and some accept ($D = 1$) but drop out of the program at a later stage. We need to modify the formulation in this section to cover this case. We now turn to that modification.

9.5. The dynamics of dropout and program participation

Actual programs are more dynamic in character than the stylized program just analyzed. Multiple actors are involved, such as the agents being studied and the groups administering the programs. People apply, are accepted, enroll, and complete the program. A fully dynamic analysis, along the lines of the models developed by Abbring

[169] Randomizations of treatment to entire geographically segmented regions can produce ATE assuming homogeneity in background conditions across regions. This is the logic behind the Progressa experiment [see Behrman, Sengupta and Todd (2005)].

and Heckman in Chapter 72, analyzes each of these decisions, accounting for the updating of agent and program administrators' information.[170] This section briefly discusses some new issues that arise in a more dynamic formulation of the dropout problem. Heckman (1992), Heckman, Smith and Taber (1998), Hotz, Mullin and Sanders (1997), and Manski (1996, 2003) discuss these issues in greater depth.

In this subsection, we analyze the effects of dropouts on inferences from social experiments and assume no attrition. Our analysis of this case is of interest both in its own right and as a demonstration of the power of our approach.

Consider a stylized multiple stage program. In stage "0", the agent (possibly in conjunction with program officials) decides to participate or not to participate in the program. This is an enrollment phase prior to treatment. Let $D_0 = 1$ denote that the agent does not choose to participate. $D_0 = 0$ denotes that the agent participates and receives some treatment among J possible program levels beyond the no treatment state. The outcome associated with state "0" is Y_0. This assumes that acts of inquiry about a program or registration in it have no effect on outcomes.[171] One could disaggregate stage "0" into recruitment, application, and acceptance stages, but for expositional simplicity we collapse these into one stage.

If the J possible treatment stages are ordered, say, by the intensity of treatment, then "1" is the least amount of treatment and "J" is the greatest amount. A more general model would allow people to transit to stage j but not complete it. The J distinct stages can be interpreted quite generally. If a person no longer participates in the program after stage j, $j = 1, \dots, J$, we set indicator $D_j = 1$. The person is assumed to receive stage j treatment. $D_J = 1$ corresponds to completion of the program in all J stages of its treatment phase. Note that, by construction, $\sum_{j=0}^{J} D_j = 1$. The sequential updating model developed by Abbring and Heckman in Chapter 72 can be used to formalize these decisions and their associated outcomes. We can also use the simple multinomial choice model developed and analyzed in Appendix B of Chapter 70.

Let $\{D_j(z)\}_{z \in \mathcal{Z}}$ be the set of potential treatment choices for choice j associated with setting $Z = z$. For each $Z = z$, $\sum_{j=0}^{J} D_j(z) = 1$. Using the methods exposited in Abbring and Heckman (Chapter 72), we could update the information sets at each stage. We keep this updating implicit. Different components of Z may determine different choice indicators. Array the collections of choice indicators evaluated at each $Z = z$ into a vector

$$D(z) = \left(\left\{D_1(z)\right\}_{z \in \mathcal{Z}}, \dots, \left\{D_J(z)\right\}_{z \in \mathcal{Z}}\right).$$

The potential outcomes associated with each of the $J + 1$ states are

$$Y_j = \mu_j(X, U_j), \quad j = 0, \dots, J.$$

[170] Heckman and Smith (1999) analyze the determinants of program participation for a job training program.

[171] Merely being interested in a program, such as an HIV treatment program, may signal information that affects certain outcomes prior to receiving any treatment. We ignore these effects, but can easily accommodate them by making application a stage of the program.

Y_0 is the no treatment state, and the Y_j, $j \geqslant 1$, correspond to outcomes associated with dropping out at various stages of the program. In the absence of randomization, the observed Y is

$$Y = \sum_{j=0}^{J} D_j Y_j,$$

the Roy–Quandt switching regime model. Let $\tilde{Y} = (Y_0, \ldots, Y_J)$ denote the vector of potential outcomes associated with all phases of the program. Through selection, the Y_j for persons with $D_j = 1$ may be different from the Y_j for persons with $D_j = 0$.

Appendix B of Chapter 70 gives conditions under which the distributions of the Y_j and the subjective evaluations R_j, $j = 0, \ldots, J$, that generate choices D_j are identified. Using the tools for multiple outcome models developed in Section 7, we can use IV and our extensions of IV to identify the treatment parameters discussed there.

In this subsection, we consider what randomizations at various stages identify. We assume that the randomizations do not disturb the program. Thus we invoke assumption (PI-3). Recall that we also assume absence of general equilibrium effects (PI-4). Let $\xi_j = 1$ denote whether the person is eligible to move beyond stage j. $\xi_j = 0$ means the person is randomized out of the program after completing stage j. A randomization at stage j with $\xi_j = 1$ means the person is allowed to continue on to stage $j + 1$, although the agent may still choose not to. We set $\xi_J \equiv 1$ to simplify the notation. The ξ_j are ordered in a natural way: $\xi_j = 1$ only if $\xi_\ell = 1$, $\ell = 0, \ldots, j - 1$. Array the ξ_j into a vector ξ and denote its support by $\tilde{\xi}$.

Because of agent self-selection, a person who does not choose to participate at stage j cannot be forced to do so. For a person who would choose k ($D_k = 1$) in a nonexperimental environment, Y_k is observed if $\prod_{\ell=0}^{k-1} \xi_\ell = 1$. Otherwise, if $\xi_{k-1} = 0$ but, say, $\prod_{\ell=0}^{k'-1} \xi_\ell = 1$ and $\prod_{\ell=0}^{k'} \xi_\ell = 0$ for $k' < k$, we observe $Y_{k'}$ for the agent. From an experiment with randomization administered at different stages, we observe

$$Y = \sum_{j=0}^{J} D_j \left(\sum_{k=0}^{j} \left(\prod_{\ell=0}^{k-1} \xi_\ell \right) (1 - \xi_k) Y_k \right).$$

To understand this formula, consider a program with three stages ($J = 3$) after the initial participation stage. For a person who would like to complete the program ($D_3 = 1$), but is stopped by randomization after stage 2, we observe Y_2 instead of Y_3. If the person is randomized out after stage 1, we observe Y_1 instead of Y_3.[172]

[172] A more descriptively accurate but notationally cumbersome framework would disaggregate the participation decision and would also recognize that enrolling in a program stage is different from completing it. Thus, let $D_{\mathrm{R}} = 1$ if a person is recruited, $D_{\mathrm{R}} = 0$ if not; $D_{\mathrm{A}} = 1$ if a person applies, $D_{\mathrm{A}} = 0$ if not; $D_{\mathrm{Acc}} = 1$ if a person is accepted, $D_{\mathrm{Acc}} = 0$ if not; $D_{1\mathrm{e}} = 1$ if a person enrolls in stage 1, $D_{1\mathrm{e}} = 0$ if not; $D_{1\mathrm{c}} = 1$ if a person completes stage 1, $D_{1\mathrm{c}} = 0$ if not; and so forth up to $D_{J\mathrm{e}} = 1$ or 0; $D_{J\mathrm{c}} = 1$ or 0. Associated with completing stage ℓ but no later stage is Y_ℓ, $\ell \in \{\mathrm{R}, \mathrm{A}, \mathrm{Acc}, 1\mathrm{e}, 1\mathrm{c}, \ldots, J\mathrm{e}, J\mathrm{c}\}$. Information can be revealed

Let A_k be the indicator that we observe the agent with a stage k outcome. This can happen if a person would have chosen to stop at stage k ($D_k = 1$) and survives randomization through k ($\prod_{\ell=0}^{k-1} \xi_\ell = 1$), or if a person would have chosen to stop at a stage later than k but was thwarted from doing so by the randomization and settles for the best attainable state given the constraint imposed by the randomization. We can express A_k as

$$A_k = D_k \prod_{\ell=0}^{k-1} \xi_\ell + \sum_{j \geqslant k} D_j \left(\prod_{\ell=0}^{k-1} \xi_\ell \right) (1 - \xi_k), \quad k = 1, \ldots, J.$$

If a person who chooses $D_k = 1$ survives all stages of randomization through $k - 1$ and hence is allowed to transit to k, we observe Y_k for that person. For persons who would choose $D_j = 1$, $j > k$, but get randomized out at k, i.e., $(\prod_{\ell=0}^{k-1} \xi_\ell)(1 - \xi_k) = 1$, we also observe Y_k.[173]

We now state the conditions under which sequential randomizations are instrumental variables for the A_j. Let $A_i(z, e_i)$ be the value of A_i when $Z = z$ and $\xi_i = e_i$. Array the A_i, $i = 1, \ldots, J$, into a vector

$$A(z, e) = \big(A_1(z, e_1), A_2(z, e_2), \ldots, A_J(z, e_J)\big).$$

A variety of randomization mechanisms are possible in which randomization depends on the information known to the randomizer at each stage of the program.

IV conditions for ξ are satisfied under the following sequential randomization assumptions. They parallel the sequential randomization conditions developed in the dynamic models analyzed by Abbring and Heckman (Chapter 72) of our contribution:

(e-3a) $\xi_i \perp\!\!\!\perp (\tilde{Y}, \{A(z, e)\}_{(z,e) \in \mathcal{Z} \times \tilde{\xi}}) \mid X, Z, D_\ell = 1$ *for* $\ell < i$, $\prod_{\ell=0}^{i-1} \xi_\ell = 1)$, *for* $i = 1, \ldots, J$,[174]

and

(e-3b) $\Pr(A_i = 1 \mid X, Z, D_\ell = 1$ for $\ell < i, \xi_i, \prod_{\ell=0}^{i-1} \xi_\ell = 1)$ *depends on* ξ_i, *for* $i = 1, \ldots, J$.

at stage ℓ. Observed Y is

$$Y = \sum_{\ell \in \{\mathrm{R}, \mathrm{A}, \mathrm{Acc}, 1\mathrm{e}, 1\mathrm{c}, \ldots, J\mathrm{e}, J\mathrm{c}\}} D_\ell Y_\ell.$$

Randomization can be administered at any stage. We write $\xi_\ell = 0$ if $\prod_{j=1}^{\ell-1} \xi_j = 1$ and a person is randomized out at stage ℓ.

[173] Assumption (PI-3) is crucial in justifying this formula. If randomization alters agent choice behavior, persons who would choose j but get randomized out at k, $k < j$, might change their valuations and decision rule (i.e., there may be randomization bias).

[174] The special case where $\xi_J \equiv 1$ satisfies (e-3a), because in that case ξ_i is a constant.

These expressions assume that the components of $\tilde{Y} = (Y_0, \ldots, Y_J)$ that are realized are known to the randomizer after the dropout decision is made, and thus cannot enter the conditioning set for the sequential randomizations.

To fix ideas, consider a randomization of eligibility ξ_0, setting $\xi_1 = \cdots = \xi_J = 1$. This is a randomization that makes people eligible for participation at all stages of the program. We investigate what this randomization identifies, assuming invariance conditions (PI-3) and (PI-4) hold. For those declared eligible,

$$E(Y \mid \xi_0 = 1) = \sum_{j=0}^{J} E(Y_j \mid D_j = 1) \Pr(D_j = 1). \tag{9.4}$$

For those declared ineligible,

$$E(Y \mid \xi_0 = 0) = \sum_{j=0}^{J} E(Y_0 \mid D_j = 1) \Pr(D_j = 1), \tag{9.5}$$

since agents cannot participate in any stage of the program and are all in the state “0” with outcome Y_0. From observed choice behavior, we can identify each of the components of (9.4). We observe $\Pr(D_j = 1)$ from observed choices of treatment, and we observe $E(Y_j \mid D_j = 1)$ from observed outcomes for each treatment choice. Except for the choice probabilities ($\Pr(D_j = 1)$, $j = 0, \ldots, J$) and $E(Y_0 \mid D_0 = 1)$, we cannot identify individual components of (9.5) for $J > 1$. When $J = 1$, we can identify all of the components of (9.5). The individual components of (9.5) cannot, without further assumptions, be identified by the experiment, although the sum can be. Comparing the treatment group with the control group, we obtain the “intention to treat” parameter with respect to the randomization of ξ_0 alone, setting $\xi_1 = \cdots = \xi_J = 1$ for anyone for whom $\xi_0 = 1$,

$$E(Y \mid \xi_0 = 1) - E(Y \mid \xi_0 = 0) = \sum_{j=1}^{J} E(Y_j - Y_0 \mid D_j = 1) \Pr(D_j = 1). \tag{9.6}$$

For $J > 1$, this simple experimental estimator does not identify the effect of full participation in the program for those who participate ($E(Y_J - Y_0 \mid D_J = 1)$) unless additional assumptions are invoked, such as the assumption that partial participation has the same mean effect as full participation for persons who drop out at the early stages, i.e., $E(Y_j - Y_0 \mid D_j = 1) = E(Y_J - Y_0 \mid D_j = 1)$ for all j. This assumption might be appropriate if just getting into the program is all that matters – a pure signaling effect.

A second set of conditions for identification of this parameter is that $E(Y_j - Y_0 \mid D_j = 1) = 0$ for all $j < J$. Under those conditions, if we divide the mean difference by $\Pr(D_J = 1)$, we obtain the “Bloom” estimator [Mallar, Kerachsky and Thorton (1980), Bloom (1984)]

$$\mathrm{IV}_{\mathrm{Bloom}} = \frac{E(Y \mid \xi_0 = 1) - E(Y \mid \xi_0 = 0)}{\Pr(D_J = 1)},$$

assuming $\Pr(D_J = 1) \neq 0$. This is an IV estimator using ξ_0 as the instrument for A_J. In general, the mean difference between the treated and the controlled identifies only the composite term shown in (9.6). In this case, the simple randomization estimator identifies a not-so-simple or easily interpreted parameter.

More generally, if we randomize persons out after completing stage k ($[\prod_{\ell=0}^{k-1} \xi_\ell](1 - \xi_k) = 1$) and for another group establish full eligibility at all stages ($\prod_{\ell=0}^{J} \xi_\ell = 1$), we obtain

$$E\left[Y \,\middle|\, \prod_{\ell=0}^{J} \xi_\ell = 1\right] - E\left[Y \,\middle|\, \left(\prod_{\ell=0}^{k-1} \xi_\ell\right)(1 - \xi_k) = 1\right]$$
$$= \sum_{j=k}^{J} E(Y_j - Y_k \mid D_j = 1) \Pr(D_j = 1),$$

and hence, since we know $E(Y_k \mid D_k = 1)$ and $\Pr(D_k = 1)$ from observational data, we can identify the combination of parameters

$$\sum_{j=k+1}^{J} E(Y_k \mid D_j = 1) \Pr(D_j = 1), \tag{9.7}$$

for each randomization that stops persons from advancing beyond level k, $k = 0, \ldots, J - 1$.

Observe that a randomization of eligibility that prevents people from going to stage $J - 1$ but not to stage J ($[\prod_{\ell=0}^{J-2} \xi_\ell](1 - \xi_{J-1}) = 1$) identifies $E(Y_J - Y_{J-1} \mid D_J = 1)$:

$$E(Y \mid \xi_0 = 1, \ldots, \xi_{J-2} = 1, \xi_{J-1} = 0)$$
$$= \left[\sum_{j=0}^{J-1} E(Y_j \mid D_j = 1) \Pr(D_j = 1)\right] + E(Y_{J-1} \mid D_J = 1) \Pr(D_J = 1).$$

Thus,

$$E(Y \mid \xi_0 = 1, \ldots, \xi_J = 1) - E(Y \mid \xi_0 = 1, \ldots, \xi_{J-1} = 1, \xi_J = 0)$$
$$= E(Y_J - Y_{J-1} \mid D_J = 1) \Pr(D_J = 1).$$

Since $\Pr(D_J = 1)$ is observed from choice data, as is $E(Y_J \mid D_J = 1)$, we can identify $E(Y_{J-1} \mid D_J = 1)$ from the experiment.

In the general case under assumptions (PI-3) and (PI-4), a randomization that prevents agents from moving beyond stage ℓ ($\xi_0 = 1, \ldots, \xi_{\ell-1} = 1, \xi_\ell = 0$) directly identifies

$$E(Y \mid \xi_0 = 1, \ldots, \xi_{\ell-1} = 1, \xi_\ell = 0)$$
$$= \underbrace{\sum_{j=0}^{\ell} E(Y_j \mid D_j = 1) \Pr(D_j = 1)}_{\text{all components known from observational data}}$$

$$+ \qquad \underbrace{\sum_{j=\ell+1}^{J} E(Y_\ell \mid D_j = 1)\Pr(D_j = 1)}_{\text{sum and probability weights known, but not individual } E(Y_\ell|D_j=1)} \qquad .$$

All of the components of the first set of terms on the right-hand side are known from observational data. The probabilities in the second set of terms are known, but the individual conditional expectations $E(Y_\ell \mid D_j = 1)$, $j = \ell + 1, \dots, J$, are not known without further assumptions.

Randomization at stage ℓ is an IV. To show this, decompose the observed outcome Y into components associated with each value of A_j, the indicator associated with observing a stage j outcome:

$$Y = \sum_{j=0}^{J} A_j Y_j.$$

We can interpret ξ_ℓ as an instrument for A_ℓ. Keeping the conditioning on X, Z implicit, we obtain

$$\begin{aligned}\mathrm{IV}_{\xi_\ell} &= \frac{E[Y \mid \xi_\ell = 0] - E[Y \mid \xi_\ell = 1]}{\Pr(A_\ell = 1 \mid \xi_\ell = 0) - \Pr(A_\ell = 1 \mid \xi_\ell = 1)} \\ &= \frac{\sum_{j=\ell+1}^{J} E[Y_\ell - Y_j \mid D_j = 1]\Pr(D_j = 1)}{\sum_{j=\ell+1}^{J}\Pr(D_j = 1)}, \qquad \ell = 0, \dots, J-1.\end{aligned}$$

By the preceding analysis, we know the numerator term but not the individual components. We know the denominator from choices measured in observational data and invariance assumption (PI-3). The IV formalism is less helpful in the general case.

Table 13 summarizes the parameters or combinations of parameters that can be identified from randomizations performed at different stages. It presents the array of factual and counterfactual conditional mean outcomes $E(Y_j \mid D_\ell = 1)$, $j = 0, \dots, J$ and $\ell = 0, \dots, J$. The conditional mean outcomes obtained from observational data are on the diagonal of the table ($E(Y_j \mid D_j = 1)$, $j = 0, \dots, J$). Because of choices of agents, experiments do not directly identify the elements in the table that are above the diagonal. Under assumptions (PI-3) and (PI-4), experiments described at the base of the table identify the *combinations* of the parameters below the diagonal. Recall that if $\xi_\ell = 0$, the agent cannot advance beyond stage ℓ.[175] If we randomly deny eligibility to move to J ($\xi_{J-1} = 0$), we point identify $E(Y_{J-1} \mid D_J = 1)$, as well as the parameters that can be obtained from observational data. In general, we can only identify the combinations of parameters shown at the base of the table. Following Balke and Pearl (1997), Manski (1989, 1990, 1996, 2003), and Robins (1989), we can use the identified combinations

[175] This definition of ξ_ℓ assumes that $\xi_0 = \dots = \xi_{\ell-1} = 1$.

Table 13

Parameters and combinations of parameters that can be identified by different randomizations

Choice probabilities (known)	Choice	Outcome						
		Y_0	Y_1	$\cdots$	Y_j	$\cdots$	Y_{J-1}	Y_J
$\Pr(D_0 = 1)$	D_0	$E(Y_0 \mid D_0 = 1)$	$E(Y_1 \mid D_0 = 1)$	$\cdots$	$E(Y_j \mid D_0 = 1)$	$\cdots$	$E(Y_{J-1} \mid D_0 = 1)$	$E(Y_J \mid D_0 = 1)$
$\Pr(D_1 = 1)$	D_1	$E(Y_0 \mid D_1 = 1)$	$E(Y_1 \mid D_1 = 1)$	$\cdots$	$E(Y_j \mid D_1 = 1)$	$\cdots$	$E(Y_{J-1} \mid D_1 = 1)$	$E(Y_J \mid D_1 = 1)$
$\Pr(D_2 = 1)$	D_2	$E(Y_0 \mid D_2 = 1)$	$E(Y_1 \mid D_2 = 1)$	$\cdots$	$E(Y_j \mid D_2 = 1)$	$\cdots$	$E(Y_{J-1} \mid D_2 = 1)$	$E(Y_J \mid D_2 = 1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\Pr(D_j = 1)$	D_j	$E(Y_0 \mid D_j = 1)$	$E(Y_1 \mid D_j = 1)$	$\cdots$	$E(Y_j \mid D_j = 1)$	$\cdots$	$E(Y_{J-1} \mid D_j = 1)$	$E(Y_J \mid D_j = 1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\Pr(D_{J-1} = 1)$	D_{J-1}	$E(Y_0 \mid D_{J-1} = 1)$	$E(Y_1 \mid D_{J-1} = 1)$	$\cdots$	$E(Y_j \mid D_{J-1} = 1)$	$\cdots$	$E(Y_{J-1} \mid D_{J-1} = 1)$	$E(Y_J \mid D_{J-1} = 1)$
$\Pr(D_J = 1)$	D_J	$E(Y_0 \mid D_J = 1)$	$E(Y_1 \mid D_J = 1)$	$\cdots$	$E(Y_j \mid D_J = 1)$	$\cdots$	$E(Y_{J-1} \mid D_J = 1)$	$E(Y_J \mid D_J = 1)$
Randomization		$\xi_0 = 0$	$\xi_1 = 0$	$\cdots$	$\xi_j = 0$	$\cdots$	$\xi_{J-1} = 0$	$\xi_J = 0$
New identified combinations of parameters		$\sum_{\ell=1}^{J}\{E(Y_0 \mid D_\ell = 1) \times \Pr(D_\ell = 1)\}$	$\sum_{\ell=2}^{J}\{E(Y_1 \mid D_\ell = 1) \times \Pr(D_\ell = 1)\}$	$\cdots$	$\sum_{\ell=j+1}^{J}\{E(Y_j \mid D_\ell = 1) \times \Pr(D_\ell = 1)\}$	$\cdots$	$E(Y_{J-1} \mid D_J = 1)$	

from different randomizations to bound the admissible values of counterfactuals below the diagonal of Table 13.

Heckman, Smith and Taber (1998) present a test for a strengthened version of the identifying assumptions made by Bloom.[176] They perform a sensitivity analysis to analyze departures from the assumption that dropouts have the same outcomes as nonparticipants. Hotz, Mullin and Sanders (1997) apply the Manski bounds in carefully executed empirical examples and show the difficulties involved in using the Bloom estimator in experiments with multiple outcomes. We next turn to some evidence on the importance of randomization bias.

9.6. *Evidence on randomization bias*

Violations of assumption (PI-3) in the general case with essential heterogeneity affect the interpretation of the outputs of social experiments. They are manifestations of a more general problem termed "Hawthorne effects" that arise from observing any population [see Campbell and Stanley (1963), Cook and Campbell (1979)]. How important is this theoretical possibility in practice? Surprisingly, very little is known about the answer to this question for the social experiments conducted in economics. This is so because randomized social experimentation has usually only been implemented on "pilot projects" or "demonstration projects" designed to evaluate new programs never previously estimated. Disruption by randomization cannot be confirmed or denied using data from these experiments. In one ongoing program evaluated by randomization by the Manpower Demonstration Research Corporation (MDRC), participation was compulsory for the target population [Doolittle and Traeger (1990)]. Hence randomization did not affect applicant pools or assessments of applicant eligibility by program administrators.

There is some information on the importance of randomization, although it is indirect. In the 1980s, the US Department of Labor financed a large-scale experimental evaluation of the ongoing, large-scale manpower training program authorized under the Job Training Partnership Act (JTPA). A study by Doolittle and Traeger (1990) gives some indirect information from which it is possible to determine whether randomization bias was present in an ongoing program.[177] Job training in the United States is organized through geographically decentralized centers. These centers receive incentive payments for placing unemployed persons and persons on welfare in "high-paying" jobs. The participation of centers in the experiment was not compulsory. Funds were set aside to compensate job centers for the administrative costs of participating in the experiment. The funds set aside range from 5 percent to 10 percent of the total operating costs of the centers.

In attempting to enroll geographically dispersed sites, MDRC experienced a training center refusal rate in excess of 90 percent. The reasons for refusal to participate are

[176] They show how to test Bloom's identifying assumption when it is made for distributions rather than just means.

[177] Hotz (1992) summarizes and extends their discussion.

Table 14
Percentage of local JTPA agencies citing specific concerns about participating in the experiment

Concern	Percentage of training centers citing the concern
1. Ethical and public relations implications of:	
a. Random assignment in social programs	61.8
b. Denial of services to controls	54.4
2. Potential negative effect of creation of a control group on achievement of client recruitment goals	47.8
3. Potential negative impact on performance standards	25.4
4. Implementation of the study when service providers do intake	21.1
5. Objections of service providers to the study	17.5
6. Potential staff administrative burden	16.2
7. Possible lack of support by elected officials	15.8
8. Legality of random assignment and possible grievances	14.5
9. Procedures for providing controls with referrals to other services	14.0
10. Special recruitment problems for out-of-school youth	10.5
Sample size: 228	

Notes: Concerns noted by fewer than 5 percent of the training centers are not listed. Percentages add up to more than 100.0 because training centers could raise more than one concern.

Source: Based on responses of 228 local JTPA agencies contacted about possible participation in the National JTPA Study.

Source: Heckman (1992), based on Doolittle and Traeger (1990).

given in Table 14. (The reasons stated there are not mutually exclusive.) Leading the list are ethical and public relations objections to randomization. Major fears (items 2 and 3) were expressed about the effects of randomization on the quality of applicant pool, which would impede the profitability of the training centers. By randomizing, the centers had to widen the available pool of persons deemed eligible, and there was great concern about the effects of this widening on applicant quality – precisely the behavior ruled out by assumptions (PI-3) and (PI-4). In attempting to entice centers to participate, MDRC had to reduce the randomized rejection probability from $\frac{1}{2}$ to as low as $\frac{1}{6}$ for certain centers. The resulting reduction in the size of the control group impairs the power of statistical tests designed to test the null hypothesis of no program effect. Compensation for participation was expanded sevenfold in order to get any centers to participate in the experiment. The MDRC analysts conclude:

> *Implementing a complex random assignment research design in an ongoing program providing a variety of services does inevitably change its operation in some ways. The most likely difference arising from a random assignment field study of program impacts is a change in the mix of clients served. Expanded recruitment efforts, needed to generate the control group, draw in additional applicants who are not identical to the people previously served. A second likely change is that the*

> *treatment categories may somewhat restrict program staff's flexibility to change service recommendations* [Doolittle and Traeger (1990), p. 121].

These authors go on to note that

> *Some* [*training centers*] *because of severe recruitment problems or up-front services cannot implement the type of random assignment model needed to answer the various impact questions without major changes in procedures* [Doolittle and Traeger (1990), p. 123].

This indirect evidence is hardly decisive even about the JTPA experiment, much less all experiments. Training centers may offer these arguments only as a means of avoiding administrative scrutiny, and there may be no "real" effect of randomization. During the JTPA experiment conducted at Corpus Christi, Texas, center administrators successfully petitioned the government of Texas for a waiver of its performance standards on the ground that the experiment disrupted center operations. Self-selection likely guarantees that participant sites are the least likely sites to suffer disruption. Such a selective participation in the experiment calls into question the validity of experimental estimates as a statement about the JTPA system as a whole, as it clearly poses a threat to external validity – problem P-2 as defined in Chapter 70. Torp et al. (1993) report similar problems in a randomized evaluation of a job training program in Norway.

Kramer and Shapiro (1984) note that subjects in drug trials were less likely to participate in randomized trials than in nonexperimental studies. They discuss one study of drugs administered to children afflicted with a disease. The study had two components. The nonexperimental phase of the study had a 4 percent refusal rate, while 34 percent of a subsample of the same parents refused to participate in a randomized subtrial, although the treatments were equally nonthreatening.

These authors cite further evidence suggesting that refusal to participate in randomization schemes is selective. In a study of treatment of adults with cirrhosis, no effect of the treatment was found for participants in a randomized trial. But the death rates for those randomized out of the treatment were substantially lower than among those individuals who refused to participate in the experiment, despite the fact that both groups were administered the same alternative treatment. Part of any convincing identification strategy by randomization requires that the agent document the absence of randomization bias. We next consider some evidence on the importance of dropping out and noncompliance with experimental protocols.

9.7. Evidence on dropping out and substitution bias

Dropouts are a feature of all social programs. Randomization may raise dropout rates, but the evidence for such effects is weak.[178] In addition, most social programs have good substitutes, so that the estimated effect of a program as typically estimated has to be

[178] See Heckman, LaLonde and Smith (1999).

defined relative to the full range of substitute activities in which nonparticipants engage. Experiments exacerbate this problem by creating a pool of persons who attempt to take training who then flock to substitute programs when they are placed in an experimental control group ($\xi = 0$ in the simple randomization analyzed in Sections 9.1–9.4).

Table 15 [reproduced from Heckman et al. (2000)] demonstrates the practical importance of both dropout and substitution bias in experimental evaluations. It reports the rates of treatment group dropout and control group substitution from a variety of social experiments. It reveals that the fraction of treatment group members receiving program services is often less than 0.7, and sometimes less than 0.5. Furthermore, the observed characteristics of the treatment group members who drop out often differ from those who remain and receive the program services.[179] With regard to substitution bias, Table 15 shows that as many as 40% of the controls in some experiments received substitute services elsewhere. In a simple one treatment experiment with full compliance ($\xi = 1 \Rightarrow A = 1$ and $\xi = 0 \Rightarrow A = 0$), all individuals assigned to the treatment group receive the treatment and there is no control group substitution, so that the difference between the fraction of treatments and controls that receive the treatment equals 1.0. In practice, this difference is often well below 1.0. Randomization reduced and delayed receipt of training in the experimental control group but by no means eliminated it. Many of the treatment group members received no treatment.

The extent of both substitution and dropout depends on the characteristics of the treatment being evaluated and the local program environment. In the NSW study, where the treatment was relatively unique and of high enough quality to be clearly perceived as valuable by participants, dropout and substitution rates were low enough to approximate the ideal case. In contrast, for the NJS and for other programs that provide low cost services widely available from other sources, substitution and dropout rates are high. In the NJS, the substitution problem is accentuated by the fact that the program relied on outside vendors to provide most of its training. Many of these vendors, such as community colleges, provided the same training to the general public, often with subsidies from other government programs such as Pell Grants. In addition, in order to help in recruiting sites to participate in the NJS, evaluators allowed them to provide control group members with a list of alternative training providers in the community. Of the 16 sites in the NJS, 14 took advantage of this opportunity to alert control group members to substitute training opportunities.

There are counterpart findings in the application of randomized clinical trials. For example, Palca (1989) notes that AIDS patients denied potentially life-saving drugs took steps to undo random assignment. Patients had the pills they were taking tested to see if they were getting a placebo or an unsatisfactory treatment, and were likely to drop out of the experiment in either case or to seek more effective medication, or both. In the MDRC experiment, in some sites qualified trainees found alternative avenues for securing exactly the same training presented by the same subcontractors by using other

[179] For the NSW shown in this table, see LaLonde (1984). For the NJS data, see Smith (1992).

Table 15
Fraction of experimental treatment and control groups receiving services in experimental evaluations of employment and training programs

Study	Authors/time period	Target group(s)	Fraction of treatments receiving services	Fraction of controls receiving services
1. NSW	Hollister et al. (1984)	Long-term AFDC women	0.95	0.11
	(9 months after RA)	Ex-addicts	NA	0.03
		17–20 year old high school dropouts	NA	0.04
2. SWIM	Friedlander and Hamilton (1993)	AFDC women: applicants and recipients		
	(Time period not reported)	a. Job search assistance	0.54	0.01
		b. Work experience	0.21	0.01
		c. Classroom training/OJT	0.39	0.21
		d. Any activity	0.69	0.30
		AFDC-U unemployed fathers		
		a. Job search assistance	0.60	0.01
		b. Work experience	0.21	0.01
		c. Classroom training/OJT	0.34	0.22
		d. Any activity	0.70	0.23
3. JOBSTART	Cave et al. (1993) (12 months after RA)	Youth high school dropouts		
		Classroom training/OJT	0.90	0.26
4. Project Independence	Kemple et al. (1995) (24 months after RA)	AFDC women: applicants and recipients		
		a. Job search assistance	0.43	0.19
		b. Classroom training/OJT	0.42	0.31
		c. Any activity	0.64	0.40
5. New chance	Quint et al. (1994)	Teenage single mothers		
	(18 months after RA)	Any education services	0.82	0.48
		Any training services	0.26	0.15
		Any education or training	0.87	0.55
6. National JTPA Study	Heckman and Smith (1998)	Self-reported from survey data		
	(18 months after RA)	Adult males	0.38	0.24
		Adult females	0.51	0.33
		Male youth	0.50	0.32
		Female youth	0.81	0.42
	Combined Administrative Survey Data			
		Adult males	0.74	0.25
		Adult females	0.78	0.34
		Male youth	0.81	0.34
		Female youth	0.81	0.42

(*Continued on next page*)

Table 15
(*Continued*)

Notes: RA = random assignment. H.S. = high school. AFDC = Aid to Families with Dependent Children. OJT=On the Job Training.

Service receipt includes any employment and training services. The services received by the controls in the NSW study are CETA and WIN jobs. For the Long Term AFDC Women, this measure also includes regular public sector employment during the period.

Sources for data: Maynard and Brown (1980), p. 169, Table A14; Masters and Maynard (1981), p. 148, Table A.15; Friedlander and Hamilton (1993), p. 22, Table 3.1; Cave et al. (1993), p. 95, Table 4.1; Quint et al. (1994), p. 110, Table 4.9; and Kemple et al. (1995), p. 58, Table 3.5; Heckman and Smith (1998) and calculations by the authors.

Source: Heckman, LaLonde and Smith (1999) and Heckman et al. (2000).

methods of financial support. Heckman, LaLonde and Smith (1999) discuss a variety of other problems that sometimes plague social experiments.

Our discussion up to this point has focused on point identification of parameters over the empirical supports. A large and emerging literature produces bounds on the parameters and distributions when point identification is not possible. We now consider bounds on the parameters within the framework of economic models of choice and the MTE.

10. Bounding and sensitivity analysis

Thus far we have assumed full support conditions and have presented conditions for identification over those supports. We now consider partial identification in the context of the MTE framework. We return to the two-outcome model to develop the basic approach in a simpler setting.

The central evaluation problem is that we observe the distribution of $(Y, D, X, Z) = (DY_1 + (1-D)Y_0, D, X, Z)$, but do not observe the distribution of all of the components that comprise it (Y_1, Y_0, D, X, Z). Let η denote a distribution for (Y_1, Y_0, D, X, Z), and let it be known that η belongs to the class $\mathcal{H} \subset \mathcal{F}$, where $\mathcal{F}$ is the space of all probability distributions on (Y_1, Y_0, D, X, Z). Let P_η denote the resulting distribution of $(DY_1 + (1 - D)Y_0, D, X, Z)$ if η is the distribution for (Y_1, Y_0, D, X, Z). Let η^0 and P_{η^0} denote the corresponding true distributions. Knowledge of the distribution of $(DY_1 + (1-D)Y_0, D, X, Z)$ allows us to infer that η lies in the set $\{\eta \in \mathcal{H}\colon\ P_\eta = P_{\eta^0}\}$. All elements of $\{\eta \in \mathcal{H}\colon\ P_\eta = P_{\eta^0}\}$ are consistent with the true distribution of the observed data.

Let $\mathcal{H}^0 = \{\eta \in \mathcal{H}\colon\ P_\eta = P_{\eta^0}\}$. Let E_η denote expectation with respect to the measure η, i.e., $E_\eta(A) = \int A\, d\eta$, so that $E(A) = E_{\eta^0}(A)$. Consider inference for ATE, $E(Y_1 - Y_0)$. Knowledge of the distribution of the observed variables allows us to infer that

$$E(Y_1 - Y_0) \in \bigl\{E_\eta(Y_1 - Y_0)\colon\ \eta \in \mathcal{H}^0\bigr\}.$$

The identification analyses of the previous sections proceed by imposing sufficient restrictions on $\mathcal{H}$ such that $\{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}^0\}$ contains only one element and thus $E(Y_1 - Y_0)$ is point identified. Bounding analysis proceeds by finding a set $\mathcal{B}$ such that $\mathcal{B} \supseteq \{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}^0\}$.[180] One goal of bounding analysis is to construct $\mathcal{B}$ such that $\mathcal{B} = \{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}^0\}$ in which case the bounds are said to be *sharp*. If the bounds are sharp, then the bounds exploit all information and no smaller bounds can be constructed without imposing additional structure. In contrast, if $\{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}^0\}$ is a proper subset of $\mathcal{B}$, then smaller bounds can be constructed. In every example we consider, the set $\{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}^0\}$ is a closed interval, so that $\{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}^0\} = [\inf_{\eta \in \mathcal{H}^0} E_\eta(Y_1 - Y_0), \sup_{\eta \in \mathcal{H}^0} E_\eta(Y_1 - Y_0)]$.

Sensitivity analysis is a commonly used procedure. It varies the parameters fixed in a model and determines the sensitivity of estimates to the perturbations of the parameter. Sensitivity analysis is formally equivalent to bounding. In particular, in sensitivity analysis, one parameterizes η and then constructs bounds based on letting the parameters vary over some set.[181] Parameterize η as $\eta(\theta)$ for some parameter vector $\theta \in \Theta$, and let θ^0 be the "true" parameter value so that $\eta^0 = \eta(\theta^0)$. θ is typically finite-dimensional, though it need not be. Let $\Theta^0 = \{\theta \in \Theta\colon P_{\eta(\theta)} = P_{\eta(\theta^0)}\}$. If θ is point identified given the observed variables, then Θ^0 will contain only one element, but if not all parameters are identified given the observed data then Θ will contain more than one element. Consider

$$\left\{E_{\eta(\theta)}(Y_1 - Y_0)\colon \theta \in \Theta^0\right\}.$$

This can trivially be seen as a special case of bounding analysis by taking $\mathcal{H} = \{\eta(\theta)\colon \theta \in \Theta\}$ and $\mathcal{H}^0 = \{\eta(\theta)\colon \theta \in \Theta^0\}$. Likewise, by taking a proper parameterization, any bounding analysis can be seen as a special case of sensitivity analysis.

We consider bounds on ATE. The corresponding bounds on treatment on the treated follow with trivial modifications.[182] We focus on bounds that exploit instrumental variable type assumptions or latent index assumptions, and we do not attempt to survey the entire literature on bounds.[183] We begin by describing the bounds that only assume that the outcome variables are bounded. We then consider imposing additional assumptions. We consider imposing the assumption of comparative advantage in the decision

[180] Examples of bounding analysis include Balke and Pearl (1997), Heckman, Smith and Clements (1997), Manski (1989, 1990, 1997, 2003) and Robins (1989).

[181] Examples of sensitivity analysis include Glynn, Laird and Rubin (1986), Smith and Welch (1986), and Rosenbaum (1995).

[182] We do not consider bounds on the joint distribution of (Y_1, Y_0). Identification of the joint distribution of (Y_1, Y_0) is substantially more difficult than identification of the ATE or treatment on the treated (TT). For example, even a perfect randomized experiment does not point identify the joint distribution of (Y_1, Y_0) without further assumptions. See Heckman and Smith (1993), Heckman, Smith and Clements (1997), and Heckman, LaLonde and Smith (1999) for an analysis of this problem.

[183] Surveys of the bounding approach include Manski (1995, 2003). Heckman, LaLonde and Smith (1999) includes an alternative survey of the bounding approach.

rule, then consider instead imposing an instrumental variables type assumption, and conclude by considering the combination of comparative advantage and instrumental variables assumptions. We examine the relative power of these alternative assumptions to narrow the very wide bounds that result from only imposing that the outcome variables are bounded.

10.1. Outcome is bounded

We first consider bounds on $E(Y_1 - Y_0)$ that only assume that the outcomes be bounded. We consider this case as a point of contrast for the later bounds that exploit instrumental variable conditions, and also for the pedagogical purpose of showing the bounding methodology in a simple context. We impose that the outcomes are bounded with probability 1.

ASSUMPTION B *(Outcome is Bounded).* For $j = 0, 1$,

$$\Pr\bigl(y^l \leqslant Y_j \leqslant y^u\bigr) = 1.$$[184]

In our notation, this corresponds to

$$\mathcal{H} = \bigl\{\eta \in \mathcal{F}\colon\ \eta\bigl[y^l \leqslant Y_1 \leqslant y^u\bigr] = 1, \eta\bigl[y^l \leqslant Y_0 \leqslant y^u\bigr] = 1\bigr\}.$$

For example, if Y is an indicator variable, then the bounds are $y^l = 0$ and $y^u = 1$.

Following Manski (1989) and Robins (1989), use the law of iterated expectations to obtain

$$E(Y_1) = \Pr[D = 1]E(Y_1 \mid D = 1) + \bigl(1 - \Pr[D = 1]\bigr)E(Y_1 \mid D = 0),$$
$$E(Y_0) = \Pr[D = 1]E(Y_0 \mid D = 1) + \bigl(1 - \Pr[D = 1]\bigr)E(Y_0 \mid D = 0).$$

$\Pr[D = 1]$, $E(Y_1 \mid D = 1)$, and $E(Y_0 \mid D = 0)$ are identified, while $E(Y_0 \mid D = 1)$ and $E(Y_1 \mid D = 0)$ are bounded by y^l and y^u, so that

$$\begin{aligned}
&\Pr[D = 1]E(Y_1 \mid D = 1) + \bigl(1 - \Pr[D = 1]\bigr)y^l \\
&\quad \leqslant E(Y_1) \leqslant \Pr[D = 1]E(Y_1 \mid D = 1) + \bigl(1 - \Pr[D = 1]\bigr)y^u, \\
&\Pr[D = 1]y^l + \bigl(1 - \Pr[D = 1]\bigr)E(Y_0 \mid D = 0) \\
&\quad \leqslant E(Y_0) \leqslant \Pr[D = 1]y^u + \bigl(1 - \Pr[D = 1]\bigr)E(Y_0 \mid D = 0)
\end{aligned}$$

[184] We assume that Y_1 and Y_0 have the same bounds for ease of exposition. The modifications required to analyze the more general case are straightforward.

and thus

$$\mathcal{B} = \left[B^L, B^U\right],$$

with

$$\begin{aligned} B^L &= \left(\Pr[D=1]E(Y \mid D=1) + \left(1-\Pr[D=1]\right)y^l\right) \\ &\quad - \left(\Pr[D=1]y^u + \left(1-\Pr[D=1]\right)E(Y \mid D=0)\right), \\ B^U &= \left(\Pr[D=1]E(Y \mid D=1) + \left(1-\Pr[D=1]\right)y^u\right) \\ &\quad - \left(\Pr[D=1]y^l + \left(1-\Pr[D=1]\right)E(Y \mid D=0)\right) \end{aligned}$$

with the width of these bounds given by

$$B^U - B^L = y^u - y^l.$$

For example, if $Y = 0, 1$, then the width of the bounds equals 1, $B^U - B^L = 1$.

These bounds are sharp. To show this, for any $M \in [B^L, B^U]$, one can trivially construct a distribution η of (Y_0, Y_1, D) which is consistent with the observed data, consistent with the restriction that the outcomes are bounded, and for which $E_\eta(Y_1 - Y_0) = M$, thus showing that $M \in [B^L, B^U]$. Since this is true for any $M \in [B^L, B^U]$, it follows that $[B^L, B^U] \subseteq \{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}_0\}$. Since we have already shown that $[B^L, B^U]$ are valid bounds, $[B^L, B^U] \supseteq \{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}_0\}$, we conclude that $[B^L, B^U] = \{E_\eta(Y_1 - Y_0)\colon \eta \in \mathcal{H}_0\}$ and thus that the bounds are sharp. This illustrates a common technique towards the construction of sharp bounds: in a first step, construct a natural set of bounds, and in a second step, use a proof by construction to show that the bounds are sharp.

Note the following features of these bounds. First, as noted by Manski (1990), these bounds always include zero. Thus, bounds that only exploit that the outcomes are bounded can never reject the null of zero average treatment effect. The bounds themselves depend on the data, but the width of the bounds, $B^U - B^L = y^u - y^l$, is completely driven by the assumed bounds on Y_1, Y_0. For example, if Y_1 and Y_0 are binary, the width of the bounds is always 1.

10.2. Latent index model: Roy model

The bounds that only impose that the outcomes are bounded are typically very wide, never provide point identification, and can never reject the null of zero average treatment effect. This lack of identifying power raises the question of whether one can impose additional structure to narrow the bounds. The central issue with bounding analysis is to explore the trade-off between assumptions and width of the resulting bounds. In this section, we discuss bounds that follow from maintaining Assumption B, that the outcomes are bounded, but also add the assumption of a Roy model for selection into treatment.[185] Such an assumption substantially narrows the width of the bounds compared to only imposing that the outcomes themselves are bounded, but does not provide point identification.

Again impose Assumption B: the outcomes are bounded. In addition, assume a model of comparative advantage, in particular,

ASSUMPTION RM *(Roy Model).*

$$D = \mathbf{1}[Y_1 \geqslant Y_0]. \tag{10.1}$$

Restriction RM imposes a special case of a latent index model, $D = \mathbf{1}[Y^* \geqslant 0]$ with $Y^* = Y_1 - Y_0$. Using the assumption of a Roy model while maintaining the assumption that the outcomes are bounded, we can narrow the bounds compared to the case where we only imposed that the outcomes are bounded. Peterson (1976) constructs the sharp bounds for the competing risks model, which is formally equivalent to a Roy model. Manski (1995) constructs the same bounds for the Roy model.

Following Peterson (1976) and Manski (1995), we have that

$$\begin{aligned} E[Y_1 \mid D = 1] &= E[Y_1 \mid Y_0 \leqslant Y_1] \\ &\geqslant E[Y_0 \mid Y_0 \leqslant Y_1] \\ &= E[Y_0 \mid D = 1] \end{aligned}$$

and by a parallel argument, $E[Y_0 \mid D = 0] \geqslant E[Y_1 \mid D = 0]$. We thus have upper bounds on $E(Y_0 \mid D = 1)$ and $E(Y_1 \mid D = 0)$. The lower bounds on $E[Y_0 \mid D = 1]$ and $E[Y_1 \mid D = 0]$ are the same as for the bounds that only imposed that the outcomes are bounded. We then have

$$E(Y_1 - Y_0) \in \mathcal{B} \equiv \left[B^L, B^U\right],$$

with

$$\begin{aligned} B^L &= \big(\Pr[D = 1]E(Y \mid D = 1) + \big(1 - \Pr[D = 1]\big)y^l\big) \\ &\quad - \big(\Pr[D = 1]E(Y \mid D = 1) + \big(1 - \Pr[D = 1]\big)E(Y \mid D = 0)\big), \\ B^U &= \big(\Pr[D = 1]E(Y \mid D = 1) + \big(1 - \Pr[D = 1]\big)E(Y \mid D = 0)\big) \\ &\quad - \big(\Pr[D = 1]y^l + \big(1 - \Pr[D = 1]\big)E(Y \mid D = 0)\big), \end{aligned}$$

[185] In contrast to the comparative advantage Roy model, one could instead impose an absolute advantage model as in the bounding analysis of Smith and Welch (1986). They assume that those with $D = 1$ have an absolute advantage over those with $D = 0$ in terms of their Y_1 outcomes: $\frac{1}{2}E(Y_1 \mid D = 1) \leqslant E(Y_1 \mid D = 0) \leqslant E(Y_1 \mid D = 1)$, and use this assumption to bound $E(Y_1)$. In their application, Y_1 is the wage and D is an indicator variable for working, so that there is not a well defined Y_0 variable. However, if one were to adapt their idea of absolute advantage to the treatment effect literature, one could assume, e.g., that $E(Y_0 \mid D = 0) \leqslant E(Y_0 \mid D = 1) \leqslant \frac{3}{2}E(Y_0 \mid D = 0)$ with the bounds on ATE following immediately from these assumptions.

and we can rewrite these bounds as

$$B^L = \bigl(1 - \Pr[D = 1]\bigr)\bigl(y^l - E(Y \mid D = 0)\bigr),$$
$$B^U = \Pr[D = 1]\bigl(E(Y \mid D = 1) - y^l\bigr),$$

with the width of the bounds given by

$$B^U - B^L = E(Y) - y^l.$$

For example, if $Y = 0, 1$, then the width of the bounds is given by $B^U - B^L = \Pr(Y = 1)$. Following an argument similar to that presented in the previous section, one can show that these bounds are sharp.

Note the following features of these bounds. First, the bounds do not involve y^u, and actually the same bounds will hold if we were to weaken the maintained assumption that $\Pr[y^l \leqslant Y_j \leqslant y^u] = 1$ for $j = 0, 1$, to instead only require that $\Pr[y^l \leqslant Y_j] = 1$. The width of the bounds imposing comparative advantage are $E(Y) - y^l$, so that the bounds will never provide point identification (as long as $E(Y) > y^l$). For example, if Y is binary, the width of the bounds is $\Pr[Y = 1]$, the bounds will not provide point identification unless all individuals have $Y = 0$. However, the bounds will always improve upon the bounds that impose only that the outcome is bounded – imposing comparative advantage shrinks the width of the bounds from $y^u - y^l$ to $E(Y) - y^l$, thus shrinking the bounds by an amount equal to $y^u - E(Y)$. For example, if Y is binary, then imposing the bounds shrinks the width of the bounds from 1 to $\Pr[Y = 1]$. Finally, note that the bounds will always include zero, so that imposing comparative advantage does not by itself allow one to ever reject the null of zero average treatment effect.

10.3. Bounds that exploit an instrument

The previous section considered bounds that exploit knowledge of the selection process, in particular that selection is determined by a Roy model. An alternative way to narrow the bounds over simply imposing that the outcome is bounded is to assume access to an instrument. We now discuss bounds with various types of instrumental variables assumptions. We begin with the Manski (1990) analysis for bounds that exploit a mean-independence condition, then consider the Balke and Pearl (1997) analysis for bounds that exploit a full statistical independence condition, and finally conclude with a discussion of Heckman and Vytlacil (1999) who combine an instrumental variable assumption with a nonparametric selection model.

10.3.1. Instrumental variables: Mean independence condition

Again impose Assumption B so that the outcomes are bounded. In addition, following Manski (1990), impose a mean-independence assumption:

ASSUMPTION IV.

$$E(Y_1 \mid Z = z) = E(Y_1),$$
$$E(Y_0 \mid Z = z) = E(Y_0)$$

for $z \in \mathcal{Z}$ where $\mathcal{Z}$ denotes the support of the distribution of Z.

For any $z \in \mathcal{Z}$, following the exact same series of steps as for the bounds that only imposed Assumption B, we have that

$$\begin{aligned} E(DY \mid Z = z) + \bigl(1 - P(z)\bigr)y^l &\leqslant E(Y_1 \mid Z = z) \\ &\leqslant E(DY \mid Z = z) + \bigl(1 - P(z)\bigr)y^u. \end{aligned}$$

By the IV assumption, we have $E(Y_1 \mid Z = z) = E(Y_1)$. Since these bounds hold for any $z \in \mathcal{Z}$, we have

$$\begin{aligned} &\sup_{z \in \mathcal{Z}}\bigl\{E(DY \mid Z = z) + \bigl(1 - P(z)\bigr)y^l\bigr\} \\ &\quad \leqslant E(Y_1) \leqslant \inf_{z \in \mathcal{Z}}\bigl\{E(DY \mid Z = z) + \bigl(1 - P(z)\bigr)y^u\bigr\}. \end{aligned}$$

Applying the same analysis for $E(Y_0)$, we have

$$E(Y_1 - Y_0) \in \mathcal{B} = \bigl[B^L, B^U\bigr],$$

with

$$\begin{aligned} B^L &= \sup_{z \in \mathcal{Z}}\bigl\{E(DY \mid Z = z) + \bigl(1 - P(z)\bigr)y^l\bigr\} \\ &\quad - \inf_{z \in \mathcal{Z}}\bigl\{E\bigl((1 - D)Y \bigm| Z = z\bigr) + P(z)y^u\bigr\}, \\ B^U &= \inf_{z \in \mathcal{Z}}\bigl\{E(DY \mid Z = z) + \bigl(1 - P(z)\bigr)y^u\bigr\} \\ &\quad - \sup_{z \in \mathcal{Z}}\bigl\{E\bigl((1 - D)Y \bigm| Z = z\bigr) + P(z)y^l\bigr\}. \end{aligned}$$

As discussed by Manski (1994), these bounds are sharp under the mean-independence condition.[186] As noted by Manski (1990), these bounds do not necessarily include zero, so that it may be possible to use the bounds to test the null of zero average treatment effect. Let $p^u = \sup_{z \in \mathcal{Z}} \Pr[D = 1 \mid Z = z]$, $p^l = \inf_{z \in \mathcal{Z}} \Pr[D = 1 \mid Z = z]$. A trivial modification to Corollaries 1 and 2 of Proposition 6 of Manski (1994) shows that

(1) $p^u \geqslant \frac{1}{2}$ and $p^l \geqslant \frac{1}{2}$ is a necessary condition for $B^L = B^U$, i.e., for point identification from the mean independence condition.

[186] See Manski and Pepper (2000) for extensions of these bounds.

(2) If Y_1, Y_0 are independent of D, then the width of the IV-bounds is $((1 - p^u) + p^l)(y^u - y^l)$. Thus, if Y_1, Y_0 are independent of D, the bounds will collapse to point identification if and only if $p^u = 1$, $p^l = 0$.

Note that it is neither necessary nor sufficient for $P(z)$ to be a nontrivial function of z for these bounds to improve upon the bounds that only imposed that the outcome is bounded. Likewise, comparing these bounds to the comparative advantage bounds shows that neither set of bounds will in general be narrower than the other. Finally, note that these bounds are relatively complicated, and to evaluate the bounds and the width of the bounds requires use of $P(z)$, $E(YD \mid Z = z)$, and $E(Y(1 - D) \mid Z = z)$ for all $z \in \mathcal{Z}$.

10.3.2. Instrumental variables: Statistical independence condition

While Manski constructs sharp bounds for mean-independence conditions, Balke and Pearl (1997) construct sharp bounds for the statistical independence condition for the case where Y and Z are binary. Balke and Pearl impose the same independence condition as the Imbens and Angrist (1994) LATE independence condition. In particular, let D_0, D_1 denote the counterfactual choices that would have been made had Z been set exogenously to 0 and 1, respectively, and impose the following assumption.

ASSUMPTION IV-BP.

$$(Y_0, Y_1, D_0, D_1) \perp\!\!\!\perp Z.$$

Note that this strengthens the Manski conditions not only in imposing that potential outcomes are statistically independent of Z instead of mean-independent of Z, but also imposing that the counterfactual choices are independent of Z.

For the case of Z and Y binary, Balke and Pearl manage to transform the problem of constructing sharp bounds into a linear programming problem. Assuming that the identified set is a closed interval, the sharp bounds are by definition $[B^L, B^U]$ with

$$B^L = \inf_{\eta \in \mathcal{H}^0} E_\eta(Y_1 - Y_0),$$
$$B^U = \sup_{\eta \in \mathcal{H}^0} E_\eta(Y_1 - Y_0).$$

In general, the constrained set of distributions, $\eta \in \mathcal{H}^0$, may be high-dimensional and nonconvex. Using the assumption that Z and Y are binary, they transform the problem into the minimization of a linear function over a finite-dimensional vector space subject to a set of linear constraints. The resulting bounds are somewhat complex. For some distributions of the observed data, they will coincide with the Manski mean-independence bounds, but for other distributions of the observed data they will be narrower than the Manski mean-independence bounds. Thus, imposing statistical independence does narrow the bounds over the mean independence bounds.

It is not immediately clear how to generalize the Balke and Pearl analysis to distributions with continuous Z or Y, or how to construct sharp bounds under the statistical independence condition for Z or Y continuous. The appropriate generalization of Balke and Pearl's analysis to a more general setting remains an open question.

10.3.3. Instrumental variables: Nonparametric selection model/LATE conditions

We started with the mean independence version of the instrumental variables condition, and then discussed strengthening the instrumental variables condition to full independence in the special case where Y and Z are binary. The result of shifting from mean independence to full independence is to sometimes reduce the width of the resulting bounds but also to have an even more complicated form for the bounds. We now consider further strengthening the instrumental variables either by imposing a nonparametric selection model for the first stage as in Heckman and Vytlacil (1999) or by imposing instrumental variable conditions of the form considered by Imbens and Angrist (1994). The sharp bounds corresponding to these strengthened versions of instrumental variables do not reduce the bounds compared to imposing a weaker form of the instrumental variables assumption but produces a much simpler form for the bounds.

Let $D(z)$ denote the counterfactual choices that would have been made had Z been set exogenously to z. Consider the LATE independence, rank, and monotonicity conditions (IV-1), (IV-2), (IV-3), respectively, of Imbens and Angrist (1994) presented in Sections 2 and 4.

Note that the LATE monotonicity assumption (IV-3) strengthens assumption [IV-BP]. The LATE independence assumption (IV-1) is exactly the same as assumption [IV-BP] except that the assumption is stated here without requiring Z to be binary. In their context of binary Z and Y, Balke and Pearl discuss the LATE monotonicity condition and show that the LATE monotonicity condition imposes constraints on the observed data which imply that the Balke and Pearl (1997) bounds and the Manski mean-independence bounds will coincide.[187]

Consider the nonparametric selection model of Heckman and Vytlacil (1999):

NONPARAMETRIC SELECTION MODEL S. $D = \mathbf{1}[\mu(Z) \geqslant U]$ and $Z \perp\!\!\!\perp (Y_0, Y_1, U)$. This is a consequence of Equations (3.3) and assumptions (A-1)–(A-5) presented in Section 4.

From Vytlacil (2002), we have that the Imbens and Angrist conditions (IV-1)–(IV-3) are equivalent to imposing a nonparametric selection model of the form S. Thus, the bounds derived under one set of assumptions will be valid under the alternative set of assumptions, and bounds that are sharp under one set will be sharp under the alternative

[187] Robins (1989) constructs the same bounds under the LATE condition for the case of Z and Y binary, though he does not prove that the bounds are sharp.

set of assumptions. This equivalence implies that the Balke and Pearl result also holds for the selection model: if Z and Y are binary, then the sharp bounds under the nonparametric selection model coincide with the sharp bounds under mean independence IV.

We now consider the more general case where neither Z nor Y need be binary. Heckman and Vytlacil (1999) derived bounds on the average treatment effect under the assumptions that the outcomes are generated from a bounded outcome nonparametric selection model for treatment without requiring that Z or Y be binary or any other restrictions on the support of the distributions of Z and Y beyond the assumption that the outcomes are bounded (Assumption B). In particular, they derived the following bounds on the average treatment effect:

$$B^L \leqslant E(Y_1 - Y_0) \leqslant B^U,$$

with

$$\begin{aligned} B^U &= E\big(DY \bigm| P(Z) = p^u\big) + \big(1 - p^u\big)y^u \\ &\quad - E\big((1-D)Y \bigm| P(Z) = p^l\big) - p^l y^l, \\ B^L &= E\big(DY \bigm| P(Z) = p^u\big) + \big(1 - p^u\big)y^l - E\big((1-D)Y \bigm| P(Z) = p^l\big) - p^l y^u. \end{aligned}$$

Note that these bounds do not necessarily include zero. The width of the bounds is

$$B^U - B^L = \big(\big(1 - p^u\big) + p^l\big)\big(y^u - y^l\big).$$

For example, if Y is binary then the width of the bounds is simply $B^U - B^L = ((1 - p^u) + p^l)$. Trivially, $p^u = 1$ and $p^l = 0$ is necessary and sufficient for the bounds to collapse to point identification, with the width of the bounds linearly related to the distance between p^u and 1 and the distance between p^l and 0. Note that it is necessary and sufficient for $P(z)$ to be a nontrivial function of z for these bounds to improve upon the bounds that only imposed that the outcomes are bounded. Evaluating the width of the bounds only requires p^u, p^l. The only additional information required to evaluate the bounds themselves is $E(DY \mid P(Z) = p^u)$ and $E((1-D)Y \mid P(Z) = p^l)$.

Heckman and Vytlacil (2001a) analyze how these bounds compare to the Manski (1990) mean independence bounds, and analyze whether these bounds are sharp. They show that the selection model imposes restrictions on the observed data such that the Manski (1990) mean independence bounds collapse to the simpler Heckman and Vytlacil (2001a) bounds. In particular, given assumption S, they show that

$$\begin{aligned} &\inf_{z\in\mathcal{Z}}\big\{E(DY \mid Z = z) + \big(1 - P(z)\big)y^u\big\} = E\big(DY \bigm| P(Z) = p^u\big) + \big(1 - p^u\big)y^u, \\ &\sup_{z\in\mathcal{Z}}\big\{E\big((1-D)Y \bigm| Z = z\big) + P(z)y^l\big\} = E\big((1-D)Y \bigm| P(Z) = p^l\big) - p^l y^l \end{aligned}$$

and thus the Manski (1990) upper bound collapses to the Heckman and Vytlacil (1999) upper bound under assumption S. The parallel result holds for the lower bounds. Furthermore, Heckman and Vytlacil (2001a) establish that the Heckman and Vytlacil (1999) bounds are sharp given Assumptions B and S. Thus, somewhat surprisingly, imposing the stronger assumption of the existence of an instrument in a nonparametric selection model does not narrow the bounds compared to the case of imposing only the weaker assumption of mean independence, but does impose structure on the data which substantially simplifies the form of the mean-independence bounds. By the Vytlacil (2002) equivalence result, the same conclusion holds for the LATE assumptions – imposing the LATE assumptions does not narrow the bounds compared to only imposing the weaker assumption of mean independence, but does impose restrictions on the data that substantially simplify the form of the bounds. Vytlacil, Santos and Shaikh (2005) extend these bounds.

10.4. Combining comparative advantage and instrumental variables

We have thus far examined bounds that impose a comparative advantage model, and bounds that exploit an instrumental variables assumption. In general, neither restriction has more identifying power than the other. We now consider combining both types of assumptions.

Assume $D = \mathbf{1}[Y_1 - Y_0 \geqslant C(Z)]$, with Z observed and $Z \perp\!\!\!\perp (Y_0, Y_1)$. This is a Roy model with a cost $C(Z)$ of treatment, with the cost of treatment a function of an "instrument" Z. For ease of exposition, assume that Z is a continuous scalar random variable and that (Y_0, Y_1) are continuous random variables.[188] Also for ease of exposition, assume that $\mathcal{Z}$ (the support of the distribution Z) is compact and that $C(\cdot)$ is a continuous function. These assumptions are only imposed for ease of exposition.

The model is a special case of the nonparametric selection model considered by Heckman and Vytlacil (2001a), but with more structure that we can now exploit. Begin by following steps similar to Heckman and Vytlacil (2001a). Using the fact that $D = \mathbf{1}[Y_1 - Y_0 \geqslant C(Z)]$ and that $Z \perp\!\!\!\perp (Y_0, Y_1)$, we have

$$P(Z) = 1 - F_{Y_1 - Y_0}\big(C(Z)\big),$$

where $F_{Y_1 - Y_0}$ is the distribution function of $Y_1 - Y_0$. Given our assumptions, we have that there will exist z^u and z^l such that

$$\begin{aligned}
&C\big(z^u\big) = \sup\big\{C(z)\colon\ z \in \mathcal{Z}\big\},\\
&P\big(z^u\big) = 1 - F_{Y_1 - Y_0}\big(C\big(z^u\big)\big) = \inf\big\{P(Z)\colon\ z \in \mathcal{Z}\big\},\\
&C\big(z^l\big) = \inf\big\{C(z)\colon\ z \in \mathcal{Z}\big\},\\
&P\big(z^l\big) = 1 - F_{Y_1 - Y_0}\big(C\big(z^l\big)\big) = \sup\big\{P(Z)\colon\ z \in \mathcal{Z}\big\}.
\end{aligned}$$

In other words, $Z = z^u$ is associated with the highest possible cost of treatment and thus the lowest possible conditional probability of $D = 1$, while $Z = z^l$ is associated with the lowest possible cost of treatment and thus the highest possible conditional

[188] More formally, impose that the distribution of Z has a density with respect to Lebesgue measure on $\mathbb{R}$, and assume that (Y_1, Y_0) has a density with respect to Lebesgue measure on $\mathbb{R}^2$.

probability of $D = 1$. Since $P(\cdot)$ for $z \in \mathcal{Z}$ is identified, we have that z^u and z^l are identified.

Consider identification of $C(z)$. Using the model and the independence assumptions, we have

$$\begin{aligned}
&\frac{\partial}{\partial z} E(Y \mid Z = z) \\
&\quad = \frac{\partial}{\partial z} E(YD \mid Z = z) + \frac{\partial}{\partial z} E\big(Y(1 - D) \bigm| Z = z\big) \\
&\quad = \frac{\partial}{\partial z} \int_{C(z)}^{\infty} E(Y_1 \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t) \\
&\qquad + \frac{\partial}{\partial z} \int_{-\infty}^{C(z)} E(Y_0 \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t) \\
&\quad = -\big[E\big(Y_1 \bigm| Y_1 - Y_0 = C(z)\big) - E\big(Y_0 \bigm| Y_1 - Y_0 = C(z)\big)\big] f_{Y_1 - Y_0}\big(C(z)\big) C'(z) \\
&\quad = -C(z) C'(z) f_{Y_1 - Y_0}\big(C(z)\big)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial z} P(z) &= \frac{\partial}{\partial z} \int_{C(z)}^{\infty} dF_{Y_1 - Y_0}(t) \\
&= -C'(z) f_{Y_1 - Y_0}\big(C(z)\big)
\end{aligned}$$

and thus

$$\left[\frac{\partial}{\partial z} E(Y \mid Z = z) \Big/ \frac{\partial}{\partial z} P(z)\right] = C(z)$$

for any $z \in \mathcal{Z}$ such that $\frac{\partial}{\partial z} P(z) \neq 0$, i.e., for any $z \in \mathcal{Z}$ such that $C'(z) \neq 0$ and $F_{Y_1 - Y_0}(C(z)) \neq 0$. We thus conclude that $C(z)$ is identified for $z \in \mathcal{Z}$.

Our goal is to identify $E(Y_1 - Y_0)$. For any $z \in \mathcal{Z}$, we have by the law of iterated expectations that

$$\begin{aligned}
E(Y_j) &= \int E(Y_j \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t) \\
&= \int_{-\infty}^{C(z)} E(Y_j \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t) \\
&\quad + \int_{C(z)}^{\infty} E(Y_j \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t)
\end{aligned}$$

for $j = 0, 1$. Using the model for D and the assumption that $Z \perp\!\!\!\perp (Y_0, Y_1)$, we have

$$\int_{C(z)}^{\infty} E(Y_1 \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t) = E(DY \mid Z = z), \tag{10.2}$$

$$\int_{-\infty}^{C(z)} E(Y_0 \mid Y_1 - Y_0 = t)\, dF_{Y_1 - Y_0}(t) = E\big((1 - D)Y \bigm| Z = z\big). \tag{10.3}$$

We identify the right-hand sides of these equations for any $z \in \mathcal{Z}$, and thus identify the left-hand sides for any $z \in \mathcal{Z}$. In particular, consider evaluating Equation (10.2) at $z = z^l$ and Equation (10.3) at $z = z^u$. Then, to bound $E(Y_1 - Y_0)$, we need to bound $\int_{-\infty}^{C(z^l)} E(Y_1 \mid Y_1 - Y_0 = t)\, dF_{Y_1-Y_0}(t)$ and $\int_{C(z^u)}^{\infty} E(Y_0 \mid Y_1 - Y_0 = t)\, dF_{Y_1-Y_0}(t)$.

We have

$$\begin{aligned}
&\int_{-\infty}^{C(z^l)} E(Y_1 \mid Y_1 - Y_0 = t)\, dF_{Y_1-Y_0}(t) \\
&\quad = \bigl(1 - P\bigl(z^l\bigr)\bigr)E\bigl[Y_1 \bigm| Z = z^l, Y_1 \leqslant Y_0 + C\bigl(z^l\bigr)\bigr] \\
&\quad \leqslant \bigl(1 - P\bigl(z^l\bigr)\bigr)E\bigl[Y_0 + C\bigl(z^l\bigr) \bigm| Z = z^l, Y_1 \leqslant Y_0 + C\bigl(z^l\bigr)\bigr] \\
&\quad = E\bigl[(1-D)Y \bigm| Z = z^l\bigr] + \bigl(1 - P\bigl(z^l\bigr)\bigr)C\bigl(z^l\bigr) \\
&\quad = E\bigl[(1-D)Y \bigm| Z = z^l\bigr] - \left[\frac{\partial}{\partial z}E(Y \mid Z = z)\Big/\frac{\partial}{\partial z}\ln\bigl(1 - P(z)\bigr)\right]\Bigg|_{z=z^l},
\end{aligned}$$

where the inequality arises from the conditioning $Y_1 \leqslant Y_0 + C(z^l)$. The final expression follows from our derivation of $C(z)$. Since $\Pr[y^l \leqslant Y_1 \leqslant y^u] = 1$ by assumption, we have

$$\begin{aligned}
&\bigl(1 - P\bigl(z^l\bigr)\bigr)y^l \\
&\quad \leqslant \int_{-\infty}^{C(z^l)} E(Y_1 \mid Y_1 - Y_0 = t)\, dF_{Y_1-Y_0}(t) \\
&\quad \leqslant E\bigl[(1-D)Y \bigm| Z = z^l\bigr] - \left[\frac{\partial}{\partial z}E(Y \mid Z = z)\Big/\frac{\partial}{\partial z}\ln\bigl(1 - P(z)\bigr)\right]\Bigg|_{z=z^l}.
\end{aligned}$$

By a parallel argument, we have

$$\begin{aligned}
P\bigl(z^u\bigr)y^l &\leqslant \int_{C(z^u)}^{\infty} E(Y_0 \mid Y_1 - Y_0 = t)\, dF_{Y_1-Y_0}(t) \\
&\leqslant E\bigl[DY \bigm| Z = z^u\bigr] + \left[\frac{\partial}{\partial z}E(Y \mid Z = z)\Big/\frac{\partial}{\partial z}\ln P(z)\right]\Bigg|_{z=z^u}.
\end{aligned}$$

We thus have the bounds

$$B^L \leqslant E(Y_1 - Y_0) \leqslant B^U,$$

with

$$\begin{aligned}
B^U = {}& E\bigl(Y \bigm| Z = z^l\bigr) \\
&- \left[\frac{\partial}{\partial z}E(Y \mid Z = z)\Big/\frac{\partial}{\partial z}\ln\bigl(1 - P(z)\bigr)\right]\Bigg|_{z=z^l} \\
&- E\bigl((1-D)Y \bigm| Z = z^u\bigr) - P\bigl(z^u\bigr)y^l,
\end{aligned}$$

$$B^L = E\big(DY \mid Z = z^l\big) + \big[1 - P\big(z^l\big)\big]y^l - E\big(Y \mid Z = z^u\big) \\ - \left[\frac{\partial}{\partial z} E(Y \mid Z = z) \Big/ \frac{\partial}{\partial z} \ln P(z)\right]\bigg|_{z=z^u}.$$

The last two terms in B^U come from the lower bound for $E(Y_0)$ and the first two terms come from the upper bound for $E(Y_1)$ just derived. The terms for B^L are decomposed in an analogous fashion, reversing the roles of the upper and lower bounds for $E(Y_1)$ and $E(Y_0)$. These bounds improve over the bounds that only impose a nonparametric selection model (Assumption S) without imposing the Roy model structure. We next consider some alternative approaches to the solution of selection and hence evaluation problems developed in the literature using replacement functions, proxy functions, and other conditions.

11. Control functions, replacement functions, and proxy variables

This chapter analyzes the main tools used to evaluate social programs in the presence of selection bias in observational data. Yet many other tools have not been analyzed. We briefly summarize these approaches. Chapter 73 (Matzkin) of this Handbook establishes conditions under which some of the methods we discuss produce identification of econometric models. Abbring and Heckman (Chapter 72) use some of these tools.

The methods of replacement functions and proxy variables all start from characterizations (U-1) and (U-2) which we repeat for convenience:

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

but

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

where θ is not observed by the analyst and (Y_0, Y_1) are not observed directly but Y is observed as are the X, Z:

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

Missing variables θ produce selection bias which creates a problem with using observational data to evaluate social programs. From (U-1), if we condition on θ, we would satisfy the condition (M-1) for matching, and hence 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:

$$\theta = \tau(X, Z, Q). \tag{11.1}$$

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

(X, Z, Q). Then

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

We can condition nonparametrically on (X, Z, Q) and do not have to know the exact functional form of τ, although knowledge of τ might reduce the dimensionality of the matching problem, 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 τ is entered into the regressions of earnings on schooling to escape the problem of ability bias, typically assuming a linear relationship between τ and θ.[189] 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. Chapter 73 (Matzkin) of this Handbook provides a rigorous proof of identification for this approach in a general nonparametric setting.

The method of replacement functions assumes that (11.1) is a perfect proxy. In many applications, this assumption is far too strong. More often, we measure θ with error. This produces a factor model or measurement error model [Aigner et al. (1984)]. Chapter 73 (Matzkin) of this Handbook surveys this method. We can represent the factor model in a general way by a system of equations:

$$Y_j = g_j(X, Z, Q, \theta, \varepsilon_j), \quad j = 1, \dots, J. \tag{11.2}$$

A linear factor model separable in the unobservables writes

$$Y_j = g_j(X, Z, Q) + \lambda_j \theta + \varepsilon_j, \quad j = 1, \dots, J, \tag{11.3}$$

where

$$(X, Z, Q) \perp\!\!\!\perp (\theta, \varepsilon_j), \quad \varepsilon_j \perp\!\!\!\perp \theta, \quad j = 1, \dots, J, \tag{11.4}$$

and the ε_j are mutually independent. Observe that under (11.2) and (11.3), Y_j controlling for X, Z, Q only imperfectly proxies θ because of the presence of ε_j. The θ are called factors, λ_j factor loadings and the ε_j "uniquenesses" [see, e.g., Aigner et al. (1984)].

A large literature, partially reviewed in Abbring and Heckman (Chapter 72), Section 1, and in Chapter 73 (Matzkin) of this Handbook, shows how to establish identification of econometric models under factor structure assumptions. Cunha, Heckman

[189] Thus if $\tau = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$, we 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 we do not need to know the coefficients $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ to implement the method. We can condition on X, Q, Z.

and Matzkin (2003), Schennach (2004) and Hu and Schennach (2006) establish identification in nonlinear models of the form (11.2).[190] The key to identification is multiple, but imperfect (because of ε_j), measurements on θ from the Y_j, $j = 1, \ldots, J$, 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 (11.4), they show how to nonparametrically identify the econometric model and the distributions of the unobservables $F_\theta(\theta)$ and $F_{\varepsilon_j}(\varepsilon_j)$. 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 [Cunha, Heckman and Matzkin (2003), Hu and Schennach (2006), Cunha, Heckman and Schennach (2006b)] extends the linear model to a nonlinear 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:

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

and (U-2) holds. For example, in the normal selection model analyzed in Chapter 70, Section 9, we broke 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 arises from the choice equation. Under normality, ε is independent of $E(U_1 \mid V)$. Further,

$$E(U_1 \mid V) = \frac{\operatorname{Cov}(U_1, V)}{\operatorname{Var}(V)} V, \tag{11.5}$$

assuming $E(U_1) = 0$ and $E(V) = 0$. In that section, we show how to construct a control function in the context of the choice model

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

Controlling for V controls for the component of θ_1 in (U-1)$'$ that gives rise to the spurious dependence. The Blundell and Powell (2003, 2004) application of the control function principle assumes functional form (11.5) but assumes that V can be perfectly proxied by a first stage equation. Thus they use a replacement function in their first stage. Their method does not work when one can only condition on D rather than on

[190] Cunha, Heckman and Schennach (2007, 2006b) apply and extend this approach to a dynamic factor setting where the θ_t are time-dependent.

$D^* = \mu_D(Z) - V$.[191] In the sample selection model, it is not necessary to use V. As developed in Chapter 70 and reviewed in Sections 4.8 and 8.3.1 of this chapter, under additive separability for the outcome equation for Y_1, we can write

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

so we "expect out" rather than solve out the effect of the component of V on U_1 and thus control for selection bias under our maintained assumptions. In terms of the propensity score, under the conditions specified in Chapter 70, we may write the preceding expression in terms of $P(Z)$:

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

where $K_1(P(Z)) = E(U_1 \mid X, Z, D = 1)$. It is not necessary to know V or be able to estimate it. The Blundell and Powell (2003, 2004) application of the control function principle assumes that the analyst can condition on and estimate V.

The Blundell–Powell method and the method of Imbens and Newey (2002) build heavily on (11.5) and implicitly make strong distributional and functional form assumptions that are not intrinsic to the method of control functions. As just noted, their method uses a replacement function to obtain $E(U_1 \mid V)$ in the first step of their procedures. The general control function method does not require a replacement function approach. 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 $(X, Z, Q) \perp\!\!\!\perp (\theta, \varepsilon_j)$, she assumes in a two equation system that $(\theta, \varepsilon_1) \perp\!\!\!\perp Y_2 \mid Y_1, X, Z$. See the discussion in Chapter 73 (Matzkin) of this Handbook.

We have not discussed panel data methods in this chapter. 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 key papers. Most of the estimators we have discussed can be adapted to a panel data setting. Heckman et al. (1998) develop difference-in-differences matching estimators. Abadie (2002) extends this work.[192] Separability between errors and observables is a key feature of the panel data approach in its standard application. Altonji and Matzkin (2005) and Matzkin (2003) present analyses of nonseparable panel data methods.

[191] Imbens and Newey (2002) extend their approach. See the discussion in Chapter 73 (Matzkin) of this Handbook.

[192] There is related work by Athey and Imbens (2006).

12. Summary

This chapter summarizes the main methods used to identify mean treatment effect parameters under semiparametric and nonparametric assumptions. We have used the marginal treatment effect as the unifying parameter to straddle a diverse econometric literature summarized in Table 1 of this chapter. For each estimator, we establish what it identifies, the economic content of the estimand and the identifying assumptions of the method.

Appendix A: Relationships among parameters using the index structure

Given the index structure, a simple relationship exists among the parameters. It is immediate from the definitions $D = \mathbf{1}(U_D \leqslant P(z))$ and $\Delta = Y_1 - Y_0$ that

$$\Delta^{\mathrm{TT}}\bigl(x, P(z)\bigr) = E\bigl(\Delta \bigm| X = x, U_D \leqslant P(z)\bigr). \tag{A.1}$$

Next consider $\Delta^{\mathrm{LATE}}(x, P(z), P(z'))$. Note that

$$\begin{aligned} &E\bigl(Y \bigm| X = x, P(Z) = P(z)\bigr) \\ &\quad = P(z)\bigl[E\bigl(Y_1 \bigm| X = x, P(Z) = P(z), D = 1\bigr)\bigr] \\ &\qquad + \bigl(1 - P(z)\bigr)\bigl[E\bigl(Y_0 \bigm| X = x, P(Z) = P(z), D = 0\bigr)\bigr] \\ &\quad = \int_0^{P(z)} E(Y_1 \mid X = x, U_D = u_D)\, du_D \\ &\qquad + \int_{P(z)}^1 E(Y_0 \mid X = x, U_D = u_D)\, du_D, \end{aligned}$$

so that

$$\begin{aligned} &E\bigl(Y \bigm| X = x, P(Z) = P(z)\bigr) - E\bigl(Y \bigm| X = x, P(Z) = P(z')\bigr) \\ &\quad = \int_{P(z')}^{P(z)} E(Y_1 \mid X = x, U_D = u_D)\, du_D \\ &\qquad - \int_{P(z')}^{P(z)} E(Y_0 \mid X = x, U_D = u_D)\, du_D, \end{aligned}$$

and thus

$$\Delta^{\mathrm{LATE}}\bigl(x, P(z), P(z')\bigr) = E\bigl(\Delta \bigm| X = x, P(z') \leqslant U_D \leqslant P(z)\bigr).$$

Notice that this expression could be taken as an alternative definition of LATE. Note that, in this expression, we could replace $P(z)$ and $P(z')$ with u_D and u'_D. No instrument needs to be available to define LATE.

We can rewrite these relationships in succinct form in the following way:

$$\Delta^{\mathrm{MTE}}(x, u_D) = E(\Delta \mid X = x, U_D = u_D),$$

$$\begin{aligned}
&\Delta^{\mathrm{ATE}}(x) = \int_0^1 E(\Delta \mid X = x, U_D = u_D)\, du_D, \\
&P(z)\big[\Delta^{\mathrm{TT}}\big(x, P(z)\big)\big] = \int_0^{P(z)} E(\Delta \mid X = x, U_D = u_D)\, du_D, \\
&\big(P(z) - P(z')\big)\big[\Delta^{\mathrm{LATE}}\big(x, P(z), P(z')\big)\big] \\
&\quad = \int_{P(z')}^{P(z)} E(\Delta \mid X = x, U_D = u_D)\, du_D.
\end{aligned} \tag{A.2}$$

We stress that everywhere in these expressions we can replace $P(z)$ with u_D and $P(z')$ with u'_D. Each parameter is an average value of MTE, $E(\Delta \mid X = x, U_D = u_D)$, but for values of U_D lying in different intervals and with different weighting functions. MTE defines the treatment effect more finely than do LATE, ATE, or TT. The relationship between MTE and LATE or TT conditional on $P(z)$ is analogous to the relationship between a probability density function and a cumulative distribution function. The probability density function and the cumulative distribution function represent the same information, but for some purposes the density function is more easily interpreted. Likewise, knowledge of TT for all $P(z)$ evaluation points is equivalent to knowledge of the MTE for all u_D evaluation points, so it is not the case that knowledge of one provides more information than knowledge of the other. However, in many choice-theoretic contexts it is often easier to interpret MTE than the TT or LATE parameters. It has the interpretation as a measure of willingness to pay on the part of people on a specified margin of participation in the program.

$\Delta^{\mathrm{MTE}}(x, u_D)$ is the average effect for people who are just indifferent between participation in the program ($D = 1$) or not ($D = 0$) if the instrument is externally set so that $P(Z) = u_D$. For values of u_D close to zero, $\Delta^{\mathrm{MTE}}(x, u_D)$ is the average effect for individuals with unobservable characteristics that make them the most inclined to participate in the program ($D = 1$), and for values of u_D close to one it is the average treatment effect for individuals with unobserved (by the econometrician) characteristics that make them the least inclined to participate. ATE integrates $\Delta^{\mathrm{MTE}}(x, u_D)$ over the entire support of U_D (from $u_D = 0$ to $u_D = 1$). It is the average effect for an individual chosen at random from the entire population. $\Delta^{\mathrm{TT}}(x, P(z))$ is the average treatment effect for persons who chose to participate at the given value of $P(Z) = P(z)$; it integrates $\Delta^{\mathrm{MTE}}(x, u_D)$ up to $u_D = P(z)$. As a result, it is primarily determined by the MTE parameter for individuals whose unobserved characteristics make them the most inclined to participate in the program. LATE is the average treatment effect for someone who would not participate if $P(Z) \leqslant P(z')$ and would participate if $P(Z) \geqslant P(z)$. The parameter $\Delta^{\mathrm{LATE}}(x, P(z), P(z'))$ integrates $\Delta^{\mathrm{MTE}}(x, u_D)$ from $u_D = P(z')$ to $u_D = P(z)$.

Using the third expression in Equation (A.2) to substitute into Equation (A.1), we obtain an alternative expression for the TT parameter as a weighted average of MTE parameters:

$$\begin{aligned}&\Delta^{\mathrm{TT}}(x)\\&\quad=\int_0^1\frac{1}{p}\biggl[\int_0^p E(\Delta\mid X=x,U_D=u_D)\,du_D\biggr]dF_{P(Z)|X,D}(p|x,D=1).\end{aligned}$$

Using Bayes' rule, it follows that

$$dF_{P(Z)|X,D}(p\mid x,1)=\frac{\Pr(D=1\mid X=x,P(Z)=p)}{\Pr(D=1\mid X=x)}\,dF_{P(Z)|X}(p|x).$$

Since $\Pr(D=1\mid X=x,P(Z)=p)=p$, it follows that

$$\begin{aligned}\Delta^{\mathrm{TT}}(x)={}&\frac{1}{\Pr(D=1\mid X=x)}\\&\times\int_0^1\biggl(\int_0^p E(\Delta\mid X=x,U_D=u_D)\,du_D\biggr)dF_{P(Z)|X}(p|x).\end{aligned}\tag{A.3}$$

Note further that since $\Pr(D=1\mid X=x)=E(P(Z)\mid X=x)=\int_0^1(1-F_{P(Z)|X}(t|x))\,dt$, we can reinterpret (A.3) as a weighted average of local IV parameters where the weighting is similar to that obtained from a length-biased, size-biased, or P-biased sample:

$$\begin{aligned}\Delta^{\mathrm{TT}}(x)={}&\frac{1}{\Pr(D=1\mid X=x)}\\&\times\int_0^1\biggl(\int_0^1\mathbf{1}(u_D\leqslant p)E(\Delta\mid X=x,U_D=u_D)\,du_D\biggr)dF_{P(Z)|X}(p|x)\\={}&\frac{1}{\int(1-F_{P(Z)|X}(t|x))\,dt}\\&\times\int_0^1\biggl(\int_0^1 E(\Delta\mid X=x,U_D=u_D)\mathbf{1}(u_D\leqslant p)\,dF_{P(Z)|X}(p|x)\biggr)du_D\\={}&\int_0^1 E(\Delta\mid X=x,U_D=u_D)\biggl(\frac{1-F_{P(Z)|X}(u_D|x)}{\int(1-F_{P(Z)|X}(t|x))\,dt}\biggr)du_D\\={}&\int_0^1 E(\Delta\mid X=x,U_D=u_D)g_x(u_D)\,du_D,\end{aligned}$$

where

$$g_x(u_D)=\frac{1-F_{P(Z)|X}(u_D|x)}{\int(1-F_{P(Z)|X}(t|x))\,dt}.$$

Thus $g_x(u_D)$ is a *weighted distribution* [Rao (1985)]. Since $g_x(u_D)$ is a nonincreasing function of u_D, we have that drawings from $g_x(u_D)$ oversample persons with low values of U_D, i.e., values of unobserved characteristics that make them the most likely to

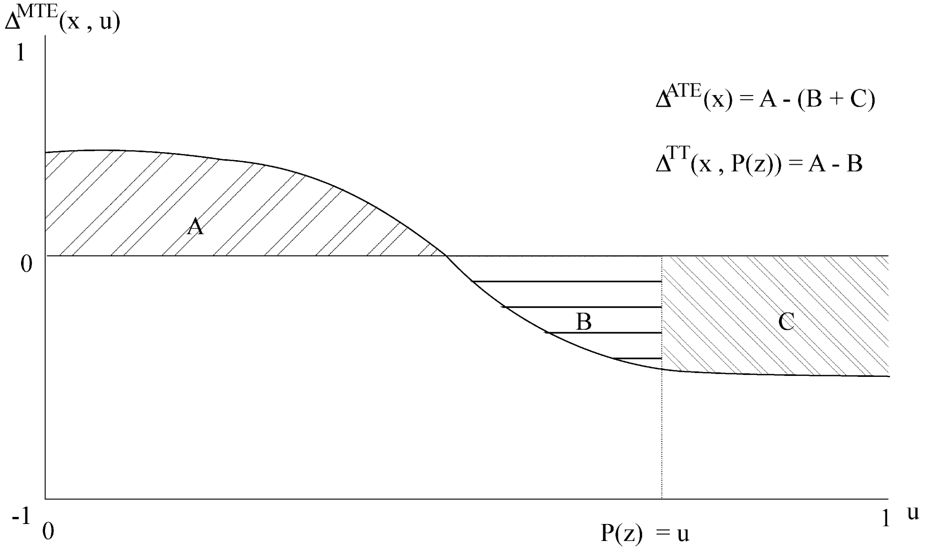


Figure A.1. MTE integrates to ATE and TT under full support (for dichotomous outcome). *Source*: Heckman and Vytlacil (2000).

participate in the program no matter what their value of $P(Z)$. Since

$$\Delta^{\mathrm{MTE}}(x, u_D) = E(\Delta \mid X = x, U_D = u_D)$$

it follows that

$$\Delta^{\mathrm{TT}}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) g_x(u_D)\, du_D.$$

The TT parameter is thus a weighted version of MTE, where $\Delta^{\mathrm{MTE}}(x, u_D)$ is given the largest weight for low u_D values and is given zero weight for $u_D \geqslant p_x^{\max}$, where $p_x^{\max}$ is the maximum value in the support of $P(Z)$ conditional on $X = x$.

Figure A.1 graphs the relationship between $\Delta^{\mathrm{MTE}}(u_D)$, Δ^{ATE} and $\Delta^{\mathrm{TT}}(P(z))$, assuming that the gains are the greatest for those with the lowest U_D values and that the gains decline as U_D increases. The curve is the MTE parameter as a function of u_D, and is drawn for the special case where the outcome variable is binary so that MTE parameter is bounded between -1 and 1. The ATE parameter averages $\Delta^{\mathrm{MTE}}(u_D)$ over the full unit interval (i.e., is the area under A minus the area under B and C in the figure). $\Delta^{\mathrm{TT}}(P(z))$ averages $\Delta^{\mathrm{MTE}}(u_D)$ up to the point $P(z)$ (is the area under A minus the area under B in the figure). Because $\Delta^{\mathrm{MTE}}(u_D)$ is assumed to be declining in u_D, the TT parameter for any given $P(z)$ evaluation point is larger then the ATE parameter.

Equation (A.2) relates each of the other parameters to the MTE parameter. One can also relate each of the other parameters to the LATE parameter. This relationship turns out to be useful later on in this chapter when we encounter conditions where LATE can be identified but MTE cannot. MTE is the limit form of LATE:

$$\Delta^{\mathrm{MTE}}(x, p) = \lim_{p' \to p} \Delta^{\mathrm{LATE}}(x, p, p').$$

Direct relationships between LATE and the other parameters are easily derived. The relationship between LATE and ATE is immediate:

$$\Delta^{\mathrm{ATE}}(x) = \Delta^{\mathrm{LATE}}(x, 0, 1).$$

Using Bayes' rule, the relationship between LATE and TT is

$$\Delta^{\mathrm{TT}}(x) = \int_0^1 \Delta^{\mathrm{LATE}}(x, 0, p) \frac{p}{\Pr(D = 1 \mid X = x)} \, dF_{P(Z)|X}(p|x). \tag{A.4}$$

Appendix B: Relaxing additive separability and independence

There are two central assumptions that underlie the latent index representation used in this chapter: that V is independent of Z, and that V and Z are additively separable in the index.[193] The latent index model with these two restrictions implies the independence and monotonicity assumptions of Imbens and Angrist (1994) and the latent index model implied by those assumptions implies a latent index model with a representation that satisfies both the independence and the monotonicity assumptions. In this appendix, we consider the sensitivity of the analysis presented in the text to relaxation of either of these assumptions.

First, consider allowing V and Z to be nonseparable in the treatment index:

$$D^* = \mu_D(Z, V),$$

$$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

while maintaining the assumption that Z is independent of (V, U_1, U_0). We do not impose any restrictions on the cross partials of μ_D. The monotonicity condition of Imbens and Angrist (1994) is that for any (z, z') pair, $\mu_D(z, v) \geqslant \mu_D(z', v)$ for all v, or $\mu_D(z, v) \leqslant \mu_D(z', v)$ for all v.[194] Vytlacil (2002) shows that monotonicity always implies one representation of μ_D as $\mu_D(z, v) = \mu_D(z) + v$. We now reconsider the analysis in the text without imposing the monotonicity condition by considering the latent index model without additive separability. Since we have imposed no structure on the $\mu_D(z, v)$ index, one can easily show that this model is equivalent to imposing the independence condition of Imbens and Angrist (1994) without imposing their

[193] Recall that $U_D = F_{V|X}(V)$.

[194] Note that the monotonicity condition is a restriction across v. For a given fixed v, it will always trivially have to be the case that either $\mu_D(z, v) \geqslant \mu_D(z', v)$ or $\mu_D(z, v) \leqslant \mu_D(z', v)$.

monotonicity condition. A random coefficient discrete choice model with $\mu_D = Z\gamma + \varepsilon$ where γ and ε are random, and γ can assume positive or negative values is an example of this case, i.e., $V = (\gamma, \varepsilon)$.

We impose the regularity condition that, for any $z \in \operatorname{Supp}(Z)$, $\mu_D(z, V)$ is absolutely continuous with respect to Lebesgue measure.[195] Let

$$\Omega(z) = \{v\colon\ \mu_D(z, v) \geqslant 0\},$$

so that

$$P(z) \equiv \Pr(D = 1 \mid Z = z) = \Pr\bigl(V \in \Omega(z)\bigr).$$

Under additive separability, $P(z) = P(z') \Leftrightarrow \Omega(z) = \Omega(z')$. This equivalence enables us to define the parameters in terms of the $P(z)$ index instead of the full z vector. In the more general case without additive separability, it is possible to have (z, z') such that $P(z) = P(z')$ and $\Omega(z) \neq \Omega(z')$. We present a random coefficient choice model example of this case in Section 4.10.1 in the text. In this case, we can no longer replace $Z = z$ with $P(Z) = P(z)$ in the conditioning sets.

Define, using $\Delta = Y_1 - Y_0$,

$$\Delta^{\mathrm{MTE}}(x, v) = E(\Delta \mid X = x, V = v).$$

For ATE, we obtain the same expression as before:

$$\Delta^{\mathrm{ATE}}(x) = \int_{-\infty}^{\infty} E(\Delta \mid X = x, V = v)\, dF_{V|X}(v).$$

For TT, we obtain a similar but slightly more complicated expression:

$$\begin{aligned}\Delta^{\mathrm{TT}}(x, z) &\equiv E(\Delta \mid X = x, Z = z, D = 1)\\ &= E\bigl(\Delta \bigm| X = x, V \in \Omega(z)\bigr)\\ &= \frac{1}{P(z)} \int_{\Omega(z)} E(\Delta \mid X = x, V = v)\, dF_{V|X}(v).\end{aligned}$$

Because it is no longer the case that we can define the parameter solely in terms of $P(z)$ instead of z, it is possible to have (z, z') such that $P(z) = P(z')$ but $\Delta^{\mathrm{TT}}(x, z) \neq \Delta^{\mathrm{TT}}(x, z')$.

Following the same derivation as used in the text for the TT parameter not conditional on Z,

$$\begin{aligned}&\Delta^{\mathrm{TT}}(x)\\ &\quad\equiv E(\Delta \mid X = x, D = 1)\\ &\quad= \int E(\Delta \mid X = x, Z = z, D = 1)\, dF_{Z|X,D}(z|x, 1)\end{aligned}$$

[195] We impose this condition to ensure that $\Pr(\mu_D(z, V) = 0) = 0$ for any $z \in \operatorname{Supp}(Z)$.

$$
\begin{aligned}
&= \frac{1}{\Pr(D=1 \mid X=x)} \\
&\quad \times \int \left[\int_{-\infty}^{\infty} \mathbf{1}\big[v \in \Omega(z)\big] E(\Delta \mid X=x, V=v)\, dF_{V|X}(v) \right] dF_{Z|X}(z|x) \\
&= \frac{1}{\Pr(D=1 \mid X=x)} \\
&\quad \times \int_{-\infty}^{\infty} \left[\int \mathbf{1}\big[v \in \Omega(z)\big] E(\Delta \mid X=x, V=v)\, dF_{Z|X}(z|x) \right] dF_{V|X}(v) \\
&= \int_{-\infty}^{\infty} E(\Delta \mid X=x, V=v) g_x(v)\, dv,
\end{aligned}
$$

where

$$
g_x(v) = \frac{\int \mathbf{1}[v \in \Omega(z)]\, dF_{Z|X}(z|x)}{\Pr(D=1 \mid X=x)} = \frac{\Pr(D=1 \mid V=v, X=x)}{\Pr(D=1 \mid X=x)}.
$$

Thus the definitions of the parameters and the relationships among them that are developed in the main text of this chapter generalize in a straightforward way to the nonseparable case. Separability allows us to define the parameters in terms of $P(z)$ instead of z and allows for slightly simpler expressions, but is not crucial for the definition of parameters or the relationship among them.

Separability is, however, crucial to the form of LATE when we allow V and Z to be additively nonseparable in the treatment index. For simplicity, we will keep the conditioning on X implicit. Define the following sets

$$
\begin{aligned}
A(z, z') &= \big\{v\colon\ \mu_D(z, v) \geqslant 0, \mu_D(z', v) \geqslant 0\big\}, \\
B(z, z') &= \big\{v\colon\ \mu_D(z, v) \geqslant 0, \mu_D(z', v) < 0\big\}, \\
C(z, z') &= \big\{v\colon\ \mu_D(z, v) < 0, \mu_D(z', v) < 0\big\}, \\
D(z, z') &= \big\{v\colon\ \mu_D(z, v) < 0, \mu_D(z', v) \geqslant 0\big\}.
\end{aligned}
$$

Monotonicity implies that either $B(z, z')$ or $D(z, z')$ is empty. Suppressing the z, z' arguments, we have

$$
\begin{aligned}
E(Y \mid Z=z) &= \Pr(A \cup B) E(Y_1 \mid A \cup B) + \Pr(C \cup D) E(Y_0 \mid C \cup D), \\
E(Y \mid Z=z') &= \Pr(A \cup D) E(Y_1 \mid A \cup D) + \Pr(B \cup C) E(Y_0 \mid B \cup C)
\end{aligned}
$$

so that

$$
\begin{aligned}
&\frac{E(Y \mid Z=z) - E(Y \mid Z=z')}{\Pr(D=1 \mid Z=z) - \Pr(D=1 \mid Z=z')} \\
&\quad = \frac{E(Y \mid Z=z) - E(Y \mid Z=z')}{\Pr(A \cup B) - \Pr(A \cup D)}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\Pr(B)E(Y_1 - Y_0 \mid B) - \Pr(D)E(Y_1 - Y_0 \mid D)}{\Pr(B) - \Pr(D)} \\
&= w_B E(\Delta \mid B) - w_D E(\Delta \mid D)
\end{aligned}
$$

with

$$
\begin{aligned}
w_B &= \frac{\Pr(B \mid B \cup D)}{\Pr(B \mid B \cup D) - \Pr(D \mid B \cup D)}, \\
w_D &= \frac{\Pr(D \mid B \cup D)}{\Pr(B \mid B \cup D) - \Pr(D \mid B \cup D)}.
\end{aligned}
$$

Under monotonicity, either $\Pr(B) = 0$ and LATE identifies $E(\Delta \mid D)$ or $\Pr(D) = 0$ and LATE identifies $E(\Delta \mid B)$. Without monotonicity, the IV estimator used as the sample analogue to LATE converges to the above weighted difference in the two terms, and the relationship between LATE and the other treatment parameters presented in the text no longer holds.

Consider what would happen if we could condition on a given v. For $v \in A \cup C$, the denominator is zero and the parameter is not well defined. For $v \in B$, the parameter is $E(\Delta \mid V = v)$, for $v \in D$, the parameter is $E(\Delta \mid V = v)$. If we could restrict conditioning to $v \in B$ (or $v \in D$), we would obtain monotonicity within the restricted sample.

Now consider LIV. For simplicity, assume z is a scalar. Assume $\mu_D(z, v)$ is continuously differentiable in (z, v), with $\mu^j(z, v)$ denoting the partial derivative with respect to the jth argument. Assume that $\mu_D(Z, V)$ is absolutely continuous with respect to Lebesgue measure. Fix some evaluation point, z_0. One can show that there may be at most a countable number of v points such that $\mu_D(z_0, v) = 0$. Let $j \in \mathcal{J} = \{1, \ldots, L\}$ index the set of v evaluation points such that $\mu_D(z_0, v) = 0$, where L may be infinity, and thus write: $\mu_D(z_0, v_j) = 0$ for all $j \in \mathcal{J}$. (Both the number of such evaluation points and the evaluation points themselves depend on the evaluation point, z_0, but we suppress this dependence for notational convenience.) Assume that there exists $\{B_k\}_{k \in \mathcal{J}}$, $\sum_{k \in \mathcal{J}} B_k < \infty$ such that $|\frac{\mu^1(z, v_k)}{\mu^2(z, v_k)}| \leqslant B_k$ for $k \in \mathcal{J}$ and all z in some neighborhood of z_0. One can show that

$$
\frac{\partial}{\partial z}\bigl[E(Y \mid Z = z)\bigr]\big|_{z=z_0} = \sum_{k=1}^{L} \frac{\mu^1(z_0, v_k)}{|\mu^2(z_0, v_k)|} E(\Delta \mid V = v_k)
$$

and

$$
\frac{\partial}{\partial z}\bigl[\Pr(D = 1 \mid Z = z)\bigr]\big|_{z=z_0} = \sum_{k=1}^{L} \frac{\mu^1(z_0, v_k)}{|\mu^2(z_0, v_k)|}.
$$

LIV is the ratio of these two terms, and does not in general equal the MTE. Thus, the relationship between LIV and MTE breaks down in the nonseparable case.

As an example, take the case where L is finite and $\frac{\mu^1(z_0,v_k)}{|\mu^2(z_0,v_k)|}$ does not vary with k. For this case,

$$\begin{aligned}\Delta^{\text{LIV}}(z_0) = {} & \Pr\bigl(\mu^1(z_0, V) > 0 \bigm| \mu(z_0, V) = 0\bigr) \\ & \cdot E\bigl(\Delta \bigm| \mu_D(z_0, V) = 0, \mu^1(z_0, V) > 0\bigr) \\ & - \Pr\bigl(\mu^1(z_0, V) < 0 \bigm| \mu(z_0, V) = 0\bigr) \\ & \cdot E\bigl(\Delta \bigm| \mu_D(z_0, V) = 0, \mu^1(z_0, V) < 0\bigr).\end{aligned}$$

Thus, while the definition of the parameters and the relationship among them does not depend crucially on the additive separability assumption, the connection between the LATE or LIV estimators and the underlying parameters crucially depends on the additive separability assumption.

Next consider the assumption that V and Z are separable in the treatment index while allowing them to be stochastically dependent:

$$D^* = \mu_D(Z) - V,$$
$$D = \begin{cases} 1 & \text{if } D^* \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

with Z independent of (U_0, U_1), but allowing Z and V to be stochastically dependent. The analysis of Vytlacil (2002) can be easily adapted to show that the latent index model with separability but without imposing independence is equivalent to imposing the monotonicity assumption of Imbens and Angrist without imposing their independence assumption.[196]

We have

$$\Omega(z) = \bigl\{v\colon \mu_D(z) \geqslant v\bigr\}$$

and

$$P(z) \equiv \Pr(D = 1 \mid Z = z) = \Pr\bigl(V \in \Omega(z) \bigm| Z = z\bigr).$$

Note that $\Omega(z) = \Omega(z') \Rightarrow \mu_D(z) = \mu_D(z')$, but $\Omega(z) = \Omega(z')$ does not imply $P(z) = P(z')$ since the distribution of V conditional on $Z = z$ need not equal the distribution of V conditional on $Z = z'$. Likewise, $P(z) = P(z')$ does not imply $\Omega(z) = \Omega(z')$. As occurred in the nonseparable case, we can no longer replace $Z = z$ with $P(Z) = P(z)$ in the conditioning sets.[197]

[196] To show that the monotonicity assumption implies a separable latent index model, one can follow the proofs of Vytlacil (2002) with the sole modification of replacing $P(z) = \Pr(D = 1 \mid Z = z)$ with $\Pr(D(z) = 1)$, where $D(z)$ is the indicator variable for whether the agent would have received treatment if Z had been externally set to z.

[197] However, we again have equivalence between the alternative conditioning sets if we assume index sufficiency, i.e., that $F_{V|Z}(v|z) = F_{V|P(Z)}(v|P(z))$.

Consider the definition of the parameters and the relationship among them. The definition of MTE and ATE in no way involves Z, nor does the relationship between them, so that both their definition and their relationship remains unchanged by allowing Z and V to be dependent. Now consider the TT parameter where now we make the dependence of X explicit:

$$\begin{aligned}\Delta^{\mathrm{TT}}(x,z) &= E\bigl(\Delta \bigm| X=x, Z=z, V \leqslant \mu_D(z)\bigr)\\ &= \frac{1}{P(z)}\int_{-\infty}^{\mu_D(z)} E(\Delta \mid X=x, V=v)\,dF_{V|Z,X}(v|z,x)\\ &= \frac{1}{P(z)}\int_{-\infty}^{\mu_D(z)} E(\Delta \mid X=x, V=v)\frac{f_{Z|V,X}(z|v,x)}{f_{Z|X}(z|x)}\,dF_{V|X}(v|x),\end{aligned}$$

where $f_{Z|X}$ and $f_{Z|V,X}$ denote the densities corresponding to $F_{Z|X}$ and $F_{Z|V,X}$ with respect to the appropriate dominating measure. We thus obtain

$$\begin{aligned}\Delta^{\mathrm{TT}}(x) &= E\bigl(\Delta \bigm| X=x, V \leqslant \mu_D(Z)\bigr)\\ &= \frac{1}{\Pr(D=1 \mid X=x)}\int\biggl[\int_{-\infty}^{\mu_D(z)} E(\Delta \mid X=x, V=v)\\ &\qquad\times \frac{f_{Z|U,X}(z|v,x)}{f_{Z|X}(z|x)}\,dF_{V|X}(v|x)\biggr]\,dF_{Z|X}(z|x)\\ &= \frac{1}{\Pr(D=1 \mid X=x)}\int_{-\infty}^{\infty}\biggl[\int \mathbf{1}\bigl[v \leqslant \mu_D(z)\bigr]E(\Delta \mid X=x, V=v)\\ &\qquad\times \frac{f_{Z|U,X}(z|v,x)}{f_{Z|X}(z|x)}\,dF_{Z|X}(z|x)\biggr]\,dF_{V|X}(v|x)\\ &= \frac{1}{\Pr(D=1 \mid X=x)}\\ &\quad\times\int_{-\infty}^{\infty}\biggl[\int \mathbf{1}\bigl[v \leqslant \mu_D(z)\bigr]\\ &\qquad\times E(\Delta \mid X=x, V=v)\,dF_{Z|V,X}(z|v,x)\biggr]\,dF_{V|X}(v|x)\\ &= \int_{-\infty}^{\infty} E(\Delta \mid X=x, V=v)g_x(v)\,dv,\end{aligned}$$

where

$$g_x(v) = \frac{\Pr(D=1 \mid V=v, X=x)}{\Pr(D=1 \mid X=x)}.$$

Thus the definitions of parameters and the relationships among the parameters that are developed in the text generalize naturally to the case where Z and V are stochastically dependent. Independence (combined with the additive separability assumption) allows us to define the parameters in terms of $P(z)$ instead of z and allows for slightly simpler expressions, but is not crucial for the definition of parameters or the relationship among them.

We next investigate LATE when we allow V and Z to be stochastically dependent. We have

$$\begin{aligned}
&E(Y \mid X=x, Z=z) \\
&\quad = P(z)\big[E(Y_1 \mid X=x, Z=z, D=1)\big] \\
&\qquad + \big(1-P(z)\big)\big[E(Y_0 \mid X=x, Z=z, D=0)\big] \\
&\quad = \int_{-\infty}^{\mu_D(z)} E(Y_1 \mid X=x, V=v)\, dF_{V|X,Z}(v|x,z) \\
&\qquad + \int_{\mu_D(z)}^{\infty} E(Y_0 \mid X=x, V=v)\, dF_{V|X,Z}(v|x,z).
\end{aligned}$$

For simplicity, take the case where $\mu_D(z) > \mu_D(z')$. Then

$$\begin{aligned}
&E(Y \mid X=x, Z=z) - E(Y \mid X=x, Z=z') \\
&\quad = \bigg[\int_{\mu_D(z')}^{\mu_D(z)} E(Y_1 \mid X=x, V=v)\, dF_{V|X,Z}(v|x,z) \\
&\qquad - \int_{\mu_D(z')}^{\mu_D(z)} E(Y_0 \mid X=x, V=v)\, dF_{V|X,Z}(v|x,z')\bigg] \\
&\qquad + \int_{-\infty}^{\mu_D(z')} E(Y_1 \mid X=x, V=v)\big(dF_{V|X,Z}(v|x,z) - dF_{V|X,Z}(v|x,z')\big) \\
&\qquad + \int_{\mu_D(z)}^{\infty} E(Y_0 \mid X=x, V=v)\big(dF_{V|X,Z}(v|x,z) - dF_{V|X,Z}(v|x,z')\big)
\end{aligned}$$

and thus

$$\begin{aligned}
&\Delta^{\mathrm{LATE}}(x,z,z') \\
&\quad = \delta_0(z)E\big(Y_1 \bigm| X=x, Z=z, \mu_D(z') \leqslant V \leqslant \mu_D(z)\big) \\
&\qquad - \delta_0(z')E\big(Y_0 \bigm| X=x, Z=z', \mu_D(z') \leqslant V \leqslant \mu_D(z)\big) \\
&\qquad + \big[\delta_1(z)E\big(Y_1 \bigm| X=x, Z=z, V \leqslant \mu_D(z')\big) \\
&\qquad - \delta_1(z')E\big(Y_1 \mid X=x, Z=z', V \leqslant \mu_D(z')\big)\big] \\
&\qquad + \big[\delta_2(z)E\big(Y_0 \bigm| X=x, Z=z, V > \mu_D(z)\big) \\
&\qquad - \delta_2(z')E\big(Y_0 \bigm| X=x, Z=z', V > \mu_D(z)\big)\big],
\end{aligned}$$

with

$$\begin{aligned}
\delta_0(t) &= \frac{\Pr(\mu_D(z') \leqslant V \leqslant \mu_D(z) \mid Z=t)}{\Pr(V \leqslant \mu_D(z) \mid Z=z, X=x) - \Pr(V \leqslant \mu_D(z') \mid Z=z', X=x)}, \\
\delta_1(t) &= \frac{\Pr(V \leqslant \mu_D(z') \mid Z=t)}{\Pr(V \leqslant \mu_D(z) \mid Z=z, X=x) - \Pr(V \leqslant \mu_D(z') \mid Z=z', X=x)},
\end{aligned}$$

$$\delta_2(t) = \frac{\Pr(V > \mu_D(z) \mid Z = t)}{\Pr(V \leqslant \mu_D(z) \mid Z = z, X = x) - \Pr(V \leqslant \mu_D(z') \mid Z = z', X = x)}.$$

Note that $\delta_0(z) = \delta_0(z') = 1$ and the two terms in brackets are zero in the case where Z and V are independent. In the more general case, δ_0 may be bigger or smaller than 1, and the terms in brackets are of unknown sign. In general, LATE may be negative even when Δ is positive for all individuals.

Now consider LIV. For simplicity, take the case where Z is a continuous scalar r.v. Let $f_{V|Z}(v|z)$ denote the density of V conditional on $Z = z$, and assume that this density is differentiable in z. Then we obtain

$$\begin{aligned}
&\frac{\partial E(Y \mid X = x, Z = z)}{\partial z}\\
&\quad = E\bigl(\Delta \mid X = x, V = \mu_D(z)\bigr)\mu_D'(z) f_{V|Z,X}\bigl(v \mid x, \mu_D(z)\bigr)\\
&\qquad + \left[\int_{-\infty}^{\mu_D(z)} E(Y_1 \mid X = x, V = v)\frac{\partial f_{V|Z,X}(v|z,x)}{\partial z}\,dv\right.\\
&\qquad \left. + \int_{\mu_D(z)}^{\infty} E(Y_0 \mid X = x, V = v)\frac{\partial f_{V|Z,X}(v|z,x)}{\partial z}\,dv\right],
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \Pr(D = 1 \mid Z = z)}{\partial z} &= f_{V|Z,X}\bigl(v \mid x, \mu_D(z)\bigr)\mu_D'(z)\\
&\quad + \int_{-\infty}^{\mu_D(z)} \frac{\partial f_{V|Z,X}(v|z,x)}{\partial z}\,dv.
\end{aligned}$$

LIV is the ratio of the two terms. Thus, without the independence condition, the relationship between LIV and the MTE breaks down.

PROOF OF EQUATION (4.20).

$$\begin{aligned}
&E(Y_p \mid X)\\
&\quad = \int E(Y_p \mid X, V = v, Z_p = z)\,dF_{V,Z_p|X}(v,z)\\
&\quad = \int \bigl(\mathbf{1}_{\Omega}(z) E(Y_1 \mid X, V = v, Z_p = z)\\
&\qquad + \mathbf{1}_{\Omega^c}(z) E(Y_0 \mid X, V = v, Z_p = z)\bigr)\,dF_{V,Z_p|X}(v,z)\\
&\quad = \int \bigl(\mathbf{1}_{\Omega}(z) E(Y_1 \mid X, V = v) + \mathbf{1}_{\Omega^c}(z) E(Y_0 \mid X, V = v)\bigr)\,dF_{V,Z_p|X}(v,z)\\
&\quad = \int \biggl[\int \bigl(\mathbf{1}_{\Omega}(z) E(Y_1 \mid X, V = v)\\
&\qquad + \mathbf{1}_{\Omega^c}(z) E(Y_0 \mid X, V = v)\bigr)\,dF_{Z_p|X}(z)\biggr]\,dF_{V|X}(v)
\end{aligned}$$

$$
\begin{aligned}
&= \int \bigl[\Pr[Z_p \in \Omega \mid X] E(Y_1 \mid X, V = v) \\
&\quad + \bigl(1 - \Pr[Z_p \in \Omega(z) \mid X]\bigr) E(Y_0 \mid X, V = v)\bigr]\, dF_{V|X}(v),
\end{aligned}
$$

where $\Omega^c(z)$ denotes the complement of $\Omega(z)$ and where the first equality follows from the law of iterated expectations; the second equality follows by plugging in our threshold crossing model for D; the third equality follows from independence $Z \perp\!\!\!\perp (Y_1, Y_0, V) \mid X$; the fourth and fifth equalities follow by an application of Fubini's Theorem and a rearrangement of terms. Fubini's Theorem may be applied by assumption (A-4). Thus comparing policy p to policy p', we obtain (4.20):

$$
\begin{aligned}
&E(Y_p \mid X) - E(Y_{p'} \mid X) \\
&\quad = \int E(\Delta \mid X, V = v)\bigl(\Pr[Z_p \in \Omega \mid X] - \Pr[Z_{p'} \in \Omega \mid X]\bigr)\, dF_{V|X}(v).
\end{aligned}
$$

□

PROOF OF EQUATION (4.21).

$$
\begin{aligned}
&E(Y_p \mid X) \\
&\quad = \int E(Y_p \mid X, V = v, Z_p = z)\, dF_{V,Z_p|X}(v, z) \\
&\quad = \int \bigl[\mathbf{1}_{[-\infty,\mu_D(z)]}(v) E(Y_1 \mid X, Z = z, V = v) \\
&\qquad + \mathbf{1}_{(\mu_D(z),\infty]}(v) E(Y_0 \mid X, Z = z, V = v)\bigr]\, dF_{V,Z_p|X}(v, z) \\
&\quad = \int \bigl[\,\mathbf{1}_{[-\infty,\mu_D(z)]}(v) E(Y_1 \mid X, V = v) \\
&\qquad + \mathbf{1}_{(\mu_D(z),\infty]}(v) E(Y_0 \mid X, V = v)\bigr]\, dF_{V,Z_p|X}(v, z) \\
&\quad = \int \biggl[\int \bigl(\mathbf{1}_{[-\infty,\mu_D(z)]}(v) E(Y_1 \mid X, V = v) \\
&\qquad + \mathbf{1}_{(\mu_D(z),\infty]}(v) E(Y_0 \mid X, V = v)\bigr)\, dF_{Z_p|V}(z|v)\biggr]\, dF_{V|X}(v) \\
&\quad = \int \bigl[\bigl(1 - \Pr\bigl[\mu_D(Z_p) < v \bigm| V = v\bigr]\bigr) E(Y_1 \mid X, V = v) \\
&\qquad + \Pr\bigl[\mu_D(Z_p) < v \bigm| V = v\bigr] E(Y_0 \mid X, V = v)\bigr]\, dF_{V|X}(v),
\end{aligned}
$$

where the first equality follows from the law of iterated expectations; the second equality follows by plugging in our model for D; the third equality follows from independence $Z \perp\!\!\!\perp (Y_1, Y_0) \mid X, V$; the fourth equality follows by an application of Fubini's Theorem; and the final equality follows immediately. Thus comparing policy p to policy p', we obtain (4.21) in the text. □

Appendix C: Derivation of PRTE and implications of noninvariance for PRTE

PROOF OF EQUATION (3.6). To simplify the notation, assume that $\Upsilon(Y) = Y$. Modifications required for the more general case are obvious. Define $\mathbf{1}_{\mathcal{P}}(t)$ to be the indicator function for the event $t \in \mathcal{P}$. Then

$$\begin{aligned}
&E(Y_p \mid X)\\
&\quad= \int_0^1 E\bigl(Y_p \bigm| X, P_p(Z_p) = t\bigr)\, dF_{P_p|X}(t)\\
&\quad= \int_0^1 \left[\int_0^1 \bigl[\mathbf{1}_{[0,t]}(u_D) E(Y_{1,p} \mid X, U_D = u_D)\right.\\
&\qquad\left. + \mathbf{1}_{(t,1]}(u_D) E(Y_{0,p} \mid X, U_D = u_D)\bigr]\, du_D\right] dF_{P_p|X}(t)\\
&\quad= \int_0^1 \left[\int_0^1 \bigl[\mathbf{1}_{[u_D,1]}(t) E(Y_{1,p} \mid X, U_D = u_D)\right.\\
&\qquad\left. + \mathbf{1}_{(0,u_D]}(t) E(Y_{0,p} \mid X, U_D = u_D)\bigr]\, dF_{P_p|X}(t)\right] du_D\\
&\quad= \int_0^1 \bigl[\bigl(1 - F_{P_p|X}(u_D)\bigr) E(Y_{1,p} \mid X, U_D = u_D)\\
&\qquad + F_{P_p|X}(u_D) E(Y_{0,p} \mid X, U_D = u_D)\bigr]\, du_D.
\end{aligned}$$ [198]

This derivation involves changing the order of integration. Note that from (A-4),

$$\begin{aligned}
&E\bigl|\mathbf{1}_{[0,t]}(u_D) E(Y_{1,p} \mid X, U_D = u_D)\\
&\quad + \mathbf{1}_{(t,1]}(u_D) E(Y_{0,p} \mid X, U_D = u_D)\bigr| \leqslant E\bigl(|Y_1| + |Y_0|\bigr) < \infty,
\end{aligned}$$

so the change in the order of integration is valid by Fubini's Theorem. Comparing policy p to policy p',

$$\begin{aligned}
&E(Y_p \mid X) - E(Y_{p'} \mid X)\\
&\quad= \int_0^1 E(\Delta \mid X, U_D = u_D)\bigl(F_{P_{p'}|X}(u_D) - F_{P_p|X}(u_D)\bigr)\, du_D,
\end{aligned}$$

which gives the required weights. (Recall $\Delta = Y_1 - Y_0$ and from (A-7) we can drop the p, p' subscripts on outcomes and errors.) □

[198] Recall that p denotes the policy in this section and t is a value assumed by $P(Z)$.

RELAXING A-7 *(Implications of noninvariance for PRTE)*. Suppose that all of the assumptions invoked up through Section 3.2 are satisfied, including additive separability in the latent index choice equation (3.3) (equivalently, the monotonicity or uniformity condition). Impose the normalization that the distribution of U_D is unit uniform ($U_D = F_{V|X}(V \mid X)$). Suppose however, contrary to (A-7), that the distribution of (Y_1, Y_0, U_D, X) is different under the two regimes p and p'. Thus, let $(Y_{1,p}, Y_{0,p}, U_{D,p}, X_p)$ and $(Y_{1,p'}, Y_{0,p'}, U_{D,p'}, X_{p'})$ denote the random vectors under regimes p and p', respectively. Following the same analysis as used to derive Equation (3.6), the PRTE conditional on X is given by

$$\begin{aligned}
&E(Y_p \mid X_p = x) - E(Y_{p'} \mid X_{p'} = x) \\
&\quad = \int_0^1 E(Y_{1,p} - Y_{0,p} \mid X_p = x, U_{D,p} = u) \\
&\qquad \times \left[F_{P_{p'}|X_{p'}}(u|x) - F_{P_p|X_p}(u|x)\right] du && \text{(I)} \\
&\qquad + \int_0^1 \left[E(Y_{0,p} \mid X_p = x, U_{D,p} = u) - E(Y_{0,p'} \mid X_{p'} = x, U_{D,p'} = u)\right] du && \text{(II)} \\
&\qquad + \int_0^1 \left[\left(1 - F_{P_{p'}|X_{p'}}(u|x)\right)\left(E(Y_{1,p} - Y_{0,p} \mid X_p = x, U_{D,p} = u)\right.\right. \\
&\qquad \left.\left. - E(Y_{1,p'} - Y_{0,p'} \mid X_{p'} = x, U_{D,p'} = u)\right)\right] du. && \text{(III)}
\end{aligned}$$

Thus, when the policy affects the distribution of (Y_1, Y_0, U_D, X), the PRTE is given by the sum of three terms: (I) the value of PRTE if the policy did not affect (Y_1, Y_0, X, U_D); (II) the weighted effect of the policy change on $E(Y_0 \mid X, U_D)$; and (III) the weighted effect of the policy change on MTE. Evaluating the PRTE requires knowledge of the MTE function in both regimes, knowledge of $E(Y_0 \mid X = x, U_D = u)$ in both regimes, as well as knowledge of the distribution of $P(Z)$ in both regimes. Note, however, that if we assume that the distribution of $(Y_{1,p}, Y_{0,p}, U_{D,p})$ conditional on $X_p = x$ equals the distribution of $(Y_{1,p'}, Y_{0,p'}, U_{D,p'})$ conditional on $X_{p'} = x$, then $E(Y_{1,p} \mid U_{D,p} = u, X_p = x) = E(Y_{1,p'} \mid U_{D,p'} = u, X_{p'} = x)$, $E(Y_{0,p} \mid U_{D,p} = u, X_p = x) = E(Y_{0,p'} \mid U_{D,p'} = u, X_{p'} = x)$, and thus the last two terms vanish and the expression for PRTE simplifies to the expression of Equation (3.6).

Appendix D: Deriving the IV weights on MTE

We consider instrumental variables conditional on $X = x$ using a general function of Z as an instrument. To simplify the notation, we keep the conditioning on X implicit. Let $J(Z)$ be any function of Z such that $\operatorname{Cov}(J(Z), D) \neq 0$. Consider the population analogue of the IV estimator,

$$\frac{\operatorname{Cov}(J(Z), Y)}{\operatorname{Cov}(J(Z), D)}.$$

First consider the numerator of this expression,

$$\begin{aligned}\operatorname{Cov}\bigl(J(Z),Y\bigr) &= E\bigl(\bigl[J(Z)-E\bigl(J(Z)\bigr)\bigr]Y\bigr)\\ &= E\bigl(\bigl(J(Z)-E\bigl(J(Z)\bigr)\bigr)\bigl(Y_0+D(Y_1-Y_0)\bigr)\bigr)\\ &= E\bigl(\bigl(J(Z)-E\bigl(J(Z)\bigr)\bigr)D(Y_1-Y_0)\bigr),\end{aligned}$$

where the second equality comes from substituting in the definition of Y and the third equality follows from conditional independence assumption (A-1). Define $\tilde{J}(Z) \equiv J(Z) - E(J(Z))$. Then

$$\begin{aligned}&\operatorname{Cov}\bigl(J(Z),Y\bigr)\\ &\quad= E\bigl(\tilde{J}(Z)\mathbf{1}\bigl[U_D \leqslant P(Z)\bigr](Y_1-Y_0)\bigr)\\ &\quad= E\bigl(\tilde{J}(Z)\mathbf{1}\bigl[U_D \leqslant P(Z)\bigr]E(Y_1-Y_0 \mid Z, U_D)\bigr)\\ &\quad= E\bigl(\tilde{J}(Z)\mathbf{1}\bigl[U_D \leqslant P(Z)\bigr]E(Y_1-Y_0 \mid U_D)\bigr)\\ &\quad= E_{U_D}\bigl(E_Z\bigl[\tilde{J}(Z)\mathbf{1}\bigl[U_D \leqslant P(Z)\bigr] \bigm| U_D\bigr]E(Y_1-Y_0 \mid U_D)\bigr)\\ &\quad= \int_0^1 \bigl\{E\bigl(\tilde{J}(Z) \bigm| P(Z) \geqslant u_D\bigr)\Pr\bigl(P(Z) \geqslant u_D\bigr)E(Y_1-Y_0 \mid U_D = u_D)\bigr\}\,du_D\\ &\quad= \int_0^1 \Delta^{\mathrm{MTE}}(x,u_D)E\bigl(\tilde{J}(Z) \bigm| P(Z) \geqslant u_D\bigr)\Pr\bigl(P(Z) \geqslant u_D\bigr)\,du_D,\end{aligned}$$

where the first equality follows from plugging in the model for D; the second equality follows from the law of iterated expectations with the inside expectation conditional on (Z, U_D); the third equality follows from conditional independence assumption (A-1); the fourth equality follows from Fubini's Theorem and the law of iterated expectations with the inside expectation conditional on $(U_D = u_D)$ (and implicitly on X); this allows to reverse the order of integration in a multiple integral; the fifth equality follows from the normalization that U_D is distributed unit uniform conditional on X; and the final equality follows from plugging in the definition of Δ^{MTE}. Next consider the denominator of the IV estimand. Observe that by iterated expectations

$$\operatorname{Cov}\bigl(J(Z),D\bigr) = \operatorname{Cov}\bigl(J(Z),P(Z)\bigr).$$

Thus, the population analogue of the IV estimator is given by

$$\int_0^1 \Delta^{\mathrm{MTE}}(u_D)\omega(u_D)\,du_D, \tag{D.1}$$

where

$$\omega(u_D) = \frac{E(\tilde{J}(Z) \mid P(Z) \geqslant u_D)\Pr(P(Z) \geqslant u_D)}{\operatorname{Cov}(J(Z),P(Z))}, \tag{D.2}$$

where by assumption $\operatorname{Cov}(J(Z),P(Z)) \neq 0$.

If $J(Z)$ and $P(Z)$ are continuous random variables, then an interpretation of the weight can be derived from (D.2) by noting that

$$\begin{aligned}\int \big(j - E\big(J(Z)\big)\big) \int_{u_D}^{1} f_{P,J}(t, j)\, dt\, dj \\ = \int \big(j - E\big(J(Z)\big)\big) f_J(j) \int_{u_D}^{1} f_{P|J}\big(t \mid J(Z) = j\big)\, dt\, dj.\end{aligned}$$

Write

$$\begin{aligned}\int_{u_D}^{1} f_{P|J}\big(t \mid J(Z) = j\big)\, dt &= 1 - F_{P|J}\big(u_D \mid J(Z) = j\big) \\ &= S_{P|J}\big(u_D \mid J(Z) = j\big),\end{aligned}$$

where $S_{P|J}(u_D \mid J(Z) = j)$ is the probability of $(P(Z) \geqslant u_D)$ given $J(Z) = j$ (and implicitly $X = x$). Likewise, $\Pr[P(Z) > U_D \mid J(Z)] = S_{P|J}(U_D \mid J(Z))$. Using these results, we may write the weight as

$$\omega(u_D) = \frac{\operatorname{Cov}(J(Z), S_{P|J}(u_D \mid J(Z)))}{\operatorname{Cov}(J(Z), S_{P|J}(U_D \mid J(Z)))}.$$

For fixed u_D and x evaluation points, $S_{P|J}(u_D \mid J(Z))$ is a function of the random variable $J(Z)$. The numerator of the preceding expression is the covariance between $J(Z)$ and the probability that the random variable $P(Z)$ is greater than the evaluation point u_D conditional on $J(Z)$.

$S_{P|J}(U_D \mid J(Z))$ is a function of the random variables U_D and $J(Z)$. The denominator of the above expression is the covariance between $J(Z)$ and the probability that the random variable $P(Z)$ is greater than the random variable U_D conditional on $J(Z)$. Thus, it is clear that if the covariance between $J(Z)$ and the conditional probability that $(P(Z) > u_D)$ given $J(Z)$ is positive for all u_D, then the weights are positive. The conditioning is trivially satisfied if $J(Z) = P(Z)$, so the weights are positive and IV estimates a gross treatment effect. If the $J(Z)$ and $P(Z)$ are discrete-valued, we obtain expressions and (4.15) and (4.16) in the text.

D.1. Yitzhaki's Theorem and the IV weights [Yitzhaki (1989)]

THEOREM. *Assume (Y, X) i.i.d., $E(|Y|) < \infty$, $E(|X|) < \infty$, $g(X) = E(Y \mid X)$, $g'(X)$ exists and $E(|g'(x)|) < \infty$. Let $\mu_Y = E(Y)$ and $\mu_X = E(X)$. Then,*

$$\frac{\operatorname{Cov}(Y, X)}{\operatorname{Var}(X)} = \int_{-\infty}^{\infty} g'(t)\omega(t)\, dt,$$

where

$$\omega(t) = \frac{1}{\operatorname{Var}(X)} \int_{t}^{\infty} (x - \mu_X) f_X(x)\, dx$$

$$= \frac{1}{\operatorname{Var}(X)} E(X - \mu_X \mid X > t) \Pr(X > t).$$

PROOF.

$$\begin{aligned}\operatorname{Cov}(Y, X) &= \operatorname{Cov}\bigl(E(Y \mid X), X\bigr) = \operatorname{Cov}\bigl(g(X), X\bigr) \\ &= \int_{-\infty}^{\infty} g(t)(t - \mu_X)\, f_X(t)\, dt.\end{aligned}$$

Integration by parts implies that

$$\begin{aligned}&= g(t) \int_{-\infty}^{t} (x - \mu_X) f_X(x)\, dx \Big|_{-\infty}^{\infty} \\ &\quad - \int_{-\infty}^{\infty} g'(t) \int_{-\infty}^{t} (x - \mu_X)\, f_X(x)\, dx\, dt \\ &= \int_{-\infty}^{\infty} g'(t) \int_{t}^{\infty} (x - \mu_X)\, f_X(x)\, dx\, dt,\end{aligned}$$

since $E(X - \mu_X) = 0$ and the first term in the first expression vanishes.

Therefore,

$$\operatorname{Cov}(Y, X) = \int_{-\infty}^{\infty} g'(t) E(X - \mu_X \mid X > t) \Pr(X > t)\, dt,$$

so

$$\omega(t) = \frac{1}{\operatorname{Var}(X)} E(X - \mu_X \mid X > t) \Pr(X > t).$$

□

Notice that:

(i) The weights are nonnegative ($\omega(t) \geqslant 0$).
(ii) They integrate to one (use an integration by parts formula).
(iii) $\omega(t) \to 0$ when $t \to -\infty$, and $\omega(t) \to 0$ when $t \to \infty$.

We get the formula in the text when we use $P(Z)$, with a suitably defined domain, in place of X. We apply Yitzhaki's result to the treatment effect model:

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

$$\begin{aligned}E\bigl(Y \bigm| P(Z)\bigr) &= \alpha + E\bigl(\beta \bigm| D = 1, P(Z)\bigr) P(Z) \\ &= \alpha + E\bigl(\beta \bigm| P(Z) > u_D, P(Z)\bigr) P(Z) \\ &= g\bigl(P(Z)\bigr).\end{aligned}$$

By the law of iterated expectations, we eliminate the conditioning on $D = 0$. Using our previous results for OLS,

$$\text{IV} = \frac{\operatorname{Cov}(Y, P(Z))}{\operatorname{Cov}(D, P(Z))} = \int g'(t)\omega(t)\,dt,$$
$$g'(t) = \frac{\partial[E(\beta \mid D = 1, P(Z))]P(Z)}{\partial P(Z)}\bigg|_{P(Z)=t},$$
$$\omega(t) = \frac{\int_t^1 [\varphi - E(P(Z))] f_P(\varphi)\,d\varphi}{\operatorname{Cov}(P(Z), D)}.$$

Under (A-1) to (A-5) and separability, $g'(t) = \Delta^{\text{MTE}}(t)$ but $g'(t) = \text{LIV}$, for $P(Z)$ as an instrument.

D.2. Relationship of our weights to the Yitzhaki weights[199]

Under our assumptions the Yitzhaki weights and ours are equivalent. Using (4.12),

$$\begin{aligned}\operatorname{Cov}\bigl(J(Z), Y\bigr) &= E(Y \cdot \tilde{J}) = E\bigl(E(Y \mid Z) \cdot \tilde{J}(Z)\bigr) \\ &= E\bigl(E\bigl(Y \bigm| P(Z)\bigr) \cdot \tilde{J}(Z)\bigr) = E\bigl(g\bigl(P(Z)\bigr) \cdot \tilde{J}(Z)\bigr).\end{aligned}$$

The third equality follows from index sufficiency and $\tilde{J} = J(Z) - E(J(Z) \mid P(Z) \geqslant u_D)$, where $E(Y \mid P(Z)) = g(P(Z))$. Writing out the expectation and assuming that $J(Z)$ and $P(Z)$ are continuous random variables with joint density $f_{P,J}$ and that $J(Z)$ has support $[\underline{J}, \bar{J}]$,

$$\begin{aligned}\operatorname{Cov}\bigl(J(Z), Y\bigr) &= \int_0^1 \int_{\underline{J}}^{\bar{J}} g(u_D)\tilde{j} f_{P,J}(u_D, j)\,dj\,du_D \\ &= \int_0^1 g(u_D) \int_{\underline{J}}^{\bar{J}} \tilde{j} f_{P,J}(u_D, j)\,dj\,du_D.\end{aligned}$$

Using an integration by parts argument as in Yitzhaki (1989) and as summarized in Heckman, Urzua and Vytlacil (2006), we obtain

$$\begin{aligned}\operatorname{Cov}\bigl(J(Z), Y\bigr) &= g(u_D) \int_0^{u_D} \int_{\underline{J}}^{\bar{J}} \tilde{j} f_{P,J}(p, j)\,dj\,dp\bigg|_0^1 \\ &\quad - \int_0^1 g'(u_D) \int_0^{u_D} \int_{\underline{J}}^{\bar{J}} \tilde{j} f_{P,J}(p, j)\,dj\,dp\,du_D \\ &= \int_0^1 g'(u_D) \int_{u_D}^1 \int_{\underline{J}}^{\bar{J}} \tilde{j} f_{P,J}(p, j)\,dj\,dp\,du_D\end{aligned}$$

[199] We thank Benjamin Moll for the derivation presented in this subsection.

$$= \int_0^1 g'(u_D) E\big(\tilde{J}(Z) \mid P(Z) \geqslant u_D\big) \Pr\big(P(Z) \geqslant u_D\big)\, du_D,$$

which is then exactly the expression given in (4.12), where

$$g'(u_D) = \left.\frac{\partial E(Y \mid P(Z) = p)}{\partial P(Z)}\right|_{p=u_D} = \Delta^{\mathrm{MTE}}(u_D).$$

Appendix E: Derivation of the weights for the mixture of normals example

Writing E_1 as the expectation for group 1, letting μ_1 be the mean of Z for population 1 and μ_{11} be the mean of the first component of Z,

$$\begin{aligned}
&E_1(Z_1 \mid \gamma' Z > v) \\
&\quad = \mu_{11} + \frac{\gamma' \Sigma_1^1}{\gamma' \Sigma_1 \gamma} E_1(Z_1 - \mu_1 \mid \gamma' Z > v) \\
&\quad = \mu_{11} + \frac{\gamma' \Sigma_1^1}{(\gamma' \Sigma_1 \gamma)^{1/2}} E_1\left(\frac{\gamma'(Z - \mu_1)}{(\gamma' \Sigma_1 \gamma)^{1/2}} \,\middle|\, \frac{\gamma'(Z - \mu_1)}{(\gamma' \Sigma_1 \gamma)^{1/2}} > \frac{(v - \gamma' \mu_1)}{(\gamma' \Sigma_1 \gamma)^{1/2}}\right) \\
&\quad = \mu_{11} + \frac{\gamma' \Sigma_1^1}{(\gamma' \Sigma_1 \gamma)^{1/2}} \lambda\left(\frac{(v - \gamma' \mu_1)}{(\gamma' \Sigma_1 \gamma)^{1/2}}\right),
\end{aligned}$$

where

$$\lambda(c) = \frac{1}{\sqrt{2\pi}} \frac{e^{-c^2/2}}{\Phi(-c)},$$

where $\Phi(\cdot)$ is the unit normal cumulative distribution function.

By the same logic, in the second group:

$$E_2(Z_1 \mid \gamma' Z > v) = \mu_{21} + \frac{\gamma' \Sigma_2^1}{(\gamma' \Sigma_2 \gamma)^{1/2}} \lambda\left(\frac{(v - \gamma' \mu_2)}{(\gamma' \Sigma_2 \gamma)^{1/2}}\right).$$

Therefore for the overall population we obtain

$$\begin{aligned}
&E\big(Z_1 - E(Z_1) \,\big|\, \gamma' Z > v\big) \Pr(\gamma' Z > v) \\
&\quad = (P_1 \mu_{11} + P_2 \mu_{21}) \Pr(\gamma' Z > v) \\
&\qquad + \frac{P_1 \gamma \Sigma_1^1}{(\gamma' \Sigma_1 \gamma)^{1/2} \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{v - \gamma' \mu_1}{(\gamma' \Sigma_1 \gamma)^{1/2}}\right)^2\right] \\
&\qquad + \frac{P_2 \gamma \Sigma_2^1}{(\gamma' \Sigma_2 \gamma)^{1/2} \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{v - \gamma' \mu_2}{(\gamma' \Sigma_2 \gamma)^{1/2}}\right)^2\right] \\
&\qquad - (P_1 \mu_{11} + P_2 \mu_{21}) \Pr(\gamma' Z > v) \\
&\quad = \frac{P_1 \gamma \Sigma_1^1}{(\gamma' \Sigma_1 \gamma)^{1/2} \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{v - \gamma' \mu_1}{(\gamma' \Sigma_1 \gamma)^{1/2}}\right)^2\right]
\end{aligned}$$

$$+\frac{P_2\gamma\Sigma_2^1}{(\gamma'\Sigma_2\gamma)^{1/2}\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{v-\gamma'\mu_2}{(\gamma'\Sigma_2\gamma)^{1/2}}\right)^2\right].$$

We need $\mathrm{Cov}(D,Z_1)$. To obtain it, observe that

$$D=\mathbf{1}[\gamma'Z-V>0],$$
$$E(Z_1D)=E\big(Z_1\mathbf{1}(\gamma'Z-V\geqslant 0)\big).$$

Let E_1 denote the expectation for group 1, and let E_2 denote the expectation for group 2.

$$\begin{aligned}
E(Z_1D)={}&\left\{P_1\left[\mu_{11}+\frac{\gamma'\Sigma_1^1}{\gamma'\Sigma_1\gamma+\sigma_V^2}E_1(Z_1-\mu_{11}\mid\gamma'Z-V\geqslant 0)\right]\right.\\
&\left.+P_2\left[\mu_{21}+\frac{\gamma'\Sigma_2^1}{\gamma'\Sigma_2\gamma+\sigma_V^2}E_2(Z_1-\mu_{21}\mid\gamma'Z-V\geqslant 0)\right]\right\}\\
&\times\Pr\big[(\gamma'Z-V)>0\big]\\
={}&(P_1\mu_{11}+P_2\mu_{21})\Pr(\gamma'Z-V\geqslant 0)\\
&+\frac{P_1\gamma'\Sigma_1^1}{(\gamma'\Sigma_1\gamma+\sigma_V^2)^{1/2}\sqrt{2\pi}}\exp\left[-\left(\frac{-\gamma'\mu_1}{(\gamma'\Sigma_1\gamma+\sigma_V^2)^{1/2}}\right)^2\right]\\
&+\frac{P_2\gamma'\Sigma_2^1}{(\gamma'\Sigma_2\gamma+\sigma_V^2)^{1/2}\sqrt{2\pi}}\exp\left[-\left(\frac{-\gamma'\mu_2}{(\gamma'\Sigma_2\gamma+\sigma_V^2)^{1/2}}\right)^2\right].
\end{aligned}$$

Because

$$E(D)E(Z_1)=\Pr(\gamma'Z-V\geqslant 0)(P_1\mu_{11}+P_2\mu_{21})$$

and

$$\begin{aligned}
&\mathrm{Cov}(D,Z_1)=E(Z_1D)-E(Z_1)E(D)\\
\therefore\quad &\mathrm{Cov}(D,Z_1)=\frac{P_1\gamma'\Sigma_1^1}{(\gamma'\Sigma_1\gamma+\sigma_V^2)^{1/2}\sqrt{2\pi}}\exp\left[-\left(\frac{-\gamma'\mu_1}{(\gamma'\Sigma_1\gamma+\sigma_V^2)^{1/2}}\right)^2\right]\\
&\qquad+\frac{P_2\gamma'\Sigma_2^1}{(\gamma'\Sigma_2\gamma+\sigma_V^2)^{1/2}\sqrt{2\pi}}\exp\left[-\left(\frac{-\gamma'\mu_2}{(\gamma'\Sigma_2\gamma+\sigma_V^2)^{1/2}}\right)^2\right].
\end{aligned}$$

Thus the IV weights for this set-up are

$$\begin{aligned}
\tilde{\omega}_{\mathrm{IV}}(v)={}&\left\{\left[\frac{P_1\gamma'\Sigma_1^1}{(\gamma'\Sigma_1\gamma)^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{v-\gamma'\mu_1}{(\gamma'\Sigma_1\gamma)^{1/2}}\right)^2\right]\right.\right.\\
&\left.\left.+\frac{P_2\gamma'\Sigma_2^1}{(\gamma'\Sigma_2\gamma)^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{v-\gamma'\mu_2}{(\gamma'\Sigma_2\gamma)^{1/2}}\right)^2\right]\right]f_V(v)\right\}
\end{aligned}$$

$$\times \left\{ \frac{P_1 \gamma' \Sigma_1^1}{(\gamma' \Sigma_1 \gamma + \sigma_V^2)^{1/2}} \exp\left[-\left(\frac{-\gamma' \mu_1}{(\gamma' \Sigma_1 \gamma + \sigma_V^2)^{1/2}} \right)^2 \right] + \frac{P_2 \gamma' \Sigma_2^1}{(\gamma' \Sigma_2 \gamma + \sigma_V^2)^{1/2}} \exp\left[-\left(\frac{-\gamma' \mu_2}{(\gamma' \Sigma_2 \gamma + \sigma_V^2)^{1/2}} \right)^2 \right] \right\}^{-1},$$

where σ_V^2 represents the variance of V. Clearly, $\tilde{\omega}_{\mathrm{IV}}(-\infty) = 0$, $\tilde{\omega}_{\mathrm{IV}}(\infty) = 0$ and the weights integrate to one over the support of $V = (-\infty, \infty)$. Observe that the weights must be positive if $P_2 = 0$. Thus the structure of the covariances of the instrument with the choice index $\gamma' Z$ is a key determinant of the positivity of the weights for any instrument. It has nothing to do with the *ceteris paribus* effect of Z_1 on $\gamma' Z$ or $P(Z)$ in the general case.

A necessary condition for $\omega_{\mathrm{IV}} < 0$ over some values of v is that $\operatorname{sign}(\gamma' \Sigma_1^1) = -\operatorname{sign}(\gamma' \Sigma_2^1)$, i.e., that the covariance between Z_1 and $\gamma' Z$ be of opposite signs in the two subpopulations so Z_1 and $P(Z)$ have different relationships in the two component populations. Without loss of generality, assume that $\gamma' \Sigma_1^1 > 0$. If it equals zero, we fail the rank condition in the first population and we are back to a one subpopulation model with positive weights. The numerator of the expression for $\omega_{\mathrm{IV}}(v)$ switches signs if for some values of v,

$$\frac{P_1 \gamma' \Sigma_1^1}{(\gamma' \Sigma_1 \gamma)^{1/2}} \exp\left[-\frac{1}{2} \left(\frac{v - \gamma' \mu_1}{(\gamma' \Sigma_1 \gamma)^{1/2}} \right)^2 \right] < -\frac{P_2 \gamma' \Sigma_2^1}{(\gamma' \Sigma_2 \gamma)^{1/2}} \exp\left[-\frac{1}{2} \left(\frac{v - \gamma' \mu_2}{(\gamma' \Sigma_2 \gamma)^{1/2}} \right)^2 \right],$$

while for other values the inequality is reversed. (Observe that the denominator is a constant.) Rewriting and taking logarithms, we obtain under the assumption that $\operatorname{sign}(\gamma' \Sigma_1^1) = -\operatorname{sign}(\gamma' \Sigma_2^1)$, the following expression:

$$\frac{1}{2} \left[\frac{(v - \gamma' \mu_2)^2}{\gamma' \Sigma_2 \gamma} - \frac{(v - \gamma' \mu_1)^2}{\gamma' \Sigma_1 \gamma} \right] < \ln\left(\frac{1 - P_1}{P_1} \right) + \ln\left[\frac{-\gamma' \Sigma_2^1}{\gamma' \Sigma_1^1} \right] + \ln\left[\frac{\gamma' \Sigma_1 \gamma}{\gamma' \Sigma_2 \gamma} \right],$$

where we assume $0 < P_1 < 1$. Observe that $\frac{1-P_1}{P_1}$ can be made as large or as small a non-negative number as we like by varying P_1. Varying (μ_1, μ_2) does not affect the right-hand side. For $\mu_1 = \mu_2 = 0$, the inequality becomes

$$\frac{1}{2} v^2 \left[\frac{1}{\gamma' \Sigma_2 \gamma} - \frac{1}{\gamma' \Sigma_1 \gamma} \right] < \ln\left(\frac{1 - P_1}{P_1} \right) + \ln\left[\frac{-\gamma' \Sigma_2^1}{\gamma' \Sigma_1^1} \right] + \ln\left[\frac{\gamma' \Sigma_1 \gamma}{\gamma' \Sigma_2 \gamma} \right].$$

Suppose that $\gamma' \Sigma_2 \gamma < \gamma' \Sigma_1 \gamma$. Then the left-hand side is positive except when $v = 0$. For any fixed $\gamma, \Sigma_1, \Sigma_2$ we can find a value of P_1 sufficiently small so that right-hand side of the equation is positive and for any such value of P_1 there will be a v sufficiently small for the inequality to be satisfied. There is also a value of v that reverses the inequality.

The inequality is satisfied for some $v^* \geqslant 0$. But with v arbitrarily large, the inequality can be reversed so that the weight will switch signs at some value of v. The key necessary condition is that $\operatorname{Cov}(Z_1, \gamma' Z)$ be of opposite signs in the two subpopulations. Using Z_1 as an IV, but not conditioning or controlling for the other components of Z, produces sometimes negative and sometimes positive movements in the components of $Z_2, \ldots, Z_k$ which can offset the *ceteris paribus* ($Z_2 = z_2, \ldots, Z_k = z_k$) movements of Z_1.

Appendix F: Local instrumental variables for the random coefficient model

Consider the model:

$$D = \mathbf{1}[Z\gamma \geqslant 0],$$

where γ is a random variable. For ease of exposition, we leave implicit the conditioning on X covariates. Assume that $(Y_0, Y_1, \gamma) \perp\!\!\!\perp Z$. Assume that γ has a density that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^K$. We have

$$E(Y \mid Z = z) = E(DY_1 \mid Z = z) + E\big((1 - D)Y_0 \bigm| Z = z\big).$$

To simplify the exposition, consider the first term, $E(DY_1 \mid Z = z)$. In this proof, let $Z^{[K]}$ denote the Kth element of Z and $Z^{[-K]}$ denote all other elements of Z, and write $Z = (Z^{[-K]}, Z^{[K]})$. Using the model, the independence assumption, and the law of iterated expectations, we have

$$\begin{aligned} E(DY \mid Z = z) &= E\big(\mathbf{1}[z\gamma \geqslant 0]Y_1\big) = E\big(\mathbf{1}[z\gamma \geqslant 0]E(Y_1 \mid \gamma)\big) \\ &= E\big(\mathbf{1}\big\{z^{[K]}\gamma^{[K]} \geqslant -z^{[-K]}\gamma^{[-K]}\big\}\, E(Y_1 \mid \gamma)\big), \end{aligned}$$

where the final outer expectation is over γ. Consider taking the derivative with respect to the Kth element of Z assumed to be continuous. Partition z, γ, and g as $z = (z^{[-K]}, z^{[K]})$, $\gamma = (\gamma^{[-K]}, \gamma^{[K]})$, and $g = (g^{[-K]}, g^{[K]})$, where z is a realization of Z and g is a realization of γ. For simplicity, suppose that the Kth element of z is positive, $z^{[K]} > 0$. We obtain

$$\begin{aligned} E(DY \mid Z = z) &= E\big[E\big(\mathbf{1}\big\{z^{[K]}\gamma^{[K]} \geqslant -z^{[-K]}\gamma^{[-K]}\big\}E(Y_1 \mid \gamma) \bigm| \gamma^{[-K]}\big)\big] \\ &= E\left[E\left(\mathbf{1}\left\{\gamma^{[K]} \geqslant \frac{-z^{[-K]}\gamma^{[-K]}}{z^{[K]}}\right\}E(Y_1 \mid \gamma) \biggm| \gamma^{[-K]}\right)\right], \end{aligned}$$

where the inside expectation is over $\gamma^{[K]}$ conditional on $\gamma^{[-K]}$, i.e., is over the Kth element of γ conditional on all other components of γ. Computing the derivative with respect to $z^{[K]}$, we obtain

$$\frac{\partial}{\partial z^{[K]}} E(DY \mid Z = z) = \int E\big(Y_1 \bigm| \gamma = M\big(g^{[-K]}\big)\big)\tilde{w}\big(g^{[-K]}\big)\, dg^{[-K]},$$

where

$$M\big(g^{[-K]}\big) = \left(\big(g^{[-K]}\big)', \frac{-z^{[-K]}g^{[-K]}}{z^{[K]}}\right)' \quad \text{and}$$

$$\tilde{w}\big(g^{[-K]}\big) = \frac{z^{[-K]}g^{[-K]}}{(z^{[K]})^2} f\left(g^{[-K]}, \frac{-z^{[-K]}g^{[-K]}}{z^{[K]}}\right),$$

with $f(\cdot)$ the density of γ (with respect to Lebesgue measure), and where for notational simplicity we suppress the dependence of the function $M(\cdot)$ and the weights $\tilde{w}(\cdot)$ on the z evaluation point. In this expression, we are averaging over $E(Y_1 \mid \gamma = g)$, but only over g evaluation points such that $zg = 0$. In particular, the expression averages over the $K-1$ space of $g^{[-K]}$, while for each potential realization of $g^{[-K]}$ it is filling in the value of $g^{[K]}$ such that $z^{[K]}g^{[K]} = -z^{[-K]}g^{[-K]}$ so that $z^{[K]}g^{[K]}+z^{[-K]}g^{[-K]} = 0$. Note that the weights $\tilde{w}(g^{[-K]})$ will be zero for any $g^{[-K]}$ such that $f(g^{[-K]}, \frac{-z^{[-K]}g^{[-K]}}{z^{[K]}}) = 0$, i.e., the weights will be zero for any $g^{[-K]}$ such that there does not exist $g^{[K]}$ in the conditional support of $\gamma^{[K]}$ with $z^{[K]}g^{[K]} = -z^{[-K]}g^{[-K]}$.

Following the same logic for $E((1-D)Y_0 \mid Z = z)$, we obtain

$$\frac{\partial}{\partial z^{[K]}} E\big((1-D)Y \mid Z = z\big) = -\int E\big(Y_0 \mid \gamma = M\big(g^{[-K]}\big)\big)\tilde{w}\big(g^{[-K]}\big)\, dg^{[-K]}$$

and likewise have

$$\frac{\partial}{\partial z^{[K]}} \Pr(D = 1 \mid Z = z) = \int \tilde{w}\big(g^{[-K]}\big)\, dg^{[-K]}$$

so that

$$\frac{\frac{\partial}{\partial z^{[K]}} E(Y \mid Z = z)}{\frac{\partial}{\partial z^{[K]}} \Pr(D = 1 \mid Z = z)} = \int E\big(Y_1 - Y_0 \mid \gamma = M\big(g^{[-K]}\big)\big) w\big(g^{[-K]}\big)\, dg^{[-K]},$$

where

$$w\big(g^{[-K]}\big) = \tilde{w}\big(g^{[-K]}\big) \Big/ \int \tilde{w}\big(g^{[-K]}\big)\, dg^{[-K]}.$$

Now consider the question of whether this expression will have both positive and negative weights. Recall that $\tilde{w}(g^{[-K]}) = \frac{z^{[-K]}g^{[-K]}}{(z^{[K]})^2} f(g^{[-K]}, \frac{-z^{[-K]}g^{[-K]}}{z^{[K]}})$. Thus,

$$\tilde{w}\big(g^{[-K]}\big) \geqslant 0 \quad \text{if } z^{[-K]}g^{[-K]} > 0, \qquad \tilde{w}\big(g^{[-K]}\big) \leqslant 0 \quad \text{if } z^{[-K]}g^{[-K]} < 0,$$

and will be nonzero if $z^{[-K]}g^{[-K]} \neq 0$ and there exists $g^{[K]}$ in the conditional support of $\gamma^{[K]}$ with $z^{[K]}g^{[K]} = z^{[-K]}g^{[-K]}$, i.e., with $zg = 0$. We thus have that there will be both positive and negative weights on the MTE if there exist values of g in the support of γ with both $z^{[-K]}g^{[-K]} > 0$ and $zg = 0$, and there exist other values of g in the support of γ with $z^{[-K]}g^{[-K]} < 0$ and $zg = 0$.

Appendix G: Generalized ordered choice model with stochastic thresholds

The ordered choice model presented in the text with parameterized, but nonstochastic, thresholds is analyzed in Cameron and Heckman (1998) who establish its nonparametric identifiability under the conditions they specify. Treating the W_s (or components of it) as unobservables, we obtain the generalized ordered choice model analyzed in Carneiro, Hansen and Heckman (2003) and Cunha, Heckman and Navarro (2007). In this appendix, we present the main properties of this more general model.

The thresholds are now written as $Q_s + C_s(W_s)$ in place of $C_s(W_s)$, where Q_s is a random variable. In addition to the order on the $C_s(W_s)$ in the text, we impose the order $Q_s + C_s(W_s) \geqslant Q_{s-1} + C_{s-1}(W_{s-1})$, $s = 2, \ldots, \bar{S} - 1$. We impose the requirement that $Q_{\bar{S}} = \infty$ and $Q_0 = -\infty$. The latent index D_s^* is as defined in the text, but now

$$\begin{aligned} D_s &= \mathbf{1}\big[C_{s-1}(W_{s-1}) + Q_{s-1} < \mu_D(Z) - V \leqslant C_s(W_s) + Q_s\big] \\ &= \mathbf{1}\big[l_{s-1}(Z, W_{s-1}) - Q_{s-1} > V \geqslant l_s(Z, W_s) - Q_s\big], \end{aligned}$$

where $l_s = \mu_D(Z) - C_s(W_s)$. Using the fact that $l_s(Z, W_s) - Q_s < l_{s-1}(Z, W_{s-1}) - Q_{s-1}$, we obtain

$$\begin{aligned} &\mathbf{1}\big[l_{s-1}(Z, W_{s-1}) - Q_{s-1} > V \geqslant l_s(Z, W_s) - Q_s\big] \\ &\quad = \mathbf{1}\big[V + Q_{s-1} < l_{s-1}(Z, W_{s-1})\big] - \mathbf{1}\big[V + Q_s \leqslant l_s(Z, W_s)\big]. \end{aligned}$$

The nonparametric identifiability of this choice model is established in Carneiro, Hansen and Heckman (2003) and Cunha, Heckman and Navarro (2007). We retain assumptions (OC-2)–(OC-6), but alter (OC-1) to

(OC-1)′ $(Q_s, U_s, V) \perp\!\!\!\perp (Z, W) \mid X, s = 1, \ldots, \bar{S}$.

Vytlacil (2006b) shows that this model with no transition specific instruments (with W_s degenerate for each s) implies and is implied by the independence and monotonicity conditions of Angrist and Imbens (1995) for an ordered model. Define $Q = (Q_1, \ldots, Q_{\bar{S}})$. Redefine $\pi_s(Z, W_s) = F_{V+Q_s}(\mu_D(Z) + C_s(W_s))$ and define $\pi(Z, W) = [\pi_1(Z, W_1), \ldots, \pi_{\bar{S}-1}(Z, W_{\bar{S}-1})]$. Redefine $U_{D,s} = F_{V+Q_s}(V + Q_s)$. We have that

$$\begin{aligned} &E(Y \mid Z, W) \\ &\quad = E\left(\sum_{s=1}^{\bar{S}} \mathbf{1}\big[l_{s-1}(Z, W_{s-1}) - Q_{s-1} > V \geqslant l_s(Z, W_s) - Q_s\big] Y_s \;\middle|\; Z, W\right) \\ &\quad = \sum_{s=1}^{\bar{S}} \Big(E\big(\mathbf{1}\big[V + Q_{s-1} < l_{s-1}(Z, W_{s-1})\big] Y_s \mid Z, W\big) \\ &\qquad - E\big(\mathbf{1}\big[V + Q_s \leqslant l_s(Z, W_s)\big] Y_s \mid Z, W\big)\Big) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=1}^{\bar{S}} \biggl(\int_{-\infty}^{l_{s-1}(Z,W_{s-1})} E(Y_s \mid V + Q_{s-1} = t)\, dF_{V+Q_{s-1}}(t) \\
&\quad - \int_{-\infty}^{l_s(Z,W_s)} E(Y_s \mid V + Q_s = t)\, dF_{V+Q_s}(t) \biggr) \\
&= \sum_{s=1}^{\bar{S}} \biggl(\int_{0}^{\pi_{s-1}(Z,W_{s-1})} E(Y_s \mid U_{D,s-1} = t)\, dt \\
&\quad - \int_{0}^{\pi_s(Z,W_s)} E(Y_s \mid U_{D,s} = t)\, dt \biggr).
\end{aligned}
$$

We thus have the index sufficiency restriction that $E(Y \mid Z, W) = E(Y \mid \pi(Z, W))$, and in the general case $\frac{\partial}{\partial \pi_s} E(Y \mid \pi(Z, W) = \pi) = E(Y_{s+1} - Y_s \mid U_{D,s} = \pi_s)$. Also, notice that we have the restriction that $\frac{\partial^2}{\partial \pi_s \partial \pi_{s'}} E(Y \mid \pi(Z, W) = \pi) = 0$ if $|s - s'| > 1$. Under full independence between U_s and $V + Q_s$, $s = 1, \ldots, \bar{S}$, we can test full independence for the more general choice model by testing for linearity of $E(Y \mid \pi(Z, W) = \pi)$ in π.

Define

$$
\Delta^{\mathrm{MTE}}_{s+1,s}(x, u) = E(Y_{s+1} - Y_s \mid X = x, U_{D,s} = u),
$$

so that our result above can be rewritten as

$$
\frac{\partial}{\partial \pi_s} E\bigl(Y \bigm| \pi(Z, W) = \pi\bigr) = \Delta^{\mathrm{MTE}}_{s+1,s}(x, \pi_s).
$$

Since $\pi_s(Z, W_s)$ can be nonparametrically identified from

$$
\pi_s(Z, W_s) = \Pr\biggl(\sum_{j=s+1}^{\bar{S}} D_j = 1 \biggm| Z, W_s \biggr),
$$

we have identification of MTE for all evaluation points within the appropriate support.

The policy relevant treatment effect is defined analogously. H_s^p is defined as the cumulative distribution function of $\mu_D(Z) - C_s(W_s)$. We have that

$$
\begin{aligned}
&E_p(Y_p) \\
&\quad = E_p\bigl(E(Y \mid V, Q, Z, W)\bigr) \\
&\quad = E_p\Biggl(\sum_{s=1}^{\bar{S}} \mathbf{1}\bigl[l_{s-1}(Z, W_{s-1}) - Q_{s-1} > V \geqslant l_s(Z, W_s) - Q_s\bigr] \\
&\qquad\qquad \times E(Y_s \mid V, Q, Z, W) \Biggr)
\end{aligned}
$$

$$
\begin{aligned}
&= E_p\left(\sum_{s=1}^{\bar{S}} \mathbf{1}\big[l_{s-1}(Z, W_{s-1}) - Q_{s-1} > V \geqslant l_s(Z, W_s) - Q_s\big] E(Y_s \mid V, Q)\right) \\
&= \sum_{s=1}^{\bar{S}} E_p\big(E(Y_s \mid V, Q)\big\{H_s^p(V + Q_s) - H_{s-1}^p(V + Q_{s-1})\big\}\big) \\
&= \sum_{s=1}^{\bar{S}} \int \big(E(Y_s \mid V = v, Q = q)\big\{H_s^p(v + q_s) \\
&\quad - H_{s-1}^p(v + q_{s-1})\big\}\big)\, dF_{V,Q}(v, q) \\
&= \sum_{s=1}^{\bar{S}} \left(\int E(Y_s \mid V + Q_s = t) H_s^p(t)\, dF_{V+Q_s}(t) \right. \\
&\quad \left. - \int E(Y_s \mid V + Q_{s-1} = t) H_{s-1}^p(t)\, dF_{V+Q_{s-1}}(t)\right),
\end{aligned}
$$

where V, Q_s enter additively, and

$$
\begin{aligned}
\Delta_{p,p'}^{\text{PRTE}} &= E_{p'}(Y) - E_p(Y) \\
&= \sum_{s=1}^{\bar{S}-1} \int \big(E(Y_{s+1} - Y_s \mid V + Q_s = t)\big\{H_s^p(t) - H_s^{p'}(t)\big\}\big)\, dF_{V+Q_s}(t).
\end{aligned}
$$

Alternatively, we can express this result in terms of MTE,

$$
\begin{aligned}
E_p(Y_p) = \sum_{s=1}^{\bar{S}} &\left(\int E(Y_s \mid U_{D,s} = t) \tilde{H}_s^p(t)\, dt \right. \\
&\left. - \int E(Y_s \mid U_{D,s-1} = t) \tilde{H}_{s-1}^p(t)\, dt\right)
\end{aligned}
$$

so that

$$
\begin{aligned}
\Delta_{p,p'}^{\text{PRTE}} &= E_{p'}(Y) - E_p(Y) \\
&= \sum_{s=1}^{\bar{S}-1} \int \big(E(Y_{s+1} - Y_s \mid U_{D,s} = t)\big\{\tilde{H}_s^p(t) - \tilde{H}_s^{p'}(t)\big\}\big)\, dt,
\end{aligned}
$$

where $\tilde{H}_s^p$ is the cumulative distribution function of the random variable $F_{U_{D,s}}(\mu_D(Z) - C_s(W_s))$.

Appendix H: Derivation of PRTE weights for the ordered choice model

To derive the $\omega_{p,p'}$ weights used in expression (7.5), let $l_s(Z, W_s) = \mu_D(Z) - C_s(W_s)$, and let $H_s^p(\cdot)$ denote the cumulative distribution function of $l_s(Z, W_s)$ under regime p,

$H_s^p(t) = \int \mathbf{1}[\mu_D(z) - C_s(w_s) \leqslant t]\, dF^p_{Z,W}(z,w)$. Because $C_0(W_0) = -\infty$ and $C_{\bar{S}}(W_{\bar{S}}) = \infty$, $l_0(Z, W_0) = \infty$ and $l_{\bar{S}}(Z, W_{\bar{S}}) = -\infty$, $H_0^p(t) = 0$ and $H_{\bar{S}}^p(t) = 1$ for any policy p and for all evaluation points. Since $l_{s-1}(Z, W_{s-1})$ is always larger than $l_s(Z, W_s)$, we obtain

$$\begin{aligned} &\mathbf{1}\bigl[l_s(Z, W_s) \leqslant V < l_{s-1}(Z, W_{s-1})\bigr] \\ &\quad = \mathbf{1}\bigl[V < l_{s-1}(Z, W_{s-1})\bigr] - \mathbf{1}\bigl[V \leqslant l_s(Z, W_s)\bigr], \end{aligned}$$

so that under assumption (OC-1),

$$E_p\bigl(\mathbf{1}\bigl[l_s(Z, W_s) \leqslant V < l_{s-1}(Z, W_{s-1})\bigr] \bigm| V\bigr) = H_s^p(V) - H_{s-1}^p(V).$$

Collecting these results we obtain

$$\begin{aligned} E_p(Y) &= E_p\bigl[E(Y \mid V, Z, W)\bigr] \\ &= \sum_{s=1}^{\bar{S}} \int \bigl[E(Y_s \mid V = v)\bigl\{H_s^p(v) - H_{s-1}^p(v)\bigr\}\bigr] f_V(v)\, dv. \end{aligned}$$[200]

Comparing two policies under p and p', the policy relevant treatment effect is $\Delta^{\text{PRTE}}_{p,p'} = E_{p'}(Y) - E_p(Y) = \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s \mid V = v)[H_s^p(v) - H_s^{p'}(v)] f_V(v)\, dv$. Alternatively, we can express this in terms of Δ^{MTE}: $\Delta^{\text{PRTE}}_{p,p'} = \sum_{s=1}^{\bar{S}-1} \int \Delta^{\text{MTE}}_{s,s+1}(u)[\tilde{H}_s^p(u) - \tilde{H}_s^{p'}(u)]\, du$ where $\tilde{H}_s^p(t)$ is the cumulative distribution function of $F_V(\mu_D(Z) - C_s(W_s))$ under policy p, $\tilde{H}_s^p(t) = \int \mathbf{1}[F_V(\mu_D(z) - C_s(w_s)) \leqslant t]\, dF^p_{Z,W_s}(z, w_s)$.

Appendix I: Derivation of the weights for IV in the ordered choice model

We first derive $\operatorname{Cov}(J(Z, W), Y)$. Its derivation is typical of the other terms needed to form (7.6) in the text. Defining $\tilde{J}(Z, W) = J(Z, W) - E(J(Z, W))$, we obtain, since $\operatorname{Cov}(J(Z, W), Y) = E(\tilde{J}(Z, W)Y)$,

[200] The full derivation is $E_p(Y) = E_p[E(Y|V, Z, W)] = E_p[\sum_{s=1}^{\bar{S}} \mathbf{1}[l_s(Z, W_s) \leqslant V < l_{s-1}(Z, W_{s-1})] \times E(Y_s|V, Z, W)] = \sum_{s=1}^{\bar{S}} E_p[\mathbf{1}[l_s(Z, W_s) \leqslant V < l_{s-1}(Z, W_{s-1})] E(Y_s|V)] = \sum_{s=1}^{\bar{S}} E_p[E(Y_s|V) \times \{H_s^p(V) - H_{s-1}^p(V)\}] = \sum_{s=1}^{\bar{S}} \int [E(Y_s|V = v)\{H_s^p(v) - H_{s-1}^p(v)\}] f_V(v)\, dv$. The first equality is from the law of iterated expectations; the second equality comes from the definition of Y; the third equality follows from linearity of expectations and independence assumption (OC-1); the fourth equality applies the law of iterated expectations; and the final equality rewrites the expectation explicitly as an integral over the distribution of V. Recalling that $H_0^p(v) = 0$ and $H_{\bar{S}}^p(v) = 1$, we may rewrite this result as $E_p(Y) = \sum_{s=1}^{\bar{S}-1} \int E(Y_s - Y_{s+1} \mid V = v) H_s^p(v) f_V(v)\, dv + \int E(Y_{\bar{S}} \mid V = v) f_V(v)\, dv$, where the last term is $E(Y_{\bar{S}})$.

$$
\begin{aligned}
&E\big(\tilde{J}(Z,W)Y\big)\\
&\quad= E\left[\tilde{J}(Z,W)\sum_{s=1}^{\bar{S}}\mathbf{1}\big[l_s(Z,W_s)\leqslant V<l_{s-1}(Z,W_{s-1})\big]E(Y_s\mid V,Z,W)\right]\\
&\quad= \sum_{s=1}^{\bar{S}} E\big[\tilde{J}(Z,W)\mathbf{1}\big[l_s(Z,W_s)\leqslant V<l_{s-1}(Z,W_{s-1})\big]E(Y_s\mid V)\big],
\end{aligned}
$$

where the first equality comes from the definition of Y and the law of iterated expectations, and the second equality follows from linearity of expectations and independence assumption (OC-1). Let $H_s(\cdot)$ equal $H_s^p(\cdot)$ for p equal to the policy that characterizes the observed data, i.e., $H_s(\cdot)$ is the cumulative distribution function of $l_s(Z,W_s)$,

$$
H_s^p(t)=\Pr\big(l_s(Z,W_s)\leqslant t\big)=\Pr\big(\mu_D(Z)-C_s(W_s)\leqslant t\big).
$$

Using the law of iterated expectations, we obtain

$$
\begin{aligned}
E\big(\tilde{J}(Z,W)Y\big) &= \sum_{s=1}^{\bar{S}} E\big[E\big(\tilde{J}(Z,W)\big\{\mathbf{1}\big[V<l_{s-1}(Z,W_{s-1})\big]\\
&\qquad -\mathbf{1}\big[V\leqslant l_s(Z,W_s)\big]\big\}\mid V\big)E(Y_s\mid V)\big]\\
&= \sum_{s=1}^{\bar{S}}\int\big[E(Y_s\mid V=v)\big\{K_{s-1}(v)-K_s(v)\big\}\big]f_V(v)\,dv\\
&= \sum_{s=1}^{\bar{S}-1}\int\big[E(Y_{s+1}-Y_s\mid V=v)K_s(v)\big]f_V(v)\,dv,
\end{aligned}
$$

where $K_s(v)=E(\tilde{J}(Z,W)\mid l_s(Z,W_s)>v)(1-H_s(v))$ and we use the fact that $K_{\bar{S}}(v)=K_0(v)=0$. Now consider the denominator of the IV estimand,

$$
\begin{aligned}
&E\big(S\tilde{J}(Z,W)\big)\\
&\quad= E\left[\tilde{J}(Z,W)\sum_{s=1}^{\bar{S}}s\mathbf{1}\big[l_s(Z,W_s)\leqslant V<l_{s-1}(Z,W_{s-1})\big]\right]\\
&\quad= \sum_{s=1}^{\bar{S}}sE\big[\tilde{J}(Z,W)\mathbf{1}\big[l_s(Z,W_s)\leqslant V<l_{s-1}(Z,W_{s-1})\big]\big]\\
&\quad= \sum_{s=1}^{\bar{S}}sE_V\big[E\big(\tilde{J}(Z,W)\big\{\mathbf{1}\big[V<l_{s-1}(Z,W_{s-1})\big]-\mathbf{1}\big[V\leqslant l_s(Z,W_s)\big]\big\}\mid V\big)\big]\\
&\quad= \sum_{s=1}^{\bar{S}}s\int\big[K_{s-1}(v)-K_s(v)\big]f_V(v)\,dv=\sum_{s=1}^{\bar{S}-1}\int K_s(v)f_V(v)\,dv.
\end{aligned}
$$

Collecting results, we obtain an expression for the IV estimand (7.6):

$$\frac{\operatorname{Cov}(J, Y)}{\operatorname{Cov}(J, S)} = \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s \mid V = v)\omega(s, v) f_V(v)\, dv,$$

where

$$\omega(s, v) = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}} s \int [K_{s-1}(v) - K_s(v)] f_V(v)\, dv} = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}-1} \int K_s(v) f_V(v)\, dv}$$

and clearly

$$\sum_{s=1}^{\bar{S}-1} \int \omega(s, v) f_V(v)\, dv = 1, \quad \omega(0, v) = 0, \quad \text{and} \quad \omega(\bar{S}, v) = 0.$$

Appendix J: Proof of Theorem 6

We now prove Theorem 6.

PROOF. The basic idea is that we can bring the model back to a two choice set up of j versus the "next best" option. We prove the result for the second assertion, that $\Delta_j^{\mathrm{LIV}}(x, z)$ recovers the marginal treatment effect parameter. The first assertion, that $\Delta_j^{\mathrm{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ recovers a LATE parameter, follows from a trivial modification to the same proof strategy. Recall that $R_{\mathcal{J}\setminus j}(z) = \max_{i \in \mathcal{J}\setminus j}\{R_i(z)\}$ and that $I_{\mathcal{J}\setminus j} = \operatorname{argmax}_{i \in \mathcal{J}\setminus j}(R_i(Z))$. We may write $Y = Y_{I_{\mathcal{J}\setminus j}} + D_{\mathcal{J},j}(Y_j - Y_{I_{\mathcal{J}\setminus j}})$. We have

$$\begin{aligned} &\Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = z) \\ &\quad = \Pr\bigl(R_j(z_j) \geqslant R_{\mathcal{J}\setminus j}(z) \bigm| X = x, Z = z\bigr) \\ &\quad = \Pr\bigl(\vartheta_j(z_j) \geqslant R_{\mathcal{J}\setminus j}(z) + V_j \bigm| X = x, Z = z\bigr). \end{aligned}$$

Using independence assumption (B-1), $R_{\mathcal{J}\setminus j}(z) - V_j$ is independent of Z conditional on X, so that

$$\Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = z) = \Pr\bigl(\vartheta_j(z_j) \geqslant R_{\mathcal{J}\setminus j}(z) + V_j \bigm| X = x\bigr).$$

$\vartheta_k(\cdot)$ does not depend on $z^{[j]}$ for $k \neq j$ by assumption (B-2b), and thus $R_{\mathcal{J}\setminus j}(z)$ does not depend on $z^{[j]}$, and we will therefore (with an abuse of notation) write $R_{\mathcal{J}\setminus j}(z^{[-j]})$ for $R_{\mathcal{J}\setminus j}(z)$. Write $F_{X|Z^{[-j]}}(\cdot; X = x, Z^{[-j]} = z^{[-j]})$ for the distribution function of $R_{\mathcal{J}\setminus j}(z^{[-j]}) + V_j$ conditional on $X = x$. Then

$$\Pr(D_{\mathcal{J},j} = 1 \mid X = x, Z = z) = F\bigl(\vartheta_j(z_j); x, z^{[-j]}\bigr),$$

and

$$\begin{aligned}&\frac{\partial}{\partial z^{[j]}}\Pr(D_{\mathcal{J},j}=1\mid X=x, Z=z)\\&\quad=\left[\frac{\partial}{\partial z^{[j]}}\vartheta_j(z_j)\right]f_{X|Z^{[-j]}}\big(\vartheta_j(z_j);\, X=x,\, Z^{[-j]}=z^{[-j]}\big),\end{aligned}$$

where $f_{X|Z^{[-j]}}(\cdot;\, X=x,\, Z^{[-j]}=z^{[-j]})$ is the density of $R_{\mathcal{J}\setminus j}(z^{[-j]})-V_j$ conditional on $X=x$. Consider

$$\begin{aligned}E(Y\mid X=x, Z=z)&=E(Y_{I_{\mathcal{J}\setminus j}}\mid X=x, Z=z)\\&\quad+E\big(D_{\mathcal{J},j}(Y_j-Y_{I_{\mathcal{J}\setminus j}})\;\big|\; X=x, Z=z\big).\end{aligned}$$

As a consequence of (B-1), (B-3)–(B-5), and (B-2b), we have that $E(Y_{I_{\mathcal{J}\setminus j}}\mid X=x, Z=z)$ does not depend on $z^{[j]}$. Using the assumptions and the law of iterated expectations, we may write

$$\begin{aligned}&E\big(D_{\mathcal{J},j}(Y_j-Y_{I_{\mathcal{J}\setminus j}})\;\big|\; X=x, Z=z\big)\\&\quad=\int_{-\infty}^{\vartheta_j(z)}E\big(Y_j-Y_{I_{\mathcal{J}\setminus j}}\;\big|\; X=x, Z=z, R_{\mathcal{J}\setminus j}\big(z^{[-j]}\big)+V_j=t\big)\\&\qquad\times f_{X|Z^{[-j]}}\big(t;\, X=x,\, Z^{[-j]}=z^{[-j]}\big)\,dt\\&\quad=\int_{-\infty}^{\vartheta_j(z)}E\big(Y_j-Y_{I_{\mathcal{J}\setminus j}}\;\big|\; X=x, Z^{[-j]}=z^{[-j]}, R_{\mathcal{J}\setminus j}\big(z^{[-j]}\big)+V_j=t\big)\\&\qquad\times f_{X|Z^{[-j]}}\big(t;\, X=x,\, Z^{[-j]}=z^{[-j]}\big)\,dt.\end{aligned}$$

Thus,

$$\begin{aligned}&\frac{\partial}{\partial z^{[j]}}E(Y\mid X=x, Z=z)\\&\quad=E\big(Y_j-Y_{I_{\mathcal{J}\setminus j}}\;\big|\; X=x, Z^{[-j]}=z^{[-j]}, R_j(z)=R_{\mathcal{J}\setminus j}(z)\big)\\&\qquad\times\left[\frac{\partial}{\partial z^{[j]}}\vartheta_j(z_j)\right]f_{X|Z^{[-j]}}\big(\vartheta_j(z_j)\mid X=x,\, Z^{[-j]}=z^{[-j]}\big).\end{aligned}$$

Combining results, we have

$$\begin{aligned}&\frac{\partial}{\partial z^{[j]}}E(Y\mid X=x, Z=z)\Big/\frac{\partial}{\partial z^{[j]}}\Pr(D_{\mathcal{J},j}=1\mid X=x, Z=z)\\&\quad=E\big(Y_j-Y_{I_{\mathcal{J}\setminus j}}\;\big|\; X=x, Z^{[-j]}=z^{[-j]}, R_j(z)=R_{\mathcal{J}\setminus j}(z)\big).\end{aligned}$$

Finally, noting that

$$\begin{aligned}&E\big(Y_j-Y_{I_{\mathcal{J}\setminus j}}\;\big|\; X=x, Z^{[-j]}=z^{[-j]}, R_j(z)=R_{\mathcal{J}\setminus j}(z)\big)\\&\quad=E\big(Y_j-Y_{I_{\mathcal{J}\setminus j}}\;\big|\; X=x, Z=z, R_j(z)=R_{\mathcal{J}\setminus j}(z)\big)\end{aligned}$$

provides the stated result. The proof for the LATE result follows from the parallel argument using discrete changes in the instrument. □

Appendix K: Flat MTE within a general nonseparable matching framework

The result in the text that conditional mean independence of Y_0 and Y_1 in terms of D given X implies a flat MTE holds in a more general nonseparable model. We establish this claim and also establish some additional restrictions implied by an IV assumption.

Assume a nonseparable selection model, $D = \mathbf{1}[\mu_D(X, Z, V) \geqslant 0]$, with Z independent of (Y_0, Y_1, V) conditional on X. Let $\Omega(x, z) = \{v\colon \mu_D(x, z, v) \geqslant 0\}$. Let $\Omega(x, z)^c$ denote the complement of $\Omega(x, z)$. Consider the mean independence assumption

(M-3) $E(Y_1 \mid X, D) = E(Y_1 \mid X), E(Y_0 \mid X, D) = E(Y_0 \mid X)$.

(M-3) implies that for $\Delta = Y_1 - Y_0$

$$E\bigl(\Delta \bigm| X = x, V \in \Omega(X, Z)\bigr) = E\bigl(\Delta \bigm| X = x, V \in \Omega(X, Z)^c\bigr),$$

where c here denotes "complement". Thus,

$$\begin{aligned}&E_{Z|X}\bigl(E\bigl(\Delta^{\mathrm{MTE}}(x, V) \bigm| X = x, V \in \Omega(x, Z)\bigr) \bigm| X = x\bigr)\\&\quad = E_{Z|X}\bigl(E\bigl(\Delta^{\mathrm{MTE}}(x, V) \bigm| X = x, V \in \Omega(x, Z)^c\bigr) \bigm| X = x\bigr)\end{aligned}$$

for all x in the support of X. (We assume $0 < \Pr(D = 1 \mid X) < 1$.) This establishes that the MTE is flat.

Now suppose that (M-3) holds, but suppose that there is an instrument Z such that

(M-3)′ $E(Y_1 \mid X, Z, D) \neq E(Y_1 \mid X), E(Y_0 \mid X, Z, D) \neq E(Y_0 \mid X)$.

(*Note*: $E(Y_j \mid X, Z) = E(Y_j \mid X)$ by assumption.) In this case, (M-3) implies that

$$\begin{aligned}&E_{Z|X}\bigl(E\bigl(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x, V \in \Omega(x, Z)\bigr) \bigm| X = x\bigr)\\&\quad = E_{Z|X}\bigl(E\bigl(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x, V \in \bigl(\Omega(x, Z)\bigr)^c\bigr) \bigm| X = x\bigr),\end{aligned}$$

but (M-3)′ implies that there exists z in the support of Z conditional on X such that

$$E\bigl(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x, V \in \Omega(x, z)\bigr) \neq E\bigl(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x\bigr)$$

and

$$E\bigl(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x, V \in \Omega(x, z)^c\bigr) \neq E\bigl(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x\bigr)$$

so that $\Delta^{\mathrm{MTE}}(X, V)$ is not constant in V. Note that, if $E(Y_1 \mid X, Z = z, D = 1) \neq E(Y_1 \mid X, Z = z', D = 1)$ for any z, z' evaluation points in the support of Z conditional on X, then $E(Y_1 \mid X, Z, D) \neq E(Y_1 \mid X)$. Thus, (M-3)′ is testable, given the maintained assumption that Z is a proper exclusion restriction. Note that (M-3)′ implies (M-3), so it is a stronger condition.

Now assume

(M-1)′ $E(Y_1 \mid X, Z, D) = E(Y_1 \mid X),\ E(Y_0 \mid X, Z, D) = E(Y_0 \mid X)$.

In this case, we get a stronger restriction on MTE than is produced from (M-3). We obtain

$$E\big(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x, V \in \Omega(x, z)\big) = E\big(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x\big)$$

and

$$E\big(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x, V \in \Omega(x, z)^c\big) = E\big(\Delta^{\mathrm{MTE}}(X, V) \bigm| X = x\big)$$

for all (x, z) in the proper support. Again, the MTE is not flat.

Appendix L: The relationship between exclusion conditions in IV and exclusion conditions in matching

We now investigate the relationship between IV and matching identification conditions. They are very distinct. We analyze mean treatment parameters. We define (U_0, U_1) by $U_0 = Y_0 - E(Y_0 \mid X)$ and $U_1 = Y_1 - E(Y_1 \mid X)$. We consider standard IV as a form of matching where matching does not hold conditional on X but does hold conditional on (X, Z), where Z is the instrument. Consider the following two matching conditions based on an exclusion restriction Z:

(M-4) *(U_0, U_1) are mean independent of D conditional on (X, Z). ($E(U_0|X, Z, D) = E(U_0 \mid X, Z)$ and $E(U_1 \mid X, Z, D) = E(U_1 \mid X, Z)$.)*

(M-5) *(U_0, U_1) are not mean independent of D conditional on X. ($E(U_0 \mid X, D) \neq E(U_0 \mid X)$ and $E(U_1 \mid X, D) \neq E(U_1 \mid X)$.)*

(M-4) says that the matching conditions hold conditional on (X, Z). However, (M-5) says that the matching conditions do not hold if one only conditions on X. By the definitions of U_0, U_1, these conditions are equivalent to stating that Y_0, Y_1 are mean independent of D conditional on (X, Z) but not mean independent of D conditional on X. These look like instrumental variable conditions. We now consider whether these assumptions are compatible with standard IV conditions as used by Heckman and Robb (1985a, 1986a) and Heckman (1997) to use IV to identify treatment parameters when responses are heterogenous (the model of essential heterogeneity). For ATE, they show that standard IV identifies ATE if:

(ATE-1) *U_0 is mean independent of Z conditional on X.*

(ATE-2) *$D(U_1 - U_0)$ is mean independent of Z conditional on X.*[201]

[201] When $Y = Y_0 + D(Y_1 - Y_0)$, assuming separability so that $Y_0 = \mu_0(X) + U_0$, $Y_1 = \mu_1(X) + U_1$, and $Y = \mu_0(X) + D(\mu_1(X) - \mu_0(X) + U_1 - U_0) + U_0$, identification of ATE by IV requires the rank condition (IV-2) plus $E(U_0 + D(U_1 - U_0) \mid X, Z) = E(U_0 + D(U_1 - U_0) \mid X)$, which is implied by (ATE-1) and (ATE-2).

They show that standard IV identifies TT if:

(TT-1) U_0 *is mean independent of* Z *conditional on* X.
(TT-2) $U_1 - U_0$ *is mean independent of* Z *conditional on* $D = 1$ *and on* X.[202]

The conventional assumption in means is that

(IV-1)′ (U_0, U_1) *are mean independent of* Z *conditional on* X.
(IV-2) *Rank condition* (IV-2) *is still required*: $\Pr(D = 1 \mid Z, X)$ *is a nondegenerate function of* Z.

Condition (IV-1)′ is a commonly invoked instrumental variable condition, even though Heckman and Robb (1986a) and Heckman (1997) show it is neither necessary nor sufficient to identify ATE or TT by linear IV. In Section 4, we used the stronger condition (IV-1): $(U_0, U_1) \perp\!\!\!\perp Z \mid X$ along with the rank conditions. Clearly, (IV-1) implies (IV-1)′.

We now show that assumptions (M-4) and (M-5) are inconsistent with any of the sets of IV assumptions. In particular, we show that assuming (M-4) and that U_0 is mean independent of Z conditional on X jointly imply that U_0 is mean independent of D conditional on X. If (M-4) and (M-5) hold, then Z cannot satisfy condition (IV-1)′ (or stronger condition (IV-1)), (ATE-1) or (TT-1). Thus matching based on an exclusion restriction and IV are distinct conditions. We show this by establishing a series of claims.

CLAIM 1. *Conditions* (M-4) *and* (IV-1)′ *jointly imply* U_0 *is mean independent of* D *conditional on* X. *Thus,* (M-4) *and* [(IV-1)′ *or* (ATE-1) *or* (TT-1)] *jointly imply that* (M-5) *cannot hold.*

PROOF. Assume (M-4) and (IV-1)′. We have

$$\begin{aligned} E(U_0 \mid D, X, Z) &= E(U_0 \mid X, Z) \\ &= E(U_0 \mid X), \end{aligned}$$

[202] In the separable model,

$$\begin{aligned} Y = \mu_0(X) + D\bigl(\overbrace{\mu_1(X) - \mu_0(X) + E(U_1 - U_0 \mid X, D = 1)}^{\Delta^{\mathrm{TT}}(X)}\bigr) \\ + U_0 + D\bigl(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1)\bigr). \end{aligned}$$

Identification requires that

$$\begin{aligned} &E\bigl(U_0 + D\bigl(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1)\bigr) \bigm| X, Z\bigr) \\ &\quad = E\bigl(U_0 + D\bigl(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1)\bigr) \bigm| X\bigr), \end{aligned}$$

which is implied by (TT-1) and (TT-2).

where the first equality follows from (M-4) and the second equality follows from (IV-1)′. Thus,

$$\begin{aligned} E(U_0 \mid D, X) &= E_Z\big[E(U_0 \mid D, X, Z) \bigm| D, X\big] \\ &= E_Z\big[E(U_0 \mid X) \bigm| D, X\big] \\ &= E(U_0 \mid X). \end{aligned}$$

□

Thus (M-4) and (M-5) are inconsistent with any of the sets of IV assumptions that we have considered. However, this analysis raises the question of whether it is still possible to invoke (M-5) and the assumption that U_1 is not mean independent of D conditional on X. The following results show that it is not possible.

CLAIM 2. (M-4) *and* (IV-1)′ *imply* U_1 *is mean independent of* D *conditional on* X.

PROOF. Follows with trivial modification from the proof to Claim 1. □

A similar claim can be shown for (TT-1) and (TT-2).

CLAIM 3. (M-4) *and* (TT-1), (TT-2) *imply* U_1 *is mean independent of* D *conditional on* X.

PROOF. Assume (M-4) and (TT-1), (TT-2). We have

(N-1) $E(U_0 \mid X, Z, D) = E(U_0 \mid X, Z) = E(U_0 \mid X)$,

where the first equality follows from (M-4) and the second equality follows from (TT-1). Using the result from the proof of Claim 1, we obtain

(N-2) $E(U_0 \mid X, Z, D) = E(U_0 \mid X, D)$.

By (TT-2), we have

$$\begin{aligned} &E(U_1 \mid X, Z, D = 1) - E(U_1 \mid X, D = 1) \\ &\quad = E(U_0 \mid X, Z, D = 1) - E(U_0 \mid X, D = 1). \end{aligned}$$

By equation (N-2), the right-hand side of the preceding expression is zero, and we thus have

(N-3) $E(U_1 \mid X, Z, D = 1) = E(U_1 \mid X, D = 1)$.

By (M-4), we have

(N-4) $E(U_1 \mid X, Z, D = 1) = E(U_1 \mid X, Z)$.

Combining equations (N-3) and (N-4), we obtain

$$E(U_1 \mid X, Z) = E(U_1 \mid X, D = 1).$$

Integrating both sides of this expression against the distribution of Z conditional on X, we obtain

$$E(U_1 \mid X) = E(U_1 \mid X, D = 1).$$

□

It is straightforward to show that (M-4) and (ATE-1), (ATE-2) jointly imply that U_1 is mean independent of D conditional on X.

In summary, (U_0, U_1) mean independent of D conditional on (X, Z) but not conditional on X implies that U_0 is dependent on Z conditional on X in contradiction to all of the assumptions used to justify instrumental variables. Thus (U_0, U_1) mean independent of D conditional on (X, Z) but not conditional on X implies that none of the three sets of IV conditions will hold. In addition, if we weaken these conditions to only consider U_1, so that we assume that U_1 is mean independent of D conditional on (X, Z) but not conditional on X, we obtain that U_1 is dependent on Z conditional on X. We have shown that this implies that (IV-1) does not hold, and implies that (TT-1), (TT-2) will not hold. A similar line of argument shows that (ATE-1), (ATE-2) will not hold. Thus, the exclusion conditioning in matching is not the same as the exclusion conditioning in IV.

Appendix M: Selection formulae 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 $D^* \geqslant 0$, $= 0$ otherwise; and $Y = DY_1 + (1 - D)Y_0$, where

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

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

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

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

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

The event $(V \lesseqgtr 0, Z = z)$ can be written as $\frac{V}{\sigma_V} \lesseqgtr -\frac{\mu_D(z)}{\sigma_V} \Leftrightarrow \frac{V}{\sigma_V} \lesseqgtr \Phi^{-1}(1 - P(z))$. 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:

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

where

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

Similarly for U_0

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

where $\eta_0 = \frac{\sigma_{0V}}{\sigma_V}$ and

$$M_1(P(z)) = \frac{\phi(\Phi^{-1}(1 - P(z)))}{P(z)} \quad \text{and} \quad M_0(P(z)) = -\frac{\phi(\Phi^{-1}(1 - P(z)))}{1 - P(z)}$$

are inverse Mills ratio terms.

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

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

References

Aakvik, A., Heckman, J.J., 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., 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 (1–2), 15–51.

Abadie, A. (2002). “Bootstrap tests of distributional treatment effects in instrumental variable models”. Journal of the American Statistical Association 97 (457), 284–292 (March).

Abadie, A., Imbens, G.W. (2006). “Large sample properties of matching estimators for average treatment effects”. Econometrica 74 (1), 235–267 (January).

Ahn, H., Powell, J. (1993). "Semiparametric estimation of censored selection models with a nonparametric selection mechanism". Journal of Econometrics 58 (1–2), 3–29 (July).

Aigner, D.J. (1979a). "A brief introduction to the methodology of optimal experimental design". Journal of Econometrics 11 (1), 7–26.

Aigner, D.J. (1979b). "Sample design for electricity pricing experiments: Anticipated precision for a time-of-day pricing experiment". Journal of Econometrics 11 (1), 195–205 (September).

Aigner, D.J. (1985). "The residential electricity time-of-use pricing experiments: What have we learned?". In: Hausman, J.A., Wise, D.A. (Eds.), Social Experimentation. University of Chicago Press, Chicago, pp. 11–41.

Aigner, D.J., Hsiao, C., Kapteyn, A., Wansbeek, T. (1984). "Latent variable models in econometrics". In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 2. Elsevier, pp. 1321–1393 (Chapter 23).

Altonji, J.G., Matzkin, R.L. (2005). "Cross section and panel data estimators for nonseparable models with endogenous regressors". Econometrica 73 (4), 1053–1102 (July).

Angrist, J.D., Imbens, G.W. (1995). "Two-stage least squares estimation of average causal effects in models with variable treatment intensity". Journal of the American Statistical Association 90 (430), 431–442 (June).

Angrist, J.D., Krueger, A.B. (1999). "Empirical strategies in labor economics". In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, New York, pp. 1277–1366.

Angrist, J.D., Graddy, K., 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 (3), 499–527 (July).

Angrist, J.D., Imbens, G.W., Rubin, D. (1996). "Identification of causal effects using instrumental variables". Journal of the American Statistical Association 91 (434), 444–455.

Athey, S., Imbens, G.W. (2006). "Identification and inference in nonlinear difference-in-differences models". Econometrica 74 (2), 431–497 (March).

Balke, A., Pearl, J. (1997). "Bounds on treatment effects from studies with imperfect compliance". Journal of the American Statistical Association 92 (439), 1171–1176 (September).

Banerjee, A.V. (2006). "Making aid work: How to fight global poverty – Effectively". Boston Review 31 (4) (July/August).

Barnow, B.S., Cain, G.G., Goldberger, A.S. (1980). "Issues in the analysis of selectivity bias". In: Stromsdorfer, E., Farkas, G. (Eds.), Evaluation Studies, vol. 5. Sage Publications, Beverly Hills, CA, pp. 42–59.

Barros, R.P. (1987). "Two essays on the nonparametric estimation of economic models with selectivity using choice-based samples". PhD thesis. University of Chicago.

Basu, A., Heckman, J.J., Navarro-Lozano, S., Urzua, S. (2007). "Use of instrumental variables in the presence of heterogeneity and self-selection: An application to treatments in breast cancer patients". Health Economics 16 (11), 1133–1157 (October).

Behrman, J.R., Sengupta, P., Todd, P. (2005). "Progressing through PROGRESA: An impact assessment of a school subsidy experiment in rural Mexico". Economic Development and Cultural Change 54 (1), 237–275 (October).

Bertrand, M., Duflo, E., Mullainathan, S. (2004). "How much should we trust differences-in-differences estimates?". Quarterly Journal of Economics 119 (1), 249–275 (February).

Bickel, P.J. (1967). "Some contributions to the theory of order statistics". In: LeCam, L., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 575–591.

Björklund, A., Moffitt, R. (1987). "The estimation of wage gains and welfare gains in self-selection". Review of Economics and Statistics 69 (1), 42–49 (February).

Bloom, H.S. (1984). "Accounting for no-shows in experimental evaluation designs". Evaluation Review 82 (2), 225–246.

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

Blundell, R., Powell, J. (2004). “Endogeneity in semiparametric binary response models”. Review of Economic Studies 71 (3), 655–679 (July).

Blundell, R., Duncan, A., Meghir, C. (1998). “Estimating labor supply responses using tax reforms”. Econometrica 66 (4), 827–861 (July).

Bresnahan, T.F. (1987). “Competition and collusion in the American automobile industry: The 1955 price war”. Journal of Industrial Economics 35 (4), 457–482 (June).

Cain, G.G., Watts, H.W. (1973). Income Maintenance and Labor Supply: Econometric Studies. Academic Press, New York.

Cameron, S.V., Heckman, J.J. (1993). “The nonequivalence of high school equivalents”. Journal of Labor Economics 11 (1, Part 1), 1–47 (January).

Cameron, S.V., 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 (2), 262–333 (April).

Campbell, D.T. (1969). “Reforms as experiments”. American Psychologist 24 (4), 409–429. Reprinted in: Struening, E.L., Guttentag, M. (Eds.), Handbook of Evaluation Research, vols. 1, 2. Sage Publication, Beverly Hills, CA, 1975, pp. 71–99 (vol. 1).

Campbell, D.T., Stanley, J.C. (1963). Experimental and Quasi-Experimental Designs for Research. Rand McNally, Chicago (originally appeared in Gage, N.L. (Ed.), Handbook of Research on Teaching).

Card, D. (1999). “The causal effect of education on earnings”. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 5. North-Holland, New York, pp. 1801–1863.

Card, D. (2001). “Estimating the return to schooling: Progress on some persistent econometric problems”. Econometrica 69 (5), 1127–1160 (September).

Carneiro, P. (2002). “Heterogeneity in the returns to schooling: Implications for policy evaluation”. PhD thesis. University of Chicago.

Carneiro, P., Hansen, K., Heckman, J.J. (2001). “Removing the veil of ignorance in assessing the distributional impacts of social policies”. Swedish Economic Policy Review 8 (2), 273–301 (Fall).

Carneiro, P., Hansen, K., 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 (2), 361–422 (May). 2001 Lawrence R. Klein Lecture.

Carneiro, P., Heckman, J.J., Vytlacil, E.J. (2006). “Estimating marginal and average returns to education”. American Economic Review. Submitted for publication.

Cave, G., Bos, H., Doolittle, F., Toussaint, C. (1993). “JOBSTART: Final report on a program for school dropouts”. Technical report, MDRC.

Chan, T.Y., Hamilton, B.H. (2006). “Learning, private information and the economic evaluation of randomized experiments”. Journal of Political Economy 114 (6), 997–1040 (December).

Chen, S. (1999). “Distribution-free estimation of the random coefficient dummy endogenous variable model”. Journal of Econometrics 91 (1), 171–199 (July).

Chen, X., Fan, Y. (1999). “Consistent hypothesis testing in semiparametric and nonparametric models for econometric time series”. Journal of Econometrics 91 (2), 373–401 (August).

Chernozhukov, V., Hansen, C. (2005). “An IV model of quantile treatment effects”. Econometrica 73 (1), 245–261 (January).

Chernozhukov, V., Imbens, G.W., Newey, W.K. (2007). “Nonparametric identification and estimation of nonseparable models”. Journal of Econometrics 139 (1), 1–3 (July).

Cochran, W.G., Rubin, D.B. (1973). “Controlling bias in observational studies: A review”. Sankyha Ser. A 35 (Part 4), 417–446.

Conlisk, J. (1973). “Choice of response functional form in designing subsidy experiments”. Econometrica 41 (4), 643–656 (July).

Conlisk, J., Watts, H. (1969). “A model for optimizing experimental designs for estimating response surfaces”. American Statistical Association Proceedings Social Statistics Section, 150–156.

Cook, T.D., Campbell, D.T. (1979). Quasi-Experimentation: Design and Analysis Issues for Field Settings. Rand McNally College Publishing Company, Chicago.

Cunha, F., 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 (6), 870–893 (December).

Cunha, F., Heckman, J.J. (2008). “A new framework for the analysis of inequality”. Macroeconomic Dynamics. Submitted for publication.

Cunha, F., Heckman, J.J., Matzkin, R. (2003). “Nonseparable factor analysis”. Unpublished manuscript. Department of Economics, University of Chicago.

Cunha, F., Heckman, J.J., Navarro, S. (2005). “Separating uncertainty from heterogeneity in life cycle earnings”. Oxford Economic Papers 57 (2), 191–261 (April). The 2004 Hicks lecture.

Cunha, F., Heckman, J.J., Navarro, S. (2006). “Counterfactual analysis of inequality and social mobility”. In: Morgan, S.L., Grusky, D.B., Fields, G.S. (Eds.), Mobility and Inequality: Frontiers of Research in Sociology and Economics. Stanford Univ. Press, Stanford, CA, pp. 290–348 (Chapter 4).

Cunha, F., Heckman, J.J., Navarro, S. (2007). “The identification and economic content of ordered choice models with stochastic cutoffs”. International Economic Review. In press, November.

Cunha, F., Heckman, J.J., Schennach, S.M. (2007). “Estimating the technology of cognitive and noncognitive skill formation”. Unpublished manuscript, University of Chicago, Department of Economics. Presented at the Yale Conference on Macro and Labor Economics, May 5–7, 2006. Econometrica. Submitted for publication.

Cunha, F., Heckman, J.J., Schennach, S.M. (2006b). “Nonlinear factor analysis”. Unpublished manuscript. Department of Economics, University of Chicago.

Dahl, G.B. (2002). “Mobility and the return to education: Testing a Roy model with multiple markets”. Econometrica 70 (6), 2367–2420 (November).

Darolles, S., Florens, J.-P., Renault, E. (2002). “Nonparametric instrumental regression”. Working Paper 05-2002. Centre interuniversitaire de recherche en économie quantitative, CIREQ.

Deaton, A. (2006). “Evidence-based aid must not become the latest in a long string of development fads”. Boston Review 31 (4) (July/August).

Domencich, T., McFadden, D.L. (1975). Urban Travel Demand: A Behavioral Analysis. North-Holland, Amsterdam. Reprinted 1996.

Doolittle, F.C., Traeger, L. (1990). Implementing the National JTPA Study. Manpower Demonstration Research Corporation, New York.

Duncan, G.M., Leigh, D.E. (1985). “The endogeneity of union status: An empirical test”. Journal of Labor Economics 3 (3), 385–402 (July).

Durbin, J. (1954). “Errors in variables”. Review of the International Statistical Institute 22, 23–32.

Ellison, G., Ellison, S.F. (1999). “A simple framework for nonparametric specification testing”. Journal of Econometrics 96, 1–23 (May).

Farber, H.S. (1983). “Worker preferences for union representation”. In: Reid, J. (Ed.), Research in Labor Economics, Volume Supplement 2: New Approaches to Labor Unions. JAI Press, Greenwich, CT.

Fisher, R.A. (1966). The Design of Experiments. Hafner Publishing, New York.

Florens, J.-P., Heckman, J.J., Meghir, C., Vytlacil, E.J. (2002). “Instrumental variables, local instrumental variables and control functions”. Technical Report CWP15/02, CEMMAP. Econometrica. Submitted for publication.

Florens, J.-P., Heckman, J.J., Meghir, C., Vytlacil, E.J. (2006). “Control functions for nonparametric models without large support”. Unpublished manuscript. University of Chicago.

Friedlander, D., Hamilton, G. (1993). “The Saturation Work Initiative Model in San Diego: A Five-Year Follow-Up Study”. Manpower Demonstration Research Corporation, New York.

Gerfin, M., Lechner, M. (2002). “A microeconomic evaluation of the active labor market policy in Switzerland”. Economic Journal 112 (482), 854–893 (October).

Gill, R.D., Robins, J.M. (2001). “Causal inference for complex longitudinal data: The continuous case”. The Annals of Statistics 29 (6), 1785–1811 (December).

Glynn, R.J., Laird, N.M., Rubin, D.B. (1986). “Selection modeling versus mixture modeling with nonignorable nonresponse”. In: Wainer, H. (Ed.), Drawing Inferences from Self-Selected Samples. Springer-Verlag, New York, pp. 115–142. Reprinted in: Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Gronau, R. (1974). “Wage comparisons – A selectivity bias”. Journal of Political Economy 82 (6), 1119–1143 (November–December).

Haavelmo, T. (1943). “The statistical implications of a system of simultaneous equations”. Econometrica 11 (1), 1–12 (January).

Hahn, J. (1998). “On the role of the propensity score in efficient semiparametric estimation of average treatment effects”. Econometrica 66 (2), 315–331 (March).

Hahn, J., Todd, P.E., Van der Klaauw, W. (2001). “Identification and estimation of treatment effects with a regression-discontinuity design”. Econometrica 69 (1), 201–209 (January).

Hall, P., Horowitz, J. (2005). “Nonparametric methods for inference in the presence of instrumental variables”. Annals of Statistics 33 (6), 2904–2929 (September).

Hansen, K.T., Heckman, J.J., Mullen, K.J. (2004). “The effect of schooling and ability on achievement test scores”. Journal of Econometrics 121 (1–2), 39–98 (July–August).

Härdle, W. (1990). Applied Nonparametric Regression. Cambridge Univ. Press, New York.

Harmon, C., Walker, I. (1999). “The marginal and average returns to schooling in the UK”. European Economic Review 43 (4–6), 879–887 (April).

Hausman, J.A. (1978). “Specification tests in econometrics”. Econometrica 46 (6), 1251–1272 (November).

Heckman, J.J. (1974a). “Effects of child-care programs on women’s work effort”. Journal of Political Economy 82 (2), S136–S163. Reprinted in: Schultz, T.W. (Ed.), Economics of the Family: Marriage, Children and Human Capital. University of Chicago Press, 1974 (March/April).

Heckman, J.J. (1974b). “Shadow prices, market wages, and labor supply”. Econometrica 42 (4), 679–694 (July).

Heckman, J.J. (1976a). “The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models”. Annals of Economic and Social Measurement 5 (4), 475–492 (December).

Heckman, J.J. (1976b). “A life-cycle model of earnings, learning, and consumption”. Journal of Political Economy 84 (4, Part 2), S11–S44 (August). Journal Special Issue: Essays in Labor Economics in Honor of H. Gregg Lewis.

Heckman, J.J. (1976c). “Simultaneous equation models with both continuous and discrete endogenous variables with and without structural shift in the equations”. In: Goldfeld, S., Quandt, R. (Eds.), Studies in Nonlinear Estimation. Ballinger Publishing Company, Cambridge, MA, pp. 235–272.

Heckman, J.J. (1980). “Addendum to sample selection bias as a specification error”. In: Stromsdorfer, E., Farkas, G. (Eds.), Evaluation Studies Review Annual, vol. 5. Sage Publications, Beverly Hills, CA.

Heckman, J.J. (1990). “Varieties of selection bias”. American Economic Review 80 (2), 313–318 (May).

Heckman, J.J. (1992). “Randomization and social policy evaluation”. In: Manski, C., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard Univ. Press, Cambridge, MA, pp. 201–230.

Heckman, J.J. (1996). “Randomization as an instrumental variable”. Review of Economics and Statistics 78 (2), 336–340 (May).

Heckman, J.J. (1997). “Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations”. Journal of Human Resources 32 (3), 441–462 (Summer). Addendum published in: Juornal of Human Resources 33 (1) (1998) (Winter).

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. German–American Academic Council Foundation, Bonn, Germany.

Heckman, J.J. (2001). “Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture”. Journal of Political Economy 109 (4), 673–748 (August).

Heckman, J.J., Honoré, B.E. (1990). “The empirical content of the Roy model”. Econometrica 58 (5), 1121–1149 (September).

Heckman, J.J., 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 (408), 862–874 (December). Rejoinder also published in: Journal of the American Statistical Association 84 (408) (1989) (December).

Heckman, J.J., LaFontaine, P. (2007). America’s Dropout Problem: The GED and the Importance of Social and Emotional Skills. University of Chicago Press, Chicago. Submitted for publication.

Heckman, J.J., Navarro, S. (2004). "Using matching, instrumental variables, and control functions to estimate economic choice models". Review of Economics and Statistics 86 (1), 30–57 (February).

Heckman, J.J., Navarro, S. (2007). "Dynamic discrete choice and dynamic treatment effects". Journal of Econometrics 136 (2), 341–396 (February).

Heckman, J.J., Robb, R. (1985a). "Alternative methods for evaluating the impact of interventions". In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data, vol. 10. Cambridge Univ. Press, New York, pp. 156–245.

Heckman, J.J., Robb, R. (1985b). "Alternative methods for evaluating the impact of interventions: An overview". Journal of Econometrics 30 (1–2), 239–267 (October–November).

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

Heckman, J.J., Robb, R. (1986b). "Postscript: A rejoinder to Tukey". In: Wainer, H. (Ed.), Drawing Inferences from Self-Selected Samples. Springer-Verlag, New York, pp. 111–114. Reprinted in: Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Heckman, J.J., Sedlacek, G.L. (1990). "Self-selection and the distribution of hourly wages". Journal of Labor Economics 8 (1, Part 2), S329–S363. Essays in Honor of Albert Rees.

Heckman, J.J., Smith, J.A. (1993). "Assessing the case for randomized evaluation of social programs". In: Jensen, K., Madsen, P. (Eds.), Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives, Proceedings from the Danish Presidency Conference "Effects and Measuring of Effects of Labour Market Policy Initiatives". Denmark Ministry of Labour, Copenhagen, pp. 35–95.

Heckman, J.J., Smith, J.A. (1998). "Evaluating the welfare state". In: Strom, S. (Ed.), Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium. Cambridge Univ. Press, New York, pp. 241–318.

Heckman, J.J., Smith, J.A. (1999). "The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies". Economic Journal 109 (457), 313–348 (July). Winner of the Royal Economic Society Prize, 1999.

Heckman, J.J., Vytlacil, E.J. (1998). "Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling". Journal of Human Resources 33 (4), 974–987 (Fall).

Heckman, J.J., 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 (April).

Heckman, J.J., Vytlacil, E.J. (2000). "The relationship between treatment parameters within a latent variable framework". Economics Letters 66 (1), 33–39 (January).

Heckman, J.J., Vytlacil, E.J. (2001a). "Instrumental variables, selection models, and tight bounds on the average treatment effect". In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Center for European Economic Research, New York, pp. 1–15.

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

Heckman, J.J., Vytlacil, E.J. (2001c). "Policy-relevant treatment effects". American Economic Review 91 (2), 107–111 (May).

Heckman, J.J., Vytlacil, E.J. (2005). "Structural equations, treatment effects and econometric policy evaluation". Econometrica 73 (3), 669–738 (May).

Heckman, J.J., Vytlacil, E.J. (2007). "Evaluating marginal policy changes and the average effect of treatment for individuals at the margin". Columbia University, Department of Economics. Unpublished manuscript.

Heckman, J.J., Ichimura, H., Todd, P.E. (1997). "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". Review of Economic Studies 64 (4), 605–654 (October).

Heckman, J.J., Ichimura, H., Todd, P.E. (1998). “Matching as an econometric evaluation estimator”. Review of Economic Studies 65 (223), 261–294 (April).

Heckman, J.J., LaLonde, R.J., Smith, J.A. (1999). “The economics and econometrics of active labor market programs”. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, New York, pp. 1865–2097 (Chapter 31).

Heckman, J.J., Lochner, L.J., Taber, C. (1998). “General-equilibrium treatment effects: A study of tuition policy”. American Economic Review 88 (2), 381–386 (May).

Heckman, J.J., Lochner, L.J., Todd, P.E. (2006). “Earnings equations and rates of return: The Mincer equation and beyond”. In: Hanushek, E.A., Welch, F. (Eds.), Handbook of the Economics of Education. North-Holland, Amsterdam, pp. 307–458.

Heckman, J.J., Smith, J.A., Clements, N. (1997). “Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts”. Review of Economic Studies 64 (221), 487–536 (October).

Heckman, J.J., Smith, J.A., Taber, C. (1998). “Accounting for dropouts in evaluations of social programs”. Review of Economics and Statistics 80 (1), 1–14 (February).

Heckman, J.J., Tobias, J.L., Vytlacil, E.J. (2003). “Simple estimators for treatment parameters in a latent variable framework”. Review of Economics and Statistics 85 (3), 748–754 (August).

Heckman, J.J., Urzua, S., Vytlacil, E.J. (2004). “Understanding instrumental variables in models with essential heterogeneity: Unpublished results”. Unpublished manuscript. Department of Economics, University of Chicago.

Heckman, J.J., Urzua, S., Vytlacil, E.J. (2006). “Understanding instrumental variables in models with essential heterogeneity”. Review of Economics and Statistics 88 (3), 389–432.

Heckman, J.J., Ichimura, H., Smith, J., Todd, P.E. (1998). “Characterizing selection bias using experimental data”. Econometrica 66 (5), 1017–1098 (September).

Heckman, J.J., Hohmann, N., Smith, J., Khoo, M. (2000). “Substitution and dropout bias in social experiments: A study of an influential social experiment”. Quarterly Journal of Economics 115 (2), 651–694 (May).

Hirano, K., Imbens, G.W., Ridder, G. (2003). “Efficient estimation of average treatment effects using the estimated propensity score”. Econometrica 71 (4), 1161–1189 (July).

Hollister, R.G., Kemper, P., Maynard, R.A. (1984). The National Supported Work Demonstration. University of Wisconsin Press, Madison, WI.

Hotz, V.J. (1992). “Designing an evaluation of the Job Training Partnership Act”. In: Manski, C., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard Univ. Press, Cambridge, MA, pp. 76–114.

Hotz, V.J., Mullin, C.H., Sanders, S.G. (1997). “Bounding causal effects using data from a contaminated natural experiment: Analysing the effects of teenage childbearing”. Review of Economic Studies 64 (4), 575–603 (October).

Hu, Y., Schennach, S.M. (2006). “Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions”. Working Paper. University of Chicago.

Hurwicz, L. (1962). “On the structural form of interdependent systems”. In: Nagel, E., Suppes, P., Tarski, A. (Eds.), Logic, Methodology and Philosophy of Science. Stanford Univ. Press, pp. 232–239.

Ichimura, H., Taber, C. (2002). “Semiparametric reduced-form estimation of tuition subsidies”. American Economic Review 92 (2), 286–292 (May).

Ichimura, H., Thompson, T.S. (1998). “Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution”. Journal of Econometrics 86 (2), 269–295 (October).

Ichimura, H., Todd, P.E. (2007). “Implementing nonparametric and semiparametric estimators”. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, Amsterdam.

Imbens, G.W. (2003). “Sensitivity to exogeneity assumptions in program evaluation”. American Economic Review 93 (2), 126–132 (May).

Imbens, G.W. (2004). “Nonparametric estimation of average treatment effects under exogeneity: A review”. Review of Economics and Statistics 86 (1), 4–29 (February).

Imbens, G.W., Angrist, J.D. (1994). “Identification and estimation of local average treatment effects”. Econometrica 62 (2), 467–475 (March).

Imbens, G.W., Newey, W.K. (2002). “Identification and estimation of triangular simultaneous equations models without additivity”. Technical Working Paper 285. National Bureau of Economic Research.

Kramer, M.S., Shapiro, S.H. (1984). “Scientific challenges in the application of randomized trials”. JAMA: The Journal of the American Medical Association 252 (19), 2739–2745 (November).

Kemple, J.J., Friedlander, D., Fellerath, V. (1995). “Florida’s Project Independence: Benefits, Costs, and Two-Year Impacts of Florida’s JOBS Program”. Manpower Demonstration Research Corporation, New York.

LaLonde, R.J. (1984). “Evaluating the econometric evaluations of training programs with experimental data”. Technical Report 183. Industrial Relations Section, Department of Economics, Princeton University.

Lechner, M. (2001). “Identification and estimation of causal effects of multiple treatments under the conditional independence assumption”. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluations of Active Labor Market Policies in Europe. Physica/Springer, Heidelberg.

Lee, L.-F. (1978). “Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables”. International Economic Review 19 (2), 415–433 (June).

Lee, L.-F. (1983). “Generalized econometric models with selectivity”. Econometrica 51 (2), 507–512 (March).

Mallar, C., Kerachsky, S., Thorton, C. (1980). “The short-term economic impact of the Job Corps program”. In: Stromsdorfer, E., Farkas, G. (Eds.), Evaluation Studies Review Annual, vol. 5. Sage Publications.

Manski, C.F. (1989). “Anatomy of the selection problem”. Journal of Human Resources 24 (3), 343–360 (Summer).

Manski, C.F. (1990). “Nonparametric bounds on treatment effects”. American Economic Review 80 (2), 319–323 (May).

Manski, C.F. (1994). “The selection problem”. In: Sims, C. (Ed.), Advances in Econometrics: Sixth World Congress. Cambridge Univ. Press, New York, pp. 143–170.

Manski, C.F. (1995). Identification Problems in the Social Sciences. Harvard Univ. Press, Cambridge, MA.

Manski, C.F. (1996). “Learning about treatment effects from experiments with random assignment of treatments”. Journal of Human Resources 31 (4), 709–733 (Autumn).

Manski, C.F. (1997). “Monotone treatment response”. Econometrica 65 (6), 1311–1334 (November).

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

Manski, C.F., Pepper, J.V. (2000). “Monotone instrumental variables: With an application to the returns to schooling”. Econometrica 68 (4), 997–1010 (July).

Mare, R.D. (1980). “Social background and school continuation decisions”. Journal of the American Statistical Association 75 (370), 295–305 (June).

Marschak, J. (1953). “Economic measurements for policy and prediction”. In: Hood, W., Koopmans, T. (Eds.), Studies in Econometric Method. Wiley, New York, pp. 1–26.

Masters, S.H., Maynard, R.A. (1981). The Impact of Supported Work on Long-Term Recipients of AFDC Benefits. Manpower Demonstration Research Corporation, New York.

Matzkin, R.L. (1993). “Nonparametric identification and estimation of polychotomous choice models”. Journal of Econometrics 58 (1–2), 137–168 (July).

Matzkin, R.L. (1994). “Restrictions of economic theory in nonparametric methods”. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North-Holland, New York, pp. 2523–2558.

Matzkin, R.L. (2003). “Nonparametric estimation of nonadditive random functions”. Econometrica 71 (5), 1339–1375 (September).

Matzkin, R.L. (2007). “Nonparametric identification”. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, Amsterdam.

Maynard, R., Brown, R.S. (1980). The Impact of Supported Work on Young School Dropouts. Manpower Demonstration Research Corporation, New York.

McFadden, D. (1974). “Conditional logit analysis of qualitative choice behavior”. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York.

Moffitt, R. (1992). “Evaluation methods for program entry effects”. In: Manski, C., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard Univ. Press, Cambridge, MA, pp. 231–252.

Newey, W.K., Powell, J.L. (2003). “Instrumental variable estimation of nonparametric models”. Econometrica 71 (5), 1565–1578 (September).

Olley, G.S., Pakes, A. (1996). “The dynamics of productivity in the telecommunications equipment industry”. Econometrica 64 (6), 1263–1297 (November).

Palca, J. (1989). “AIDS drug trials enter new age”. Science, New Series 246 (4926), 19–21 (October 6).

Pearl, J. (2000). Causality. Cambridge Univ. Press, Cambridge, England.

Pessino, C. (1991). “Sequential migration theory and evidence from Peru”. Journal of Development Economics 36 (1), 55–87 (July).

Peterson, A.V. (1976). “Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks”. Proceedings of the National Academy of Sciences 73 (1), 11–13 (January).

Powell, J.L. (1994). “Estimation of semiparametric models”. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. Elsevier, Amsterdam, pp. 2443–2521.

Prescott, E.C., Visscher, M. (1977). “Sequential location among firms with foresight”. Bell Journal of Economics 8 (2), 378–893 (Autumn).

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 (284), 873–880 (December).

Quandt, R.E. (1972). “A new approach to estimating switching regressions”. Journal of the American Statistical Association 67 (338), 306–310 (June).

Quint, J.C., Polit, D.F., Bos, H., Cave, G. (1994). New Chance Interim Findings on a Comprehensive Program for Disadvantaged Young Mothers and Their Children. Manpower Demonstration Research Corporation, New York.

Rao, C.R. (1985). “Weighted distributions”. In: Atkinson, A., Fienberg, S. (Eds.), A Celebration of Statistics: The ISI Centenary Volume. Springer-Verlag, New York.

Robins, J.M. (1989). “The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies”. In: Sechrest, L., Freeman, H., Mulley, A. (Eds.), Health Services Research Methodology: A Focus on AIDS. United States Department of Health and Human Services, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD, pp. 113–159.

Robins, J.M. (1997). “Causal inference from complex longitudinal data”. In: Berkane, M. (Ed.), Latent Variable Modeling and Applications to Causality. In: Lecture Notes in Statistics. Springer-Verlag, New York, pp. 69–117.

Robinson, C. (1989). “The joint determination of union status and union wage effects: Some tests of alternative models”. Journal of Political Economy 97 (3), 639–667.

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

Rosenbaum, P.R., Rubin, D.B. (1983). “The central role of the propensity score in observational studies for causal effects”. Biometrika 70 (1), 41–55 (April).

Roy, A. (1951). “Some thoughts on the distribution of earnings”. Oxford Economic Papers 3 (2), 135–146 (June).

Rubin, D.B. (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies”. Journal of Educational Psychology 66 (5), 688–701 (October).

Rubin, D.B. (1978). “Bayesian inference for causal effects: The role of randomization”. Annals of Statistics 6 (1), 34–58 (January).

Rubin, D.B. (1979). “Using multivariate matched sampling and regression adjustment to control bias in observational studies”. Journal of the American Statistical Association 74 (366), 318–328 (June).

Rudin, W. (1974). Real and Complex Analysis, second ed. McGraw–Hill, New York.

Schennach, S.M. (2004). “Estimation of nonlinear models with measurement error”. Econometrica 72 (1), 33–75 (January).

Shaked, A., Sutton, J. (1982). “Relaxing price competition through product differentiation”. Review of Economic Studies 49 (1), 3–13 (January).

Silvey, S.D. (1970). Statistical Inference. Penguin, Harmondsworth.

Smith, J.A. (1992). “The JTPA selection process: A descriptive analysis”. Unpublished working paper. Department of Economics, University of Chicago.

Smith, V.K., Banzhaf, H.S. (2004). “A diagrammatic exposition of weak complementarity and the Willig condition”. American Journal of Agricultural Economics 86 (2), 455–466 (May).

Smith, J.P., Welch, F.R. (1986). Closing the Gap: Forty Years of Economic Progress for Blacks. RAND Corporation, Santa Monica, CA.

Smith, J., Whalley, A., Wilcox, N. (2006). "Are program participants good evaluators?" Unpublished manuscript. Department of Economics, University of Michigan.

Telser, L.G. (1964). "Iterative estimation of a set of linear regression equations". Journal of the American Statistical Association 59 (307), 845–862 (September).

Todd, P.E. (1999). "A practical guide to implementing matching estimators". Unpublished manuscript. Department of Economics, University of Pennsylvania. Prepared for the IADB meeting in Santiago, Chile. (October).

Todd, P.E. (2007). "Evaluating social programs with endogenous program placement and selection of the treated". In: Handbook of Development Economics. Elsevier, Amsterdam. In press.

Todd, P.E. (2008). "Matching estimators". In: Durlauf, S., Blume, L.E. (Eds.), The New Palgrave Dictionary of Economics. Palgrave Macmillan, New York. In press.

Torp, H., Raaum, O., Hernæs, E., Goldstein, H. (1993). "The first Norwegian experiment". In: Burtless, G. (Ed.), Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policy Initiatives, Proceedings from the Danish Presidency Conference "Effects and Measuring of Effects of Labour Market Policy Initiatives". Kolding, May 1993. Denmark Ministry of Labour, Copenhagen.

Tunali, I. (2000). "Rationality of migration". International Economic Review 41 (4), 893–920 (November).

Vijverberg, W.P.M. (1993). "Measuring the unidentified parameter of the extended Roy model of selectivity". Journal of Econometrics 57 (1–3), 69–89 (May–June).

Vytlacil, E.J. (2002). "Independence, monotonicity, and latent index models: An equivalence result". Econometrica 70 (1), 331–341 (January).

Vytlacil, E.J. (2006a). "A note on additive separability and latent index models of binary choice: Representation results". Oxford Bulletin of Economics and Statistics 68 (4), 515–518 (August).

Vytlacil, E.J. (2006b). "Ordered discrete choice selection models: Equivalence, nonequivalence, and representation results". Review of Economics and Statistics 88 (3), 578–581 (August).

Vytlacil, E.J., Yildiz, N. (2006). "Dummy endogenous variables in weakly separable models". Unpublished manuscript. Department of Economics, Columbia University.

Vytlacil, E.J., Santos, A., Shaikh, A.M. (2005). "Limited dependent variable models and bounds on treatment effects: A nonparametric analysis". Unpublished manuscript. Department of Economics, Columbia University.

White, H. (1984). Asymptotic Theory for Econometricians. Academic Press, Orlando, FL.

Willis, R.J., Rosen, S. (1979). "Education and self-selection". Journal of Political Economy 87 (5, Part 2), S7–S36 (October).

Wooldridge, J.M. (1997). "On two stage least squares estimation of the average treatment effect in a random coefficient model". Economics Letters 56 (2), 129–133 (October).

Wooldridge, J.M. (2003). "Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model". Economics Letters 79 (2), 185–191 (May).

Wu, D. (1973). "Alternative tests of independence between stochastic regressors and disturbances". Econometrica 41 (4), 733–750 (July).

Yitzhaki, S. (1989). "On using linear regression in welfare economics". Working Paper 217. Department of Economics, Hebrew University.

Yitzhaki, S. (1996). "On using linear regressions in welfare economics". Journal of Business and Economic Statistics 14 (4), 478–486 (October).

Yitzhaki, S., Schechtman, E. (2004). "The Gini Instrumental Variable, or the "double instrumental variable" estimator". Metron 62 (3), 287–313.

Zellner, A., Rossi, P.E. (1987). "Evaluating the methodology of social experiments". In: Munnell, A.H. (Ed.), Lessons from the Income Maintenance Experiments: Proceedings of a Conference Held at Melvin Village. New Hampshire, September, 1986. In: Federal Reserve Bank of Boston Conference Series, vol. 30. Brookings Institution, Washington, DC, pp. 131–157.

Zheng, J.X. (1996). "A consistent test of functional form via nonparametric estimation techniques". Journal of Econometrics 75, 263–289 (December).