TECHNICAL WORKING PAPER SERIES

LOCAL INSTRUMENTAL VARIABLES

James J. Heckman
Edward J. Vytlacil

Technical Working Paper 252
http://www.nber.org/papers/T0252

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2000

We would like to thank Arild Aakvik, Jaap Abbring, Victor Aguirregabiria, Lars Hansen, Justin Tobias, two anonymous referees, and especially Cheng Hsiao for close and helpful readings of this manuscript. We thank participants in the Canadian Econometric Studies Group (September, 1998), The Midwest Econometrics Group (September, 1998), the University of Upsalla (November, 1998), the University of Chicago (December, 1998), University College London (December, 1998), the University of California-Berkeley (April, 1999), and the North American Summer Meetings of the Econometric Society (June, 1999). This research was supported by NIH:R01-HD3498-01, NIH:R01-HD3208-03, NSF97-09-873, and the Donner Foundation. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Local Instrumental Variables
James J. Heckman and Edward J. Vytlacil
NBER Technical Working Paper No. 252
April 2000
JEL No. C50, H43

ABSTRACT

This paper unites the treatment effect literature and the latent variable literature. The economic questions answered by the commonly used treatment effect parameters are considered. We demonstrate how the marginal treatment effect parameter can be used in a latent variable framework to generate the average treatment effect, the effect of treatment on the treated and the local average treatment effect, thereby establishing a new relationship among these parameters. The method of local instrumental variables directly estimates the marginal treatment effect parameters, and thus can be used to estimate all of the conventional treatment effect parameters when the index condition holds and the parameters are identified. When they are not, the method of local instrumental variables can be used to produce bounds on the parameters with the width of the bounds depending on the width of the support for the index generating the choice of the observed potential outcome.

James J. Heckman
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
j-heckman@uchicago.edu

Edward J. Vytlacil
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
e-vytlacil@uchicago.edu

1 Introduction

Latent variable models arise in many well-posed economic problems. The latent variables can be utilities, potential wages (or home wages as in Gronau, 1974 and Heckman, 1974) or potential profitability. This class of models, which originates in psychology in the work of Thurstone (1930, 1959), has been widely developed in the econometrics of discrete choice (see McFadden, 1981, and the survey of index models in labor economics presented by Heckman and MaCurdy, 1985).

This paper uses the latent variable or index model of econometrics and psychometrics to impose structure on the Neyman (1923) - Fisher (1935) - Cox (1958) - Rubin (1978) model of potential outcomes used to define treatment effects. That model is isomorphic to the Roy model (1951) as summarized by Heckman and Honoré (1990) and to Quandt's switching regression model (1972, 1988). For a comprehensive discussion of these models, see Heckman and Vytlacil (2000b).

A recent development in econometrics has been an emphasis on the estimation of certain features of economic models under weaker assumptions about functional forms of estimating equations and error distributions than are conventionally maintained in estimating structural econometric models. The recent "treatment effect" literature is the most agnostic in this regard, focusing on the estimation of certain "treatment effects" that can be nonparametrically identified under general conditions. Two major limitations of this literature are (a) that the economic questions answered by the estimated "treatment effects" are usually not clearly stated and (b) that the connection of this literature to the traditional parametric index function literature is not well established. The parameters estimated in the classical parametric discrete choice literature can be used to answer a variety of policy questions. In contrast, the parameters in the modern "treatment effect" literature answer only narrowly focused questions, but typically under weaker conditions than are postulated in the parametric literature.

This paper unites the treatment effect literature and the latent variable literature. The economic questions answered by the commonly used treatment effect parameters are considered. We demonstrate how the marginal treatment effect (MTE) parameter introduced in Heckman (1997) can be used in a latent variable framework to generate the average treatment effect (ATE), the effect of treatment on the treated (TT) and the local average treatment effect (LATE) of Imbens and Angrist (1994), thereby establishing a new relationship among these parameters. The method of local instrumental variables (LIV) introduced in Heckman and Vytlacil (1999b) directly estimates the MTE parameter, and thus can be used to estimate all of the conventional treatment effect parameters when the index condition holds and the parameters are identified. When they are not, LIV can be used to produce bounds on the parameters with the width of the bounds depending on the width of the support for the index generating the choice of the observed potential outcome.

As a consequence of the analysis of Vytlacil (1999a), the latent variable framework used in this paper is more general than might first be thought. He establishes that the assumptions used by Imbens and Angrist (1994) to identify LATE using linear instrumental variables both imply and are implied by the latent variable set up used in this paper. Thus our analysis applies to the entire class of models estimated by LATE.

LATE analysis focuses on what linear instrumental variables can estimate. LIV extends linear IV analysis and estimates (or bounds) a much wider class of treatment parameters. Under conditions presented in this paper, suitably weighted versions of LIV identify the Average Treatment Effect (ATE) and Treatment on the Treated (TT) even in the general case where responses to treatment are heterogeneous and agents participate in the program being evaluated at least in part on the basis of this heterogeneous response. Heckman (1997) shows that in this case the linear instrumental variable estimator does not identify ATE or TT. We establish conditions under which LIV identifies those parameters.

The plan of this paper is as follows. Section 2 presents a model of potential outcomes in a latent variable framework. Section 3 defines four different mean treatment parameters within the latent variable framework. Section 4 establishes a new relationship among the parameters using MTE as a unifying device. Section 5 presents conditions for identification of treatment effect parameters, and presents bounds for them when they are not identified. The LIV estimator is introduced as the empirical analog to MTE that operationalizes our identification analysis. Section 6 compares the LIV estimator to the linear IV estimator. Weighted versions of LIV identify the treatment on the treated parameter (TT) and the average treatment effect (ATE) in cases where the linear instrumental variable estimator does not. The leading case where this phenomenon arises is the additively separable correlated random coefficient model which is developed in this section. Section 7 extends the analysis of the additively separable case and applies it to classical results in the selection bias literature. Section 8 concludes the paper. The Appendix explores the sensitivity of the analysis presented in the text to the assumptions imposed on the latent index model.

2 Models of Potential Outcomes in a Latent Variable Framework

For each person i, assume two potential outcomes (Y_{0i}, Y_{1i}) corresponding, respectively, to the potential outcomes in the untreated and treated states. Our methods generalize to the case of multiple outcomes, but in this paper we consider only two outcomes.[1] Let $D_i = 1$ denote the receipt of treatment; $D_i = 0$ denotes nonreceipt. Let Y_i be the measured outcome variable so that

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

This is the Neyman-Fisher-Cox-Rubin model of potential outcomes. It is also the switching regression model of Quandt (1972) and the Roy model of income distribution (Roy, 1951; Heckman and Honoré, 1990).

[1] See Heckman and Vytlacil (2000b) for the multi-outcome extension.

This paper assumes that a latent variable model generates the indicator variable D_i. Specifically, we assume that the assignment or decision rule for the indicator is generated by a latent variable D_i^*:

$$
\begin{aligned}
D_i^* &= \mu_D(Z_i) - U_{Di} \\
D_i &= 1 \text{ if } D_i^* \geq 0, \quad = 0 \text{ otherwise,}
\end{aligned} \tag{1}
$$

where Z_i is a vector of observed random variables and U_{Di} is an unobserved random variable. D_i^* is the net utility or gain to the decision-maker from choosing state 1. The index structure underlies many models in econometrics (see, *e.g.*, the survey in Amemiya 1985) and in psychometrics (see, e.g., Junker and Ellis, 1997).

The potential outcome for the program participation state is

$$Y_{1i} = \mu_1(X_i, U_{1i}),$$

and the potential outcome for the nonparticipation state is

$$Y_{0i} = \mu_0(X_i, U_{0i}),$$

where X_i is a vector of observed random variables and (U_{1i}, U_{0i}) are unobserved random variables. No assumptions are imposed restricting the joint distribution of (U_{0i}, U_{1i}) beyond the regularity conditions presented below.[2] It is assumed that Y_{0i} and Y_{1i} are defined for everyone and that these outcomes are independent across persons so that there are no interactions among agents.[3] Important special cases include models with (Y_{0i}, Y_{1i}) generated by latent variables. These models include $\mu_j(X_i, U_{ji}) = \mu_j(X_i) + U_{ji}$ if Y is continuous and $\mu_j(X_i, U_{ji}) = 1(X_i\beta_j + U_{ji} \geq 0)$ if Y is binary, where X_i is independent of U_{ji} and where $1(A)$ is the indicator function that takes the value 1 if the event A is true and takes the value 0 otherwise. We do not restrict the (μ_1, μ_0) functions except for the regularity conditions noted below. Let Δ_i denote the treatment effect for individual i:

$$\Delta_i = Y_{1i} - Y_{0i}. \tag{2}$$

The treatment effect is a person-specific counterfactual. For person i it answers the question, what would be the outcome if the person received the treatment compared to the case where the person had not received the treatment? For notational convenience, we will henceforth suppress the i subscript.

In this paper we assume that:

[2] A common restriction is that $U_{0i} = U_{1i}$. If U_{1i}, U_{0i} are additively separable from X_i, this restriction generates a common treatment effect model (Heckman and Robb, 1986; Heckman, 1997). The restriction $U_{0i} = U_{1i}$ also aids in identification of treatment effects in nonseparable models if at least one instrument is continuous (Vytlacil, 1999b). We do not impose these restrictions in this paper.

[3] The problem of interactions among agents in the analysis of treatment effects was extensively discussed by Lewis (1963) although he never used the term "treatment effect." See the papers by Davidson and Woodbury (1993) and Heckman, Lochner and Taber (1998) for empirical demonstrations of the importance of these social interaction effects, and the general discussion of general equilibrium treatment effects in Heckman, LaLonde and Smith (1999).

(i) $\mu_D(Z)$ is a nondegenerate random variable conditional on X

(ii) (U_D, U_1), and (U_D, U_0) are absolutely continuous with respect to Lebesgue measure on $\Re^2$

(iii) (U_D, U_1) and (U_D, U_0) are independent of (Z, X)

(iv) Y_1 and Y_0 have finite first moments

(v) $1 > \Pr(D = 1|X = x) > 0$ for every $x \in \text{Supp}(X)$

Assumption (i) requires an exclusion restriction: there exists a variable that determines the treatment decision but does not directly affect the outcome. Variables that satisfy these conditions are commonly called instrumental variables. Assumption (ii) is imposed for convenience, both to simplify the notation and to impose smoothness on certain conditional expectations; it can readily be relaxed. Assumption (iii) can be weakened to the assumption that (U_D, U_0) and (U_D, U_1) are independent of Z conditional on X. We work with the stronger condition (iii) to simplify the notation. The modifications required for the more general case are trivial. If assumption (v) is relaxed so that $\Pr(D = 1|X = x) = 1$ or 0 for some x values, then the analysis of this paper will still hold for any x value for which $1 > \Pr(D = 1|X = x) > 0$.

For any random variable A, let F_A denote the variable's distribution function and let a denote a possible realization of A. Let $P(z)$ denote the probability of receiving treatment conditional on the observed covariates,

$$P(z) \equiv \Pr(D=1|Z=z) = F_{U_D}(\mu_D(z)). \tag{3}$$

$P(z)$ is sometimes called the "propensity score" by statisticians following Rosenbaum and Rubin (1983) and is called a "choice probability" by economists[4]

Without loss of generality, we assume that $U_D \sim \text{Unif}[0,1]$, in which case $\mu_D(z) = P(z)$. To see that there is no loss of generality, note that if the underlying index is $D^* = \nu(Z) - V$, with assumptions (ii) and (iii) satisfied for V, taking $\mu(Z) = F_V(\nu(Z))$ and $U = F_V(V)$ equates the two models. This transformation is innocuous, since any CDF is left-continuous and non-decreasing and thus $\mu(z) \geq U_D \Leftrightarrow \nu(z) \geq V$.[5] In addition, since U_D is distributed Unif$[0,1]$ and independent of Z, we have $\mu(z) = P(z)$.

Note that the latent variable assumption does not impose testable restrictions on choice behavior.[6] If we take Z to include all observed covariates, and define $\mu(z) = \Pr(D = 1|Z = z)$ and $U_D \sim \text{Unif}[0,1]$, the latent index assumption imposes no restrictions on the observed choice behavior.[7] However, it does impose two restrictions on counterfactual

---

[4]Heckman and Robb (1986) present a discussion of how the propensity score is used differently in selection models and in matching models for program evaluation. See also Heckman, LaLonde and Smith (1999) and Heckman and Vytlacil (2000b).

[5]More precisely, $\nu(z) \geq V \Rightarrow F_V(\nu(z)) \geq F_V(V)$ and $F_V(\nu(z)) \geq F_V(V) \Rightarrow \nu(z) \geq V$ for any $V \in \text{Supp}(V)$, so that the equivalence holds w.p.1.

[6]However, it does impose a testable restriction on the conditional distribution of the outcome variable. In particular, it imposes the index sufficiency restriction that, for any set A, $\Pr(Y_j \in A|Z=z, D=j) = \Pr(Y_j \in A|P(Z)=P(z), D=j)$. See Heckman, Ichimura, Smith and Todd (1998) for a nonparametric test of this restriction.

[7]This argument is also used in Das and Newey (1998).

outcomes. First, consider the following hypothetical intervention. If we take a random sample of individuals, and externally set Z at level z, with what probability would they have $D = 1$ after the intervention? Using the assumption that U_D is independent of Z, the answer is $\Pr(D = 1|Z = z)$, that is, the probability that $D = 1$ among those individuals who were observed to have $Z = z$. Second, if we instead took individuals with $Z = z$ and externally set their Z characteristics to z', where $\Pr(D = 1|Z = z) < \Pr(D = 1|Z = z')$, then the threshold crossing model implies that some individuals who would have had $D = 0$ with the $Z = z$ characteristic will now have $D = 1$ with the $Z = z'$ characteristics, but that no individual who would have had $D = 1$ with the $Z = z$ characteristics will have $D = 0$ with the $Z = z'$ characteristics. These two properties are the *only* restrictions imposed by the threshold crossing model as we have defined it. Under these conditions on counterfactual outcomes, there is no loss of generality in imposing a threshold crossing model (Vytlacil, 1999a).

Imbens and Angrist (1994) invoke these two properties of independence and monotonicity in their LATE analysis. Thus our latent variable model is equivalent to the LATE model. If we remove the assumption that U_D is independent of Z, while leaving the assumptions otherwise unchanged, then the monotonicity property continues to hold but not the independence property. If instead we remove the assumption that U_D is additively separable from Z (i.e., we consider $\mu_D(Z, U_D)$ instead of $\mu_D(Z) - U_D$), while leaving the model otherwise unchanged, then the independence property will hold but not in general, the monotonicity property.[8] If we remove both assumptions, while imposing no additional restrictions, then the threshold crossing model becomes completely vacuous, imposing no restrictions on the observed outcomes or counterfactual outcomes.

The Appendix investigates the sensitivity of the analysis presented in the text to the index assumption by alternatively dropping the independence or monotonicity assumptions (allowing U_D and Z to be dependent or allowing U_D and Z to be additively nonseparable in the latent index in a general way). We show that the definition of the parameters and the relationships among them as described below in Sections 3 and 4 generalize with only minor modifications to the case where U_D and Z are additively nonseparable or are stochastically dependent. The separability and independence assumptions allow us to define the parameters in terms of $P(z)$ instead of z and allow for slightly simpler expressions, but are not crucial for the definition of parameters or the relationship among them. However, the assumptions of independence and monotonicity *are* essential for establishing the connection between the LATE or LIV estimators and the underlying parameters as described in Section 5, and *are* essential to the identification analysis that accompanies that discussion.

[8] Note monotonicity is implied by additively separability, but additive separability is not required; supermodular $\mu_D(Z, U_D)$ is all that is required, *i.e.* in the differentiable case a uniform positive (or negative) cross partial of μ_D with respect to Z and U_D is all that is required. However, Vytlacil (1999a) shows that any latent index that satisfies the monotonicity condition will have an additively separable representation, so that having an additively separable representation is required.

3 Definition of Parameters

A recent shift in econometric research is toward estimating features of a model rather than estimating the full model, as is emphasized in structural econometrics. In the context of the program evaluation problem, this comes down to estimating parameters like ATE, TT, LATE and MTE directly, rather than estimating all the ingredients of the underlying structural model separately that can be built up to estimate these parameters. In general, these special parameters can be identified under weaker conditions than are required to estimate the full structural parameters, but at the same time, they cannot generate the complete array of policy counterfactuals produced from estimates of the full model. The weaker identifying assumptions make estimators based on them more widely accepted. At the same time, the estimates produce answers to more narrowly focused questions.

In this paper, we consider four different mean treatment parameters within this framework: the average treatment effect (ATE), the effect of treatment on the treated (TT), the marginal treatment effect (MTE), and the local average treatment effect (LATE). ATE and TT are the traditional parameters. Each of these parameters is a mean of the individual treatment effect, $\Delta = Y_1 - Y_0$, but with different conditioning sets. The average treatment effect is defined as

$$\Delta^{ATE}(x) \equiv E(\Delta | X = x).^{9}$$

The mean effect of treatment on the treated is the most commonly estimated parameter for both observational data and social experiments (see Heckman and Robb, 1985, 1986, and Heckman, LaLonde and Smith, 1999). It is defined as

$$\Delta^{TT}(x, D=1) \equiv E(\Delta|X=x, D=1).^{10}$$

It will be useful for the analysis of this paper to define a version of $\Delta^{TT}(x, D=1)$ that conditions on the propensity score, $P(z)$, defined in equation (3):

$$\Delta^{TT}(x, P(z), D=1) \quad \equiv \quad E(\Delta|X=x, P(Z)=P(z), D=1)^{11}$$

so that

$$\Delta^{TT}(x, D=1) \quad = \quad \int_0^1 \Delta^{TT}(x, p, D=1) dF_{P(Z)|X,D}(p|x,1). \tag{4}$$

The third parameter that we analyze is the marginal treatment parameter introduced in Heckman (1997) and defined in the context of a latent variable model as:

$$\Delta^{MTE}(x, u) \equiv E(\Delta|X=x, U_D=u).^{12}$$

[9] From assumption (iv), it follows that $E(\Delta|X=x)$ exists and is finite a.e. F_X.

[10] From assumption (iv), $\Delta^{TT}(x, D=1)$ exists and is finite a.e. $F_{X|D=1}$, where $F_{X|D=1}$ denotes the distribution of X conditional on $D=1$.

[11] From our assumptions, $\Delta^{TT}(x, P(z), D=1)$ exists and is finite a.e. $F_{X,P(Z)|D=1}$.

The final parameter we analyze is the LATE parameter of Imbens and Angrist (1994) defined by an instrumental variable. Using $P(Z)$ as the instrument:

$$\Delta^{LATE}(x, P(z), P(z')) \equiv \frac{E(Y|X=x, P(Z)=P(z)) - E(Y|X=x, P(Z)=P(z'))}{P(z)-P(z')}.$$

We require that $P(z) \neq P(z')$ for any (z, z') where the parameter is defined. We will assume that $P(z) > P(z')$, with no loss of generality given the restriction that $P(z) \neq P(z')$. This definition of the treatment parameter, while consistent with that used by Imbens and Angrist, is somewhat peculiar because it is based on an estimator rather than a property of a model. An alternative, more traditional, definition is given below.

A more general framework defines the parameters in terms of Z. As a consequence of the latent variable or index structure, defining the parameters in terms of Z or $P(Z)$ results in equivalent expressions.[13] In the index model, Z enters the model only through the index, so that for any measurable set A,

$$\begin{aligned}
\Pr(Y_j \in A|X=x, Z=z, D=1) &= \Pr(Y_j \in A|X=x, U_D \leq P(z)) \\
&= \Pr(Y_j \in A|X=x, P(Z)=P(z), D=1) \\
\Pr(Y_j \in A|X=x, Z=z, D=0) &= \Pr(Y_j \in A|X=x, U_D > P(z))
\end{aligned}$$

[12] From our assumptions, $\Delta^{MTE}(x,u)$ exists and is finite a.e. F_{X,U_D}.

[13] As discussed in the Appendix, we can equivalently define the parameters in terms of Z or $P(Z)$ because of both our additive separability and independence assumptions.

$$= \Pr(Y_j \in A | X = x, P(Z) = P(z), D = 0).$$

4 Relationship Among Parameters Using the Index Structure

Given the index structure, a simple relationship exists among the four parameters. From the definition it is immediate that

$$\Delta^{TT}(x, P(z), D = 1) = E(\Delta | X = x, U_D \leq P(z)). \tag{5}$$

Next consider $\Delta^{LATE}(x, P(z), P(z'))$. Note that

$$\begin{aligned} &E(Y|X = x, P(Z) = P(z)) \\ &= P(z)\Big[E(Y_1|X = x, P(Z) = P(z), D = 1)\Big] \\ &\quad +(1 - P(z))\Big[E(Y_0|X = x, P(Z) = P(z), D = 0)\Big] \\ &= \int_0^{P(z)} E(Y_1|X = x, U_D = u)du + \int_{P(z)}^1 E(Y_0|X = x, U_D = u)du, \end{aligned} \tag{6}$$

so that

$$\begin{aligned} &E(Y|X = x, P(Z) = P(z)) - E(Y|X = x, P(Z) = P(z')) \\ &= \int_{P(z')}^{P(z)} E(Y_1|X = x, U_D = u)du - \int_{P(z')}^{P(z)} E(Y_0|X = x, U_D = u)du, \end{aligned}$$

and thus

$$\Delta^{LATE}(x, P(z), P(z')) \quad = \quad E(\Delta | X = x, P(z') \leq U_D \leq P(z)).$$

Notice that this expression could be taken as an alternative definition of LATE.

We can rewrite these relationships in succinct form in the following way:

$$\begin{aligned} \Delta^{MTE}(x,u) &= E(\Delta|X = x, U_D = u) \\ \Delta^{ATE}(x) &= \int_0^1 E(\Delta|X = x, U_D = u)du \\ P(z)[\Delta^{TT}(x, P(z), D = 1)] &= \int_0^{P(z)} E(\Delta|X = x, U_D = u)du \\ (P(z) - P(z'))[\Delta^{LATE}(x, P(z), P(z'))] &= \int_{P(z')}^{P(z)} E(\Delta|X = x, U_D = u)du. \end{aligned} \tag{7}$$

Each parameter is an average value of MTE, $E(\Delta \mid X = x, U_D = u)$, 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.[14]

$\Delta^{MTE}(x,u)$ 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

[14] As suggested by a referee, the relationship between MTE and LATE or TT conditional on $P(z)$ is closely 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 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 for people on a specified margin of participation in the program.

$P(Z) = u$. $\Delta^{MTE}(x, u)$ for values of u close to zero is the average effect for individuals with unobservable characteristics that make them the most inclined to participate in the program ($D = 1$), and $\Delta^{MTE}(x, u)$ for values of u close to one is the average treatment effect for individuals with unobserved (by the econometrician) characteristics that make them the least inclined to participate. ATE integrates $\Delta^{MTE}(x, u)$ over the entire support of U_D (from $u = 0$ to $u = 1$). It is the average effect for an individual chosen at random from the entire population. $\Delta^{TT}(x, P(z), D = 1)$ is the average treatment effect for persons who chose to participate at the given value of $P(Z) = P(z)$. $\Delta^{TT}(x, P(z), D = 1)$ integrates $\Delta^{MTE}(x, u)$ up to $u = P(z)$. 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 $P(Z) \leq P(z')$ and would participate if $P(Z) \geq P(z)$. $\Delta^{LATE}(x, P(z), P(z'))$ integrates $\Delta^{MTE}(x, u)$ from $u = P(z')$ to $u = P(z)$.

Using the third expression in equation (7) to substitute into equation (4), we obtain an alternative expression for the TT parameter as a weighted average of MTE parameters:

$$\Delta^{TT}(x, D = 1) = \int_0^1 \frac{1}{p} \left[\int_0^p E(\Delta | X = x, U_D = u) du \right] dF_{P(Z)|X,D}(p|x, 1).$$

Using Bayes' rule, it follows that

$$dF_{P(Z)|X,D}(p|x,1) \quad = \quad \frac{\Pr(D=1|X=x,P(Z)=p)}{\Pr(D=1|X=x)} dF_{P(Z)|X}(p|x). \tag{8}$$

Since $\Pr(D=1|X=x,P(Z)=p)=p$, it follows that

$$\Delta^{TT}(x,D=1)=\frac{1}{\Pr(D=1|X=x)}\int_0^1\left[\int_0^p E(\Delta|X=x,U_D=u)du\right]dF_{P(Z)|X}(p|x). \tag{9}$$

Note further that since $\Pr(D=1|X=x)=E(P(Z)|X=x)=\int_0^1(1-F_{P(Z)|X}(t|x))dt$, we can reinterpret (9) 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:

$$\begin{aligned}
\Delta^{TT}(x,D=1) &= \frac{1}{\Pr(D=1|X=x)}\int_0^1\left[\int_0^1 1(u\le p)E(\Delta|X=x,U_D=u)du\right]dF_{P(Z)|X}(p|x)\\
&= \frac{1}{\int(1-F_{P(Z)|X}(t|x))dt}\int_0^1\left[\int_0^1 E(\Delta|X=x,U_D=u)1(u\le p)dF_{P(Z)|X}(p|x)\right]du\\
&= \int_0^1 E(\Delta|X=x,U_D=u)\left[\frac{1-F_{P(Z)|X}(u|x)}{\int(1-F_{P(Z)|X}(t|x))dt}\right]du\\
&= \int_0^1 E(\Delta|X=x,U_D=u)g_x(u)du
\end{aligned}$$

where $g_x(u)=\frac{1-F_{P(Z)|X}(u|x)}{\int(1-F_{P(Z)|X}(t|x))dt}$. Thus $g_x(u)$ is a *weighted distribution* (Rao, 1985). Since $g_x(u)$ is a nonincreasing function of u, we have that drawings from $g_x(u)$ oversample persons with low values of U_D, i.e., values of unobserved characteristics that make them the most likely to participate in the program no matter what their value of $P(Z)$. Since

$$\Delta^{MTE}(x,u) = E(\Delta|X = x, U_D = u)$$

it follows that

$$\Delta^{TT}(x, D = 1) \quad = \quad \int_0^1 \Delta^{MTE}(x,u) g_x(u) du.$$

The TT parameter is thus a weighted version of MTE, where $\Delta^{MTE}(x,u)$ is given the largest weight for low u values and is given zero weight for $u \geq p_x^{max}$, where p_x^{max} is the maximum value in the support of $P(Z)$ conditional on $X = x$.

Figure 1 graphs the relationship between $\Delta^{MTE}(u)$, Δ^{ATE} and $\Delta^{TT}(P(z), D = 1)$, 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, 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 $\Delta^{MTE}(u)$ over the full unit interval (*i.e.* is the area under A minus the area under B and C in the figure). $\Delta^{TT}(P(z), D = 1)$ averages $\Delta^{MTE}(u)$ up to the point $P(z)$ (is the area under A minus the area under B in the figure). Because $\Delta^{MTE}(u)$ is assumed to be declining in u, the TT parameter for any given $P(z)$ evaluation point is larger then the ATE parameter.

Equation (7) 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 paper when we encounter conditions where LATE can be identified but MTE cannot. MTE is the limit form of LATE:

$$\Delta^{MTE}(x,p) = \lim_{p' \to p} \Delta^{LATE}(x,p,p').$$

Direct relationships between LATE and the other parameters are easily derived. The relationship between LATE and ATE is immediate:

$$\Delta^{ATE}(x) = \Delta^{LATE}(x,0,1).$$

Using Bayes' rule, the relationship between LATE and TT is

$$\Delta^{TT}(x, D=1) = \int_0^1 \Delta^{LATE}(x,0,p) \frac{p}{\Pr(D=1|X=x)} dF_{P(Z)|X}(p|x). \tag{10}$$

5 Identification and Bounds for the Treatment Effect Parameters

Assume access to an infinite i.i.d. sample of (D,Y,X,Z) observations, so that the joint distribution of (D,Y,X,Z) is known. Let $\mathcal{P}_x$ denote the closure of the support of $P(Z)$ conditional on $X=x$, and let $\mathcal{P}_x^c = (0,1) \setminus \mathcal{P}_x$. Let p_x^{max} and p_x^{min} be the maximum and minimum values in $\mathcal{P}_x$. We show that the identification of the treatment parameters and the width of the bounds on the unidentified parameters depend critically on $\mathcal{P}_x$.[15]

We define the Local IV (LIV) estimand to be

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

LIV is the limit form of the LATE expression as $P(z) \to P(z')$.[16] In equation (6), $E(Y_1|X = x, U_D)$ and $E(Y_0|X = x, U_D)$ are integrable with respect to dF_U a.e. F_X. Thus, $E(Y_1|X = x, P(Z) = P(z))$ and $E(Y_0|X = x, P(Z) = P(z))$ are differentiable a.e. with respect to $P(z)$, and thus $E(Y|X = z, P(Z) = P(z))$ is differentiable a.e. with respect to $P(z)$ with derivative given by:

$$\frac{\partial E(Y|X=x, P(Z)=P(z))}{\partial P(z)} = E(Y_1 - Y_0|X = x, U_D = P(z)),^{17} \qquad (11)$$

[15]Heckman, Ichimura, Smith and Todd (1996), Heckman, Ichimura, Smith and Todd (1998), and Heckman, Ichimura, and Todd (1998) emphasize that the identification of treatment parameters critically depends on the support of the propensity score and present empirical evidence that failure of a full support condition is a major source of evaluation bias.

[16]The limit form of LATE was introduced in this context by Heckman (published 1997; first draft, 1995) and Heckman and Smith (published 1998; first draft, 1995). Those authors introduced the limit form of LATE within the context of a selection model as a way to connect the LATE parameter to economic theory and to policy analysis. Angrist, Graddy and Imbens (NBER Working Paper, 1995) also develop a limit form of LATE within the context of a model of supply and demand, but use it only as a device for interpreting the linear IV estimand and place no direct economic interpretation on the limit LATE parameter. Bjorklund and Moffitt (1987) consider a parametric version of this parameter for the Roy model. These papers do not develop the relationships among the parameters or the identification analysis that are the primary concerns of this paper and of Heckman and Vytlacil (1999b, 2000a).

and thus

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

Note that while $\Delta^{LIV}(x, P(z))$ is defined for individuals with a given value of $P(Z)$, $P(Z) = P(z)$, it gives the marginal effect for individuals with a given value of U_D, $U_D = P(z)$. In other words, it is estimated conditional on an observed $P(Z) = P(z)$ but defines a treatment effect for those with a given unobserved proclivity to participate that is equal to the evaluation point, $U_D = P(z)$.

LATE and LIV are defined as functions (Y, X, Z), and are thus straightforward to identify. $\Delta^{LATE}(x, P(z), P(z'))$ is identified for any $(P(z), P(z')) \in \mathcal{P}_x \times \mathcal{P}_x$ such that $P(z) \neq P(z')$. $\Delta^{LIV}(x, P(z))$ is identified for any $P(z)$ that is a limit point of $\mathcal{P}_x$. The larger the support of $P(Z)$ conditional on $X = x$, the bigger the set of LIV and LATE parameters that can be identified.

ATE and TT are not defined directly as functions of (Y, X, Z), so a more involved discussion of their identification is required. We can use LIV or LATE to identify ATE and TT under the appropriate support conditions:

(i) If $\mathcal{P}_x = [0,1]$, then $\Delta^{ATE}(x)$ is identified from $\{\Delta^{LIV}(x,p) : p \in [0,1]\}$. If $\{0,1\} \in \mathcal{P}_x$,

[17] See, *e.g.*, Kolmogorov and Fomin (1970), Theorem 9.8 for one proof. From assumption (iv), the derivative in (11) is finite a.e. F_{X,U_D}. The same argument could be used to show that $\Delta^{LATE}(x, P(z), P(z'))$ is continuous and differentiable in $P(z)$ and $P(z')$.

then $\Delta^{ATE}(x)$ is identified from $\Delta^{LATE}(x, 0, 1)$.

(ii) If $[0, P(z)] \subset \mathcal{P}_x$, then $\Delta^{TT}(x, P(z), D = 1)$ is identified from $\{\Delta^{LIV}(x, p) : p \in [0, P(z)]\}$. If $\{0, P(z)\} \in \mathcal{P}_x$, then $\Delta^{TT}(x, P(z), D = 1)$ is identified from $\Delta^{LATE}(x, 0, P(z))$.

(iii) If $\mathcal{P}_x = [0, p_x^{max}]$, then $\Delta^{TT}(x, D = 1)$ is identified from $\{\Delta^{LIV}(x, p) : p \in [0, p_x^{max}]\}$. If $\{0\} \in \mathcal{P}_x$, then $\Delta^{TT}(x, D = 1)$ is identified from $\{\Delta^{LATE}(x, 0, p) : p \in \mathcal{P}_x\}$. (see equation 10).

Note that TT (not conditional on $P(z)$) is identified under weaker conditions than is ATE. To identify TT, one needs to observe $P(z)$ arbitrarily close to 0 ($p_x^{min} = 0$) and to observe some positive $P(z)$ values, while to identify ATE one needs to observe $P(z)$ arbitrarily close to 0 and also $P(z)$ arbitrarily close to 1 ($p_x^{max} = 1$ and $p_x^{min} = 0$).

5.1 Bounds for the Parameters

When the preceding support conditions do not hold, it is still possible to construct bounds for the treatment parameters if Y_1 and Y_0 are known to be bounded w.p.1. To simplify the notation, assume that Y_1 and Y_0 have the same bounds, so that:

$$\Pr(y_x^l \leq Y_1 \leq y_x^u | X = x) = 1$$

and

$$\Pr(y_x^l \leq Y_0 \leq y_x^u | X = x) = 1.$$[18]

For example, if Y is an indicator variable, then the bounds are $y_x^l = 0$ and $y_x^u = 1$ for all x. One set of bounds follows directly from the fact that MTE can be integrated up to the other parameters. For example, for ATE

$$\Delta^{ATE}(x) = \int_{\mathcal{P}_x} \Delta^{LIV}(x,p)dp + \int_{\mathcal{P}_x^c} \Delta^{LIV}(x,p)dp.$$

Since $\Delta^{LIV}(x, P(z))$ is bounded by $(y_x^l - y_x^u)$ and $(y_x^u - y_x^l)$, we obtain

$$\begin{aligned} \Delta^{ATE}(x) &\leq \textstyle\int_{\mathcal{P}_x} \Delta^{LIV}(x,p)dp + (y_x^u - y_x^l) \int_{\mathcal{P}_x^c} dp \\ \Delta^{ATE}(x) &\geq \textstyle\int_{\mathcal{P}_x} \Delta^{LIV}(x,p)dp + (y_x^l - y_x^u) \int_{\mathcal{P}_x^c} dp. \end{aligned} \tag{12}$$

A similar analysis applies to the expressions for TT.

Figures 2 and 3 present a graphical analysis of these bounds for ATE, drawn for the same assumptions about MTE as is used in Figure 1. As before the outcome variable is assumed binary so that $y_x^l = 0$ and $y_x^u = 1$. Figures 2 and 3 are drawn assuming that the support of the propensity score is an interval, with the vertical dotted lines in the figure denoting the end points of the interval. Figure 2 is drawn to represent the lower bound on ATE, with the lower bound being the area under the curve after the curve has been set to $y_x^l - y_x^u = 0 - 1 = -1$ for u outside the support. Figure 3 is drawn to represent the upper

[18] The modifications required to analyze the more general case are straightforward.

bound on ATE, with the upper bound being the area under the curve after the curve has been set to $y_x^u - y_x^l = 1 - 0 = 1$ for u outside the support.

These bounds can be tightened substantially by virtue of the following argument. We do not identify $E(Y_1|X = x, P(Z) = p)$ or $E(Y_0|X = x, P(Z) = p)$ pointwise for $p < p_x^{min}$ or $p > p_x^{max}$, and thus cannot use LIV to identify $\Delta^{MTE}(x, u)$ for $u < p_x^{min}$ or $u > p_x^{max}$. However, it turns out that the latent index structure allows us to identify an averaged version of $E(Y_1|X = x, U = u)$ for $u < p_x^{min}$ and averaged version of $E(Y_0|X = x, U = u)$ for $u > p_x^{max}$, and we can use this additional information to construct tighter bounds. Our argument is of interest in its own right, because it shows how to use information on $E(Y_0|X = x, P(Z) = p, D = 0)$, something we can observe within the proper support, to attain at least partial information on $E(Y_0|X = x, P(Z) = p, D = 1)$, an unobserved, counterfactual quantity.

To apply this type of reasoning, note that $DY = DY_1$ is an observed random variable, and thus for any $x \in \text{Supp}(X)$, $P(z) \in \mathcal{P}_x$, we identify the expectation of DY_1 given $X = x, P(Z) = P(z)$:

$$\begin{aligned} E\Big(DY_1\Big|X = x, P(Z) = P(z)\Big) &= E(Y_1|X = x, D = 1, P(Z) = P(z))P(z) \\ &= E(Y_1|X = x, P(z) \geq U_D)P(z) \\ &= \int_0^{P(z)} E(Y_0|X = x, U_D = u)du. \end{aligned} \tag{13}$$

By similar reasoning,

$$E\Big((1-D)Y_0|X=x, P(Z)=P(z)\Big) = \int_{P(z)}^{1} E(Y_0|X=x, U_D=u)du. \qquad (14)$$

We can evaluate (13) at $P(z)=p_x^{max}$ and evaluate (14) at $P(z)=p_x^{min}$. The distribution of (D,Y,X,Z) contains no information on $\int_{p_x^{max}}^{1} E(Y_1|X=x,U_D=u)du$ and $\int_0^{p_x^{min}} E(Y_0|X=x,U_D=u)du$, but we can bound these quantities:

$$\begin{array}{rcccl} (1-p_x^{max})y_x^l & \leq & \int_{p_x^{max}}^{1} E(Y_1|X=x,U_D=u)du & \leq & (1-p_x^{max})y_x^u \\ p_x^{min}y_x^l & \leq & \int_0^{p_x^{min}} E(Y_0|X=x,U_D=u)du & \leq & p_x^{min}y_x^u. \end{array} \qquad (15)$$

We can thus bound $\Delta^{ATE}(x)$ by:

$$\begin{aligned} \Delta^{ATE}(x) \;\leq\; & p_x^{max}\Big[E(Y_1|X=x,P(Z)=p_x^{max},D=1)\Big] + (1-p_x^{max})y_x^u \\ & -(1-p_x^{min})\Big[E(Y_0|X=x,P(Z)=p_x^{min},D=0)\Big] - p_x^{min}y_x^l \\ \Delta^{ATE}(x) \;\geq\; & p_x^{max}\Big[E(Y_1|X=x,P(Z)=p_x^{max},D=1)\Big] + (1-p_x^{max})y_x^l \\ & -(1-p_x^{min})\Big[E(Y_0|X=x,P(Z)=p_x^{min},D=0)\Big] - p_x^{min}y_x^u. \end{aligned}$$ [19]

The width of the bounds is thus

[19] The bounds on ATE can also be derived by applying Manski's (1990) bounds for "Level-Set Restrictions on the Outcome Regressions." The bounds for the other parameters discussed in this paper cannot be derived by applying his results.

$$\left[(1-p_x^{max})+p_x^{min}\right](y_x^u-y_x^l).$$

If $\mathcal{P}_x$ is an interval, then the width of these bounds is half the width of the bounds given by equation (12). If $\mathcal{P}(x)$ is not an interval, then the width of these bounds is less than half the width of the bounds given by equation (12).

The width of the bounds is linearly related to the distance between p_x^{max} and 1 and the distance between p_x^{min} and 0. These bounds are directly related to the "identification at infinity" results of Heckman (1990) and Heckman and Honoré (1990). Such identification at infinity results require the condition that $\mu_D(Z)$ takes arbitrarily large and arbitrarily small values if the support of U_D is unbounded. This type of identifying condition is sometimes criticized as not being credible. However, as is made clear by the width of the bounds just presented, the proper metric for measuring how close one is to identification at infinity is the distance between p_x^{max} and 1 and the distance between p_x^{min} and 0. It is credible that these distances may be small. In practice, semiparametric econometric methods that use identification at infinity arguments to identify ATE implicitly extrapolate $E(Y_1|X=x,U_D=u)$ for $u>p_x^{max}$ and $E(Y_0|X=x,U_D=u)$ for $u<p_x^{min}$.

We can construct analogous bounds for $\Delta^{TT}(x,P(z),D=1)$ for $P(z)\in\mathcal{P}_x$ in terms of observed objects in an analogous fashion. Recall that

$$\begin{aligned}\Delta^{TT}(x,P(z),D=1) &= E(Y_1-Y_0|X=x,P(Z)=P(z),D=1)\\ &= E(Y_1|X=x,P(Z)=P(z),D=1)-E(Y_0|X=x,P(Z)=P(z),D=1)\end{aligned}$$

We know the first term for $P(z) \in \mathcal{P}(x)$. The second term is the missing counterfactual, which we can rewrite as

$$E(Y_0|X=x, P(Z)=P(z), D=1) = \frac{1}{P(z)}\int_0^{P(z)} E(Y_0|X=x, U=u)du.$$

For $P(z) \in \mathcal{P}_x$, we identify

$$E(Y_0|X=x, P(Z)=P(z), D=0) = \frac{1}{1-P(z)}\int_{P(z)}^{1} E(Y_0|X=x, U=u)du,$$

and we identify

$$E(Y_0|X=x, P(Z)=p_x^{min}, D=0) = \frac{1}{1-p_x^{min}}\int_{p_x^{min}}^{1} E(Y_0|X=x, U=u)du.$$

We therefore identify

$$(1-p_x^{min})E(Y_0|X=x, P(Z)=p_x^{min}, D=0) - (1-P(z))E(Y_0|X=x, P(Z)=p_x^{min}, D=0)$$

$$= \int_{p_x^{min}}^{P(z)} E(Y_0|X=x, U=u)du.$$

We do not identify $\int_0^{p_x^{min}} E(Y_0|X=x, U=u)du$. However, we can bound it by

$$y_x^l p_x^{min} \leq \int_0^{p_x^{min}} E(Y_0|X=x, U=u)du \leq y_x^u p_x^{min}$$

We thus have that

$$\begin{aligned}
E(Y_0|X=x,P(Z)=P(z),D=1) &\leq \frac{1}{P(z)}\Big[p_x^{min}y_x^l+(1-p_x^{min})E(Y_0|X=x,P(Z)=p_x^{min},D=0) \\
&\quad -(1-P(z))E(Y_0|X=x,P(Z)=P(z),D=0)\Big] \\
E(Y_0|X=x,P(Z)=P(z),D=1) &\geq \frac{1}{P(z)}\Big[p_x^{min}y_x^u+(1-p_x^{min})E(Y_0|X=x,P(Z)=p_x^{min},D=0) \\
&\quad -(1-P(z))E(Y_0|X=x,P(Z)=P(z),D=0)\Big].
\end{aligned}$$

Using these inequalities, we can bound $\Delta^{TT}(x,P(z),D=1)$ as follows:

$$\begin{aligned}
\Delta^{TT}(x,P(z),D=1) &\leq E(Y_1|X=x,P(Z)=P(z),D=1) \\
&\quad -\frac{1}{P(z)}\Big[p_x^{min}y_x^l+(1-p_x^{min})E(Y_0|X=x,P(Z)=p_x^{min},D=0) \\
&\quad -(1-P(z))E(Y_0|X=x,P(Z)=P(z),D=0)\Big] \\
\Delta^{TT}(x,P(z),D=1) &\geq E(Y_1|X=x,P(Z)=P(z),D=1) \\
&\quad -\frac{1}{P(z)}\Big[p_x^{min}y_x^u+(1-p_x^{min})E(Y_0|X=x,P(Z)=p_x^{min},D=0) \\
&\quad -(1-P(z))E(Y_0|X=x,P(Z)=P(z),D=0)\Big].
\end{aligned}$$

The width of the bounds for $\Delta^{TT}(x,P(z),D=1)$ is thus

$$\frac{p_x^{min}}{P(z)}(y_x^u-y_x^l).$$

As in the analysis of ATE, the width of the bounds is linearly decreasing in the distance between p_x^{min} and 0. Note that the bounds are tighter for larger $P(z)$ evaluation points, because the higher the $P(z)$ evaluation point, the less weight is placed on the unidentified quantity $\int_0^{p_x^{min}} E(Y_0|X=x,U_D=u)du$.

We can integrate the bounds on $\Delta^{TT}(x,P(z),D=1)$ to bound $\Delta^{TT}(x,D=1)$:

$$\begin{aligned}
\Delta^{TT}(x,D=1) \quad &\leq \quad \int_0^{p_x^{max}} \Big[E(Y_1|X=x,P(Z)=p,D=1) \\
&\qquad -\frac{1}{p}\Big(p_x^{min}y_x^l + (1-p_x^{min})E(Y_0|X=x,P(Z)=p_x^{min},D=0) \\
&\qquad -(1-p)E(Y_0|X=x,P(Z)=p,D=0)\Big)\Big]dF_{P(Z)|X,D}(p|x,1) \\
\Delta^{TT}(x,D=1) \quad &\geq \quad \int_0^{p_x^{max}} \Big[E(Y_1|X=x,P(Z)=p,D=1) \\
&\qquad -\frac{1}{p}\Big(p_x^{min}y_x^u + (1-p_x^{min})E(Y_0|X=x,P(Z)=p_x^{min},D=0) \\
&\qquad -(1-p)E(Y_0|X=x,P(Z)=p,D=0)\Big)\Big]dF_{P(Z)|X,D}(p|x,1).
\end{aligned}$$

The width of the bounds on $\Delta^{TT}(x,D=1)$ is thus

$$p_x^{min}(y_x^u - y_x^l)\int_{p_x^{min}}^{p_x^{max}} \frac{1}{p}dF_{P(Z)|X,D}(p|x,1).$$

Using (8), we obtain

$$p_x^{min}(y_x^u - y_x^l)\int_{p_x^{min}}^{p_x^{max}} \frac{1}{p}dF_{P(Z)|X,D}(p|x,1)$$

$$
\begin{aligned}
&= p_x^{min}(y_x^u - y_x^l)\int_{p_x^{min}}^{p_x^{max}} \frac{1}{\Pr(D=1|X=x)} dF_{P(Z)|X}(p|x) \\
&= p_x^{min}(y_x^u - y_x^l)\frac{1}{\Pr(D=1|X=x)}.^{20}
\end{aligned}
$$

Unlike the bounds on ATE, the bounds on TT depend on the distribution of $P(Z)$, in particular, on $\Pr(D = 1|X = z) = E(P(Z)|X = x)$. The width of the bounds is linearly related to the distance between p_x^{min} and 0, holding $\Pr(D = 1|X = x)$ constant. The larger $\Pr(D = 1|X = x)$, the tighter the bounds, since the larger $P(Z)$ is on average, the less probability weight is being placed on the unidentified quantity $\int_0^{p_x^{min}} E(Y_0|X = x, U_D = u)du$.

6 Linear IV vs. Local IV

Our analysis differs from that of Imbens and Angrist (1994) because we use a local version of instrumental variables and not linear instrumental variables. This section compares these two approaches. As before, we condition on X, and use the propensity score, $P(Z)$, as the instrument.

The linear IV estimand is

$$
\Delta^{IV}(x) = \frac{\mathrm{COV}(Y, P(Z) \mid X = x\ )}{\mathrm{COV}(D, P(Z) \mid X = x)}.
$$

[20] Recall that, by the definition of p_x^{max} and p_x^{min}, we have that the support of $P(Z)$ is a subset of the interval $[p_x^{min}, p_x^{max}]$.

Using the law of iterated expectations, we obtain

$$\frac{\mathrm{COV}(Y, P(Z) \mid X = x)}{\mathrm{COV}(D, P(Z) \mid X = x)} = \frac{\mathrm{COV}(Y, P(Z) \mid X = x)}{\mathrm{VAR}(P(Z) \mid X = x)}.$$

The linear IV estimand can thus be interpreted as the slope term on p from the linear least squares approximation to $E(Y|P(Z) = p, X = x)$, holding x fixed. The local IV estimand can be interpreted in the following way. Take a Taylor series expansion of $E(Y|P(Z) = p, X = x)$ for a fixed x in a neighborhood of $p = p_0$:

$$\begin{aligned} E(Y|P(Z) = p, X = x) &= E(Y|P(Z) = p_0, X = x) + (p - p_0)\frac{\partial E(Y|X = x, P(Z) = p)}{\partial p}\bigg|_{p=p_0} \\ &\quad + o(|p - p_0|) \\ &= E(Y|P(Z) = p_0, X = x) + (p - p_0)\Delta^{LIV}(x, p_0) + o(|p - p_0|). \end{aligned}$$

Local IV evaluated at the point p_0 estimates the slope term from a linear approximation to $E(Y|P(Z) = p, X = x)$ for p in a neighborhood of the point p_0, conditional on $X = x$. Linear IV is the slope term from a global (but conditional on X), linear least squares approximation to $E(Y|P(Z) = p, X = x)$, while local IV is the slope term from a local, linear approximation to $E(Y|P(Z) = p, X = x)$ at a prespecified point. If the treatment effect, $\Delta = Y_1 - Y_0$, does not vary over individuals conditional on X, then it can easily be shown that

$$E(Y|P(Z)=p, X=x) \quad = \quad E(Y_0|X=x) + E(Y_1 - Y_0|X=x)p.^{21}$$

Thus, if the treatment effect does not vary over individuals, $E(Y|P(Z)=p, X=x)$ will be a linear function of p with slope given by $E(Y_1 - Y_0|X=x)$ and the linear IV estimand and the local IV estimand coincide. If the treatment effect varies over individuals, then in the general case

$$\begin{aligned} E(Y|P(Z)=p, X=x) &= E(Y_0|X=x) + E(Y_1 - Y_0|X=x, Z=z, D=1)p \\ &= E(Y_0|X=x) + E(Y_1 - Y_0|X=x, P(Z)=p, D=1)p. \end{aligned}$$

Then, if $Y_1 - Y_0$ is not mean independent of $P(Z)$ conditional on (X, D), we have that $E(Y|P(Z)=p, X=x)$ will not be a linear function of p and the linear IV estimand and the local IV estimand will not coincide. We present an economic interpretation of this condition later on in this section.

By estimating the slope term locally, LIV is able to identify the MTE parameter pointwise:

$$\Delta^{LIV}(x,p) = \Delta^{MTE}(x,p).$$

[21] Thus $Y = Y_0 + D(Y_1 - Y_0)$, $E(Y|X=x, P(Z)=p) = E(Y_0|X=x) + E(Y_1 - Y_0|X=x, P(Z)=p, D=1)p$, and the result follows using that $E(Y_1 - Y_0|X=x, P(Z)=p, D=1) = E(Y_1 - Y_0|X=x)$ from the assumption that the treatment effect does not vary over individuals conditional on X.

In contrast, linear IV identifies a weighted average of the LIV estimands, and thus a weighted average of the MTE parameter. Using the law of iterated expectations, and assuming that the relevant second moments exist and are finite, one can show that

$$\Delta^{IV}(x) = \int_0^1 \Delta^{LIV}(x,u) h_x(u) du = \int_0^1 \Delta^{MTE}(x,u) h_x(u) du, \tag{16}$$

where

$$h_x(u) = \frac{(E(P(Z)|P(Z) \geq u, X = x)) \Pr(P(Z) \geq u, X = x)}{E[(P(Z))^2 | X = x]}, \tag{17}$$

and

$$\int_0^1 h_x(u) du = 1.^{22} \tag{18}$$

As shown by Imbens and Angrist (1994), linear IV converges to some weighted average of treatment effects. However, the implicit weighting of the MTE parameter implied by linear IV does not in general equal the weighting corresponding to the parameters defined in Section 3. One can show that the IV weighting equals the weighting corresponding to LATE only if the support of $P(Z)$ conditional on $X = x$ only contains two values, $\mathcal{P}_x = \{p', p\}$. The linear IV weighting equals the weighting of MTE corresponding to TT in the special case where $\mathcal{P}_x = \{0, p\}$, and equals the weighting corresponding to ATE in the special case where $\mathcal{P}_x = \{0, 1\}$. The IV weighting cannot correspond to the MTE parameter at a given

[22] This result is essentially the same as Theorem 2 of Imbens and Angrist (1994) except that it is explicitly based on the latent index representation.

evaluation point. Linear IV is a global approach (conditional on X) which estimates a weighted average of the MTE parameters. Local IV is a form of instrumental variables, conditional on $X = x_0$ and within a neighborhood of $P(Z) = p_0$, that estimates a given MTE parameter, which when suitably weighted and integrated, produces both TT and ATE.

To summarize the discussion of this section and to link the treatment effect parameters defined and discussed in the previous section to the familiar switching regression model, or correlated random coefficient model, it is useful to consider the following separable version of it. Let

$$Y_1 = \mu_1(X) + U_1,$$

$$Y_0 = \mu_0(X) + U_0,$$

and define D as arising from a latent index crossing a threshold as before. Assume $E(U_1) = 0$ and $E(U_0) = 0$.[23] We maintain assumptions (i) - (v). Let F denote the distribution function of U_D, and let $\tilde{U}_D = F(U_D)$. In this case, we can write the outcome equation as a random coefficient model:

$$\begin{aligned} Y &= DY_1 + (1-D)Y_0 \\ &= \mu_0(X) + [\mu_1(X) - \mu_0(X) + U_1 - U_0]D + U_0, \end{aligned}$$

[23] Because of the additive separability assumption, it is more natural to consider (Y_0, Y_1) to be continuous variables in this example.

where the coefficient of D is a random or variable coefficient. From the definition, it follows that

$$\begin{aligned} E(Y_1|X=x,Z=z,D=1) &= \mu_1(X)+E(U_1|X=x,Z=z,D=1) \\ &= \mu_1(X)+E(U_1|U_D\leq\mu_D(z)) \\ &= \mu_1(X)+K_{11}(P(z)), \end{aligned}$$

and

$$\begin{aligned} E(Y_0|X=x,Z=z,D=0) &= \mu_0(X)+E(U_0|X=x,Z=z,D=0) \\ &= \mu_0(X)+E(U_1|U_D>\mu_D(z)) \\ &= \mu_0(X)+K_{00}(P(z)), \end{aligned}$$

where $K_{ij}(P(z)) = E(U_i|Z=z,D=j)$ is a control function (see Heckman, 1980 and Heckman and Robb, 1985, 1986).

In this setup,

$$\begin{aligned} \Delta^{ATE}(x) &= E(Y_1-Y_0 \mid X=x)=\mu_1(x)-\mu_0(x) \\ \Delta^{TT}(x,z,D=1) &= E(Y_1-Y_0 \mid X=x,Z=z,D=1) \\ &= \mu_1(x)-\mu_0(x)+E\Big(U_1-U_0\Big|X=x,Z=z,D=1\Big) \\ &= \mu_1(x)-\mu_0(x)+E\Big(U_1-U_0\Big|U_D\leq\mu_D(z)\Big). \end{aligned}$$

We develop these expressions further in the next section.

Again consider the relationship between the LIV and linear IV estimators. Using the additive separability assumption, we obtain

$$\Delta^{LIV}(x,p) = \Delta^{MTE}(x,p) = \mu_1(x) - \mu_0(x) + E(U_1 - U_0|\tilde{U}_D = p),$$

and using equation (16),

$$\Delta^{IV}(x) = \mu_1(x) - \mu_0(x) + \int_0^1 E(U_1 - U_0|\tilde{U}_D = u)h_x(u)du, \tag{19}$$

where $h_x(u)$ is defined in equation (17).

To compare what LIV and the linear instrumental variable estimator identify, it is useful to follow Heckman (1997) in comparing three cases. Case 1 (C-1) is the homogeneous response model:

$$U_1 = U_0. \tag{C-1}$$

In this case, $\Delta^{MTE}(x,u) = \mu_1(x) - \mu_0(x)$ for all u. The effect of treatment does not vary over individuals given X. Using equation (19) and the fact that $h_x(u)$ integrates to one, we obtain $\Delta^{IV}(x) = \Delta^{ATE}(x) = \Delta^{TT}(x, D=1)$. Because $\Delta^{MTE}(x,u)$ does not vary with u, this is a case where linear IV will identify the parameters of interest even though the implicit weighting of MTE implied by linear IV will not generally equal the weighting corresponding to the parameters of interest.

In the case of heterogeneous responses, $U_1 \neq U_0$, it is useful to distinguish two further cases: (C-2) where agent enrollment into the program does not depend on $U_1 - U_0$ and case (C-3) where it does. Case (C-2) is (C-2a) plus (C-2b) or (C-2b$'$):

$$U_1 \neq U_0 \tag{C-2a}$$

and

$$E(U_1 - U_0 | X = x, Z = z, D = 1) = E(U_1 - U_0 | X = x, D = 1) \tag{C-2b}$$

for z a.e. $F_{Z|X=x}$ s.t. $P(z) \neq 0$, or

$$E(U_1 - U_0 | X = x, Z = z, D = 1) = M \tag{C-2b$'$}$$

for z a.e. $F_{Z|X=x}$ s.t. $P(z) \neq 0$, for some constant M. Under (C-2a) and (C-2b), $\Delta^{IV}(x) = \Delta^{TT}(x, D = 1)$; under (C-2a) and (C-2b$'$), $\Delta^{IV}(x) = \Delta^{ATE}(x)$ (see Heckman, 1997; Heckman and Smith, 1998; and Heckman, Smith and LaLonde, 1999). From assumption (iii), ((U_D, U_1) and (U_D, U_0) are independent of (Z, X)), we obtain that $U_1 - U_0$ is mean independent of X conditional on (D, Z), and thus (C-2b) can be rewritten as

$$E(U_1 - U_0 | Z = z, D = 1) = E(U_1 - U_0 | X = x, D = 1)$$

and (C-2b′) can be rewritten as

$$E(U_1 - U_0|Z = z, D = 1) = M.^{24}$$

Although we impose assumption (iii) to simplify the argument, it can be weakened substantially. Since all the IV analysis is performed conditional on $X = x$, assumption (iii) can be weakened to allow U_1, U_0 to depend on X (see Heckman, 1997; Heckman and Smith, 1998; Heckman, Smith and LaLonde, 1999).

A sufficient condition for both (C-2b) and (C-2b′) is that $\Pr(D = 1 \mid X, Z, U_1 - U_0) = \Pr(D = 1 \mid X, Z)$, i.e., U_D is independent of $U_1 - U_0$. This condition implies that $\Delta^{MTE}(x, u) = \mu_1(x) - \mu_0(x)$ for all u, and thus $\Delta^{IV}(x) = \Delta^{ATE}(x) = \Delta^{TT}(x, D = 1)$. This is another case where the implicit weighting of MTE implied by linear IV need not equal the weighting corresponding to ATE or TT and linear IV still identifies the parameters of interest since $\Delta^{MTE}(x, u)$ does not vary with u. In this case, the treatment effect varies over individuals conditional on X, but individuals do not participate in the program being evaluated on the basis of this variation.

The conditions under which (C-2b) and (C-2b′) hold, and whether these conditions are equivalent to each other or not, depend on the support of $P(Z)$ conditional on $X = x$ ($\mathcal{P}_x$). First consider the case where the support of $P(Z)$ conditional on $X = x$ is the full unit

[24] Conditioning on X is necessary on the right hand side of assumption (C-2b) since, in general, $U_1 - U_0$ will be mean dependent on X conditional on D even though $U_1 - U_0$ are mean independent of X conditional on (D, Z).

interval ($\mathcal{P}_x = [0, 1]$). Then both conditions are equivalent to the assumption that $U_1 - U_0$ is mean independent of U_D. To see this, note that the left hand side of (C-2b) can be rewritten as

$$E(U_1 - U_0 | X = x, Z = z, D = 1) = E(U_1 - U_0 | Z = z, D = 1) = E(U_1 - U_0 | U_D \leq \mu_D(z)).$$

By assumption, the support of $\mu_D(Z)$ contains the support of U_D conditional on $X = x$, so we have that (C-2b) is equivalent to assuming that $U_1 - U_0$ is mean independent of U_D. Now consider the left hand side of (C-2b$'$),

$$E(U_1 - U_0 | X = x, Z = z, D = 1) = E(U_1 - U_0 | Z = z, D = 1) = E(U_1 - U_0 | U_D \leq \mu_D(z)).$$

Again using the fact that the support of $\mu_D(Z)$ contains the support of U_D conditional on $X = x$, we have that condition (C-2b$'$) is equivalent to assuming that $U_1 - U_0$ is mean independent of U_D. This is another case where the implicit weighting of MTE implied by linear IV will not equal the weighting corresponding to ATE or TT but linear IV still identifies these parameters since $\Delta^{MTE}(x, u)$ does not vary with u.

Next consider the case where the support of $P(Z)$ conditional on $X = x$ is a set of two values, p' and p (i.e., $\mathcal{P}_x = \{p', p\}$). We then have that $\Delta^{IV}(x) = \Delta^{LATE}(x, p', p)$. Following the analysis of Section 5, if $p' = 0$, $p = 1$, we have that $\Delta^{IV}(x) = \Delta^{LATE}(x, 0, 1) = \Delta^{ATE}(x) = \Delta^{TT}(x, D = 1)$, and both conditions (C-2b) and (C-2b$'$) are satisfied allowing

$\Delta^{MTE}(x,u)$ to vary freely with u (i.e., allowing arbitrary dependence between $U_1 - U_0$ and U_D). In this special case, the implicit weighting of MTE implied by linear IV equals the weighting corresponding to ATE and TT, so that the parameters are equal even though $\Delta^{MTE}(x,u)$ varies freely with u.

Next consider the case where $p' = 0$ and $0 < p < 1$. The implicit weighting implied by linear IV equals the weighting corresponding to TT, $\Delta^{IV}(x) = \Delta^{LATE}(x,0,p) = \Delta^{TT}(x, D = 1)$, and condition (C-2b) is satisfied, even though $\Delta^{MTE}(x,u)$ may vary arbitrarily with u, thus allowing $U_1 - U_0$ to be arbitrarily dependent on U_D.[25] However, in this case, the weighting implied by $\Delta^{IV}(x)$ is not the same as the weighting corresponding to $\Delta^{ATE}(x)$. Thus $\Delta^{IV}(x) = \Delta^{ATE}(x)$ and condition (C-2b$'$) is satisfied iff $\Delta^{ATE}(x) = \Delta^{TT}(x, D = 1)$, i.e., iff $\frac{1-p}{p}\int_0^p E(U_1 - U_0|U_D = u)du = \int_p^1 E(U_1 - U_0|U_D = u)du$. This will not be satisfied in general. Imposing this condition is equivalent to imposing a particular restriction on the dependence of $U_1 - U_0$ on U_D. This restriction does not require mean independence, but is implied by mean independence. Thus, in the case where the support of $P(Z)$ conditional on $X = x$ is the set of two values, 0 and p ($\mathcal{P}_x = \{0, p\}$), conditions (C-2b) and (C-2b$'$) differ with (C-2b) satisfied without restrictions on the dependence of $U_1 - U_0$ on U_D but (C-2b$'$) is satisfied only if a peculiar restriction is imposed on this dependence. Finally, consider the case where $0 < p' < p < 1$. In this case, $\Delta^{IV}(x) = \Delta^{LATE}(x, p', p)$, and the implicit weighting produced by linear IV will not equal the weighting correspond-

[25] Arbitrary subject to the maintained assumption (ii) in Section 2.

ing to TT or ATE, and conditions (C-2b) and (C-2b′) will not hold, unless a particular restriction is imposed on the dependence between $U_1 - U_0$ and U_D, with the restriction being weaker than a mean independence assumption.

We have thus considered two extreme cases. First, where the support of $P(Z)$ conditional on $X = x$ is the full unit interval ($\mathcal{P}_x = [0,1]$), in which case (C-2b) and (C-2b′) are both equivalent to the assumption that $U_1 - U_0$ is mean independent of U_D. Second, we have also considered the case where the support of $P(Z)$ conditional on $X = x$ is a set of two points ($\mathcal{P}_x = \{p', p\}$), in which case (C-2b) and (C-2b′) will generally be different assumptions and will require restrictions on the dependence of $U_1 - U_0$ on U_D weaker than a mean independence assumption. Following the same arguments, we can consider any other support for the propensity score. In general, (C-2b) is a weaker condition than (C-2b′). Both conditions will be satisfied under a condition weaker than mean independence of $U_1 - U_0$ on U_D if the support of $P(Z)$ conditional on $X = x$ is a strict subset of the unit interval ($\mathcal{P}_x \subset [0,1]$). Both conditions will impose some restriction on the dependence of $U_1 - U_0$ on U_D except in the extreme case where the support of $P(Z)$ conditional on $X = x$ is only two points, with one of the two points being 0 ($\mathcal{P}_x = \{0, p\}$) for (C-2b) or only two points with one point being 0 and one point being 1 ($\mathcal{P}_x = \{0, 1\}$) for (C-2b′).

In the general case of a correlated random coefficient model where agents select into the program partly based on $U_1 - U_0$, or information correlated with it, we have

$$U_1 \neq U_0 \tag{C-3a}$$

and

$$E(U_1 - U_0|Z = z, D = 1) \neq E(U_1 - U_0|X = x, D = 1) \tag{C-3b}$$

for a set of z values having positive probability ($F_{Z|X=x}$), or

$$E(U_1 - U_0|Z = z, D = 1) = M(z) \tag{C-3b$'$}$$

where $M(z)$ is a nondegenerate function of z for z in the support of Z conditional on $X = x$. In this case the linear instrumental variable estimator does not identify TT if (C-3b) holds or ATE if (C-3b$'$) applies (Heckman, 1997). These cases arise in the Roy model discussed in the next section.

Even though the linear instrumental variable estimator breaks down for identifying TT or ATE in the third case, it still identifies LATE if the support of $P(Z)$ conditional on $X = x$ only contains two elements. In this general case, LIV can be used to identify all three treatment parameters. Under the support conditions presented in section 5, suitably reweighted and integrated versions of LIV identify or bound TT and ATE while linear IV does not.

We now show how to relate this analysis to the familiar case of the Roy model.

7 Additive Separability and the Roy Model

Pursuing the additively separable case further we obtain the following representation:

$$\Delta^{TT}(x, z, D = 1) = \mu_1(x) - \mu_0(x) + K(P(z))^{26}$$

where $K(P(z)) = E(U_1 - U_0|\tilde{U}_D \leq P(z)) = K_{11}(P(z)) - K_{01}(P(z))$.

Integrating out over the support of Z given $X = x$ and $D = 1$,

$$\begin{aligned}
\Delta^{TT}(x, D = 1) &= E(Y_1 - Y_0 \mid X = x, D = 1) \\
&= \int \Delta^{TT}(x, z, D = 1) dF_{Z|X,D}(z|x, 1) \\
&= \mu_1(x) - \mu_0(x) + \textstyle\int K(p) dF_{P(Z)|X,D}(p|x, 1) \\
&= \mu_1(x) - \mu_0(x) + \textstyle\int K(p) \dfrac{p}{E(P(Z)|X = x)} dF_{P(Z)|X}(p|x)
\end{aligned} \tag{20}$$

where $\int K(p) \dfrac{p}{E(P(Z)|X = x)} dF_{P(Z)|X}(p|x) = E(U_1 - U_0|\tilde{U}_D \leq P(Z))$.

For a fixed X, and two distinct values of Z,

$$\Delta^{LATE}(x, P(z), P(z')) = \mu_1(x) - \mu_0(x) + \frac{P(z)K(P(z)) - P(z')K(P(z'))}{P(z) - P(z')}.$$

In the limit as $z' \to z$,

[26] This representation is due to Heckman (1980), and Heckman and Robb (1985, 1986).

$$\lim_{z' \to z} \Delta^{LATE}(x, P(z), P(z')) = \mu_1(x) - \mu_0(x) + K(P(z))\frac{\partial P(z)}{\partial z} + P(z)\frac{\partial K(P(z))}{\partial z}.$$

This limit exists as a consequence of our assumptions. Working directly with limit $P(z') \to P(z)$,

$$\begin{aligned} \lim_{P(z') \to P(z)} \Delta^{LATE}(x, P(z), P(z')) &= \mu_1(x) - \mu_0(x) + K(P(z)) + \frac{\partial K(P(z))}{\partial P(z)} P(z) \\ &= \mu_1(x) - \mu_0(x) + K(P(z))[1 + \eta], \end{aligned}$$

where $\eta = \frac{\partial \ln K(P(z))}{\partial \ln P(z)}$ is the elasticity of the conditional mean of the unobservables with respect to the propensity score. The above is LIV as previously defined, and since $\Delta^{LIV}(x, P(z)) = \Delta^{MTE}(x, P(z))$, we obtain

$$\Delta^{MTE}(x, u) = \mu_1(x) - \mu_0(x) + K(u)[1 + \eta].$$

As noted by Heckman (1997) and Heckman and Smith (1998), $\Delta^{MTE}(x, u)$ is the "treatment on the treated" parameter for those with characteristics $X = x$ who would be indifferent between sector 1 and sector 0 if the instrument were externally set so that $P(Z) = u$. This is the parameter required for evaluating the marginal gross gain (exclusive of any costs of making the move) for persons of characteristics $P(Z) = u$ at the margin of indifference between sector 1 and sector 0. This is the parameter required for evaluating the effect of a marginal change in Z on the persons induced into (or out) of the program by the change.

It is *one* of the three parameters required in a cost-benefit analysis of the change.[27]

For the case of the classical selection model, (U_0, U_1, U_D) are mean zero joint normal random variables, independent of (Z, X). In this case, we obtain the following expressions for $P(z)$ and the treatment parameters:

$$
\begin{aligned}
P(z) &= \Phi\left(\frac{\mu_D(z)}{\sigma_{DU}}\right) \\
K(P(z)) &= \left[\frac{\sigma_{1U}-\sigma_{0U}}{\sigma_U}\right]\frac{\exp\{-\frac{1}{2}[\Phi^{-1}(P(z))]^2\}}{P(z)} \\
\Delta^{ATE}(x) &= \mu_1(x) - \mu_0(x) \\
\Delta^{TT}(x, P(z), D=1) &= \mu_1(x) - \mu_0(x) + \left[\frac{\sigma_{1U}-\sigma_{0U}}{\sigma_U}\right]\frac{1}{\sqrt{2\pi}}\frac{\exp\{-\frac{1}{2}[\Phi^{-1}(P(z)]^2\}}{P(z)} \\
\Delta^{LATE}(x, P(z), P(z')) &= \mu_1(x) - \mu_0(x) + \\
& \left[\frac{\sigma_{1U}-\sigma_{0U}}{\sigma_U}\right]\frac{1}{\sqrt{2\pi}}\left[\frac{\exp\{-\frac{1}{2}(\Phi^{-1}(P(z))^2\} - \exp\{(-\frac{1}{2}(\Phi^{-1}(P(z'))^2))\}}{P(z)-P(z')}\right] \\
\Delta^{MTE}(x, u) &= \mu_1(x) - \mu_0(x) - \left[\frac{\sigma_{1U}-\sigma_{0U}}{\sigma_{DU}}\right]\Phi^{-1}(u)
\end{aligned}
$$

where $\sigma_{jU} = [Var(U_j)]^{\frac{1}{2}}$. The expression for $\Delta^{TT}(x, D=1)$ is obtained from Equation (20) and depends on the distribution of $P(Z)$. Observe that in the normal case, $\Delta^{MTE}(x, \frac{1}{2}) = \Delta^{ATE}(x)$ and $\Delta^{LATE}(x, p, 1-p) = \Delta^{ATE}(x)$ as a result of the symmetry of the normal distribution.[28]

[27]The other parameters are the cost of the change $C'(Z)$ (where $C(Z)$ is the cost function) and the effects of the change in Z on the outcomes of persons who are not affected by the change in Z. For the complete definition, see Heckman (1997) and Heckman and Smith (1998) or see the discussion in Heckman and Vytlacil (2000b).

[28]These equalities will hold for any distribution such that (U_D, U_1) and (U_D, U_0) are jointly symmetric around their means. As in the work of Powell (1987) and Chen (1999), symmetry is one assumption that can be exploited to achieve identification without large support assumptions.

7.1 The Roy Model

Next consider the special case of a Roy model with sector-specific costs. In particular, consider a case where

$$\begin{aligned} D^* &= (Y_1 - Y_0) - \varphi(W) - V \\ &= \Big[(\mu_1(x) - \mu_0(x)) - \varphi(W)\Big] + \Big[(U_1 - U_0) - V\Big], \end{aligned}$$

and

$$D = 1 \text{ if } D^* \geq 0, \quad = 0 \text{ otherwise,}$$

where $\varphi(W) + U_D$ is the cost of being in sector 1, $Y_1 - Y_0$ is the gross benefit of being in sector 1, and D^* is the net benefit of being in sector 1. The agent will choose to be in sector 1 if the gross benefit exceeds the cost, i.e., if $Y_1 - Y_0 \geq \varphi(W) + U_D$. This model is similar to models used to analyze labor supply, unionism and educational and occupational choice. Taking $Z = (W, X)$, $U_D = U_1 - U_0 - V$, and $\mu_D(z) = \mu_1(x) - \mu_0(x) - \varphi(w)$, this model is a special case of the model previously developed. If we continue to maintain assumptions (i) - (v) from Section 2, then all analysis of section 5 continues to hold.[29] Recall assumption (i), that $\mu_D(Z)$ is a nondegenerate random variable conditional on $X = x$. In the Roy model considered here, $\mu_D(Z) = \mu_1(X) - \mu_0(X) - \varphi(W)$, so that this assumption requires that

[29] In addition, by imposing the Roy model structure, one can identify average cost of treatment parameters without observing any direct information on the cost of treatment. See Heckman and Vytlacil (2000b).

$\varphi(W)$ is a nondegenerate random variable conditional on X. In other words, we require variables that do not affect the outcome equations but that affect the cost of selecting into treatment. The requirement for $\varphi(W)$ to be a nondegenerate random variable conditional on X rules out the special case of the original Roy model with no costs. To analyze such a model, restrict $\varphi(W) = 0$ and $V = 0$, so $D^* = Y_1 - Y_0 = \mu_1(X) - \mu_0(X) + U_1 - U_0$ and $U_D = -(U_1 - U_0)$. Assumption (i) no longer holds, but we impose conditions (ii) - (v).

The treatment parameters are still well defined, the relationship between the parameters continues to hold, but we are no longer able to identify the parameters from LIV. To establish these claims, note that

$$\begin{aligned}
\Delta^{MTE}(x,u) &= \mu_1(x) - \mu_0(x) + E(U_1 - U_0 | U_1 - U_0 = -u) \\
&= \mu_1(x) - \mu_0(x) - u \\
\Delta^{ATE}(x,u) &= \mu_1(x) - \mu_0(x) \\
&= \int \Delta^{MTE}(x,u) dF_U \\
\Delta^{TT}(x, D=1) &= \mu_1(x) - \mu_0(x) + E\Big(U_1 - U_0 \Big| -(U_1 - U_0) \leq \mu_1(x) - \mu_0(x)\Big) \\
&= \tfrac{1}{\Pr(D=1|X=x)} \int_{-\infty}^{\mu_1(x)-\mu_0(x)} \Delta^{MTE}(x,u) dF_U
\end{aligned}$$

so that the same relationships exist among the parameters as established in the case when exclusion restrictions are present. However, there is no longer a well defined LIV estimator, since it is no longer possible to shift the index for the decision rule while holding X constant. Hence the instrumental variable argument breaks down in this case.

8 Conclusion

This paper uses an index model or latent variable model to model selection variable D to impose structure on a model of potential outcomes that originates with Neyman (1923), Fisher (1935), and Cox (1958). We introduce the Marginal Treatment Effect (MTE) parameter and its sample analogue, the local IV (LIV) estimator, as devices for unifying different treatment parameters. Different treatment effect parameters are averaged versions of the marginal treatment parameter which differ according to how they weight the marginal parameter. ATE weights all marginal parameters equally. LATE gives equal weight to the marginal treatment parameters within a given interval. TT gives a larger weight to those marginal treatment parameters corresponding to the treatment effect for individuals who have larger values of the unobserved proclivity to participate in the program. The weighting of the marginal treatment parameter for the treatment on the treated parameter is like that obtained in length-biased or sized-biased samples.

The local IV estimator identifies the marginal treatment effect under conditions (i)-(v). Identification of LATE and LIV depends on the support of the propensity score, $P(Z)$. The larger the support of $P(Z)$, the larger the set of LATE and LIV parameters that are identified. Identification of ATE depends on observing $P(Z)$ values arbitrarily close to 1 and $P(Z)$ values arbitrarily close to 0. When such $P(Z)$ values are not observed, ATE can be bounded and the width of the bounds is linearly related to the distance between 1 and the largest $P(Z)$ and the distance between 0 and the smallest $P(Z)$ value. For TT, identification requires that one observe $P(Z)$ values arbitrarily close to 0. If this condition does not hold, then the TT parameter can be bounded and the width of the bounds will be linearly related to the distance between 0 and the smallest $P(Z)$ value, holding $\Pr(D=1|X)$ constant.

Under full support conditions, the local IV estimator, when suitably weighted, can be used to estimate LATE, TT, and ATE in cases where linear IV cannot estimate these parameters. When support conditions fail, local IV can be use to bound LATE, TT, and ATE. Local IV is thus a more general and flexible estimation principle than linear IV, which in case of heterogeneous response to treatment, at most identifies LATE.

Implementation of the methods developed in this paper through either parametric or nonparametric estimators is straightforward. In joint work with Arild Aakvik, we have developed the sampling theory for the LIV estimator and empirically estimated and bounded various treatment parameters for a Norwegian vocational rehabilitation program (Aakvik, Heckman and Vytlacil, 1999). We have discussed several economic models where the LIV estimator can be fruitfully applied and we have examined the economic questions that the treatment effect parameters answer.

References

Aakvik, A., Heckman, J., and Vytlacil, E. (1999), "Local Instrumental Variables and Latent Variable Models for Estimating Treatment Effects," unpublished manuscript, University of Chicago.

Amemiya, T. (1985), *Advanced Econometrics* (Cambridge MA: Harvard University Press).

Angrist, J., Graddy, K., and Imbens, G. (1995), "Non-Parametric Demand Analysis with an Application to the Demand for Fish," NBER working paper No. T0178.

Angrist, J., Imbens, G., and Rubin, D. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association,* **91**, 444-455.

Bjorklund, A. and Moffitt, R. (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, **69**, 42-49.

Boadway, R. and Bruce, N. (1984), *Welfare Economics*, (Oxford: U.K. Blackwell).

Chen, S. (1999), "Distribution-Free Estimation of the Random Coefficient Dummy Endogenous Variable Model," *Journal of Econometrics,* **91**, 171-199.

Chipman, J. and Moore, J. (1976), "Why an Increase in GNP Need Not Imply an Improvement in Potential Welfare," *Kylos*, **29,** 391-418.

Cox, D. R. (1958), *The Planning of Experiments*, (New York: Wiley).

Das, M., and Newey, W. (1998), "Non-parametric Estimation of the Sample Selection Model," unpublished working paper, Massachusetts Institute of Technology.

Davidson, C. and Woodbury, S. (1993), "The Displacement Effect of Reemployment Bonus Programs," *Journal of Labor Economics*, **11**(4), 575-607.

Fisher, R. A. (1935), *Design of Experiments*, (London: Oliver and Boyd).

Gronau, R. (1974), "Wage Comparisons: A Selectivity Bias," *Journal of Political Economy*, **82**, 1119-1143.

Harberger, A. (1971), "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *Journal of Economic Literature*, **9**, 785-797.

Heckman, J. (1974), "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, **42**, 679-694.

_____ (1980), "Addendum to Sample Selection Bias as a Specification Error," in G. Farkas, ed., *Evaluation Studies Review Annual , Vol. 5,* Sage Publications.

_____ (1990), "Varieties of Selection Bias," *American Economic Review*, **80**, 313-318.

_____ (1992), "Randomization and Social Policy Evaluation," in C. Manski and I. Garfinkel, eds., *Evaluating Welfare and Training Programs*, (Cambridge, MA: Harvard University Press).

_____ (1997, first draft 1995), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* **32**, 441-462.

Heckman, J., and Honoré, B. (1990), "The Empirical Content of the Roy Model," *Econometrica*, **58**, 1121-1149.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1996), "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences,* **93**, 13416-13420.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, **66**, 1017-1098.

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

Heckman, J., LaLonde, R., Smith, J. (1999), "The Economics and Econometrics of Active Labor Market Programs," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics, Volume III*, (Amsterdam: Elsevier), forthcoming.

Heckman, J., Lochner, L., and Taber, C. (1998), "General Equilibrium Treatment Effects: A Study of Tuition Policy," *American Economic Review*, **88**, 381-386.

Heckman, J., and MaCurdy, T. (1985), "A Simultaneous Equations Linear Probability Model," *Canadian Journal of Economics*, **18**, 28-37.

Heckman, J., and Robb, R. (1985), "Alternative Methods for Estimating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, (New York: Cambridge University Press), 156-245.

_____ (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in H. Wainer, ed., *Drawing Inference From Self-Selected Samples*, (NY: Springer-Verlag), 63-107.

Heckman, J., and Smith, J. (1998, first draft 1995), "Evaluating The Welfare State," in S. Strom, ed., *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Econometric Society Monograph Series, (Cambridge, UK: Cambridge University Press).

Heckman, J., and Vytlacil, E. (1999a), "Identifying and Estimating Correlated Random Coefficient Models of the Impact of Schooling and Training on Earnings," unpublished working paper, University of Chicago.

_____ (1999b), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, **96**, 4730-4734.

——— (2000a) "The Relationship Between Treatment Parameters within a Latent Variable Framework," forthcoming, *Economic Letters.*

——— (2000b), "Econometric Evaluations of Social Programs," forthcoming in J. Heckman and E. Leamer, eds., *Handbook of Econometrics, Volume 5*, (North-Holland: Amsterdam).

Imbens, G., and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, **62**, 467-476.

Junker, B., and Ellis, J. (1997), "A Characterization of Monotone Unidimensional Latent Variable Models," *Annals of Statistics,* **25**, 1327-1343.

Kolmogorov, A. N., and Fomin, S. V. (1970), *Introductory Real Analysis*, R. Silverman, trans. (Mineola, NY: Dover Publications).

Laffont, J. J. (1989), *Fundamentals of Public Economics*, (Cambridge, MA: MIT Press).

Lewis, H. G. (1963), *Unionism and Relative Wages*, (Chicago: University of Chicago Press).

McFadden, D. (1981), "Econometric Models of Probabilistic Choice," in D. McFadden and C. Manski, eds., *Structural Analysis of Discrete Data,* (Cambridge, MA: MIT Press).

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

Moffitt, R., (1992), "Evaluation of Program Entry Effects," in C. Manski and I. Garfinkel, eds., *Evaluating Welfare and Training Programs*, (Cambridge, MA: Harvard University Press).

Neyman, J. (1923), "Statistical Problems in Agricultural Experiments," Supplement to the *Journal of Royal Statistical Society*, **2**(2), 107-180.

Powell, J. L. (1987), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, **54**, 1435-1460.

Quandt, R. (1972), "Methods for Estimating Switching Regressions," *Journal of The American Statistical Association*, **67**(338), 306-310.

_____ (1988), *The Econometrics of Disequilibrium*, (Oxford: Blackwell).

Rao, C. R. (1985), "Weighted Distributions," in A. Atkinson, and S. Feinberg, eds, *A Celebration of Statistics: The ISI Centenary Volume*, (Berlin: Springer-Verlag), 543-569.

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

Roy, A. (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, **3**, 135-146.

Rubin, D. (1978), "Bayesian Inference for Casual Effects: The Role of Randomization," *Annals of Statistics*, **6**, 34-58.

Thurstone, L. (1930), *The Measurement of Values*, (Chicago: University of Chicago Press).

—— (1959), "A Law of Comparative Judgement," *Psychological Review*, **34**, 273-286.

Vytlacil, E. (1999a), "Independence, Monotonicity, and Latent Variable Models: An Equivalence Result," working paper, University of Chicago.

—— (1999b), "Semiparametric Identification of the Average Treatment Effect in Nonseparable Models," working paper, University of Chicago.

Appendix: Relaxing Additive Separability and Independence

There are two central assumptions that underlie the latent index representation used in this paper: that U_D is independent of Z, and that U_D and Z are additively separable in the index. 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 the Imbens and Angrist assumptions imply 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 U_D and Z to be nonseparable in the treatment index:

$$\begin{aligned} D^* &= \mu_D(Z, U_D) \\ D &= 1 \text{ if } D^* \geq 0, \quad = 0 \text{ otherwise} \end{aligned}$$

while maintaining the assumption that Z is independent of (U_D, 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, $\mu_D(z, u) \geq \mu_D(z', u)$ for all u, or $\mu_D(z, u) \leq \mu_D(z', u)$ for all u.[30] Vytlacil (1999a) shows that monotonicity always implies one representation of μ_D as $\mu_D(z, u) = \mu(z) + u$. We now reconsider the analysis in the text without imposing the

[30]Note that the monotonicity condition is a restriction across u. For a given fixed u, it will always trivially have to be the case that either $\mu_D(z, u) \geq \mu_D(z', u)$ or $\mu_D(z, u) \leq \mu_D(z', u)$.

monotonicity condition by considering the latent index model without additive separability. Since we have imposed no structure on the $\mu_D(z, u)$ index, one can easily show that this model is equivalent to imposing the independence condition of Imbens and Angrist (1994) without imposing their monotonicity condition. A random coefficient discrete choice model with $\mu_D = Z\beta + \varepsilon$ where β and ε are random, and β can assume positive or negative values is an example of this case, *i.e.* $U_D = (\beta, \varepsilon)$.

We impose the regularity condition that, for any $z \in \text{Supp}(Z)$, $\mu_D(z, U_D)$ is absolutely continuous with respect to Lebesgue measure.[31] Let

$$\Omega(z) = \{u : \mu_D(z, u) \geq 0\},$$

so that

$$P(z) \equiv \Pr(D = 1 | Z = z) = \Pr(U_D \in \Omega(z)).$$

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') s.t. $P(z) = P(z')$ and $\Omega(z) \neq \Omega(z')$. In this case, we can no longer replace $Z = z$ with $P(Z) = P(z)$ in the conditioning sets.

[31] We impose this condition to ensure that $\Pr\big(\mu_D(z, U_D) = 0\big) = 0$ for any $z \in \text{Supp}(Z)$.

Define as before

$$\Delta^{MTE}(x,u) = E(\Delta|X=x, U_D=u).$$

For ATE, we obtain the same expression as before:

$$\Delta^{ATE}(x) = \int_0^1 E(\Delta|X=x, U_D=u)du.$$

For TT, we obtain a similar but slightly more complicated expression:

$$\begin{aligned}
\Delta^{TT}(x,z,D=1) &\equiv E(\Delta|X=x, Z=z, D=1) \\
&= E(\Delta|X=x, U_D \in \Omega(z)) \\
&= \frac{1}{P(z)}\int_{\Omega(z)} E(\Delta|X=x, U_D=u)du.
\end{aligned}$$

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 $\Delta^{TT}(x,z,D=1) \neq \Delta^{TT}(x,z',D=1)$.

Following the same derivation as used in the text for the TT parameter not conditional on Z,

$$\begin{aligned}
\Delta^{TT}(x,D=1) &\equiv E(\Delta|X=x, D=1) \\
&= \int E(\Delta|X=x, Z=z, D=1)dF_{Z|X,D}(z|x,1)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\Pr(D=1|X=x)}\int\left[\int_0^1 \mathbf{1}[u\in\Omega(z)]E(\Delta|X=x,U_D=u)du\right]dF_{Z|X}(z|x)\\
&= \frac{1}{\Pr(D=1|X=x)}\int_0^1\left[\int \mathbf{1}[u\in\Omega(z)]E(\Delta|X=x,U_D=u)dF_{Z|X}(z|x)\right]du\\
&= \int_0^1 E(\Delta|X=x,U_D=u)g_x(u)du
\end{aligned}
$$

where

$$
g_x(u) = \frac{\int \mathbf{1}[u\in\Omega(z)]dF_{Z|X}(z|x)}{\Pr(D=1|X=x)} = \frac{\Pr(D=1|U_D=u,X=x)}{\Pr(D=1|X=x)}.
$$

Thus the definitions of the parameters and the relationships among them that are developed in the main text of this paper 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 U_D and Z to be additively nonseparable in the treatment index. For simplicity, we will keep the conditioning on X implicit. This analysis essentially replicates the analysis of Imbens of Angrist (1994) using a latent index representation. Define the following sets

$$
A(z,z') = \{u : \mu_D(z,u)\geq 0, \mu_D(z',u)\geq 0\}
$$

$$
B(z,z') = \{u : \mu_D(z,u)\geq 0, \mu_D(z',u) < 0\}
$$

$$C(z,z') = \{u : \mu_D(z,u) < 0, \mu_D(z',u) < 0\}$$

$$D(z,z') = \{u : \mu_D(z,u) < 0, \mu_D(z',u) \geq 0\}.$$

Monotonicity implies that either $B(z,z')$ or $D(z,z')$ is empty. Suppressing the z, z' arguments, we have:

$$\begin{aligned}
E(Y|Z=z) &= \Pr(A \bigcup B)E(Y_1|A \bigcup B) + \Pr(C \bigcup D)E(Y_0|C \bigcup D) \\
E(Y|Z=z') &= \Pr(A \bigcup D)E(Y_1|A \bigcup D) + \Pr(B \bigcup C)E(Y_0|B \bigcup C)
\end{aligned}$$

so that

$$\begin{aligned}
\frac{E(Y|Z=z) - E(Y|Z=z')}{\Pr(D=1|Z=z) - \Pr(D=1|Z=z')} &= \frac{E(Y|Z=z) - E(Y|Z=z')}{\Pr(A \bigcup B) - \Pr(A \bigcup D)} \\
&= \frac{\Pr(B)E(Y_1 - Y_0|B) - \Pr(D)E(Y_1 - Y_0|D)}{\Pr(B) - \Pr(D)} \\
&= w_B E(\Delta|B) - w_D E(\Delta|D)
\end{aligned}$$

with

$$\begin{aligned}
w_B &= \frac{\Pr(B|B \bigcup D)}{\Pr(B|B \bigcup D) - \Pr(D|B \bigcup D)} \\
w_D &= \frac{\Pr(D|B \bigcup D)}{\Pr(B|B \bigcup D) - \Pr(D|B \bigcup D)}.
\end{aligned}$$

Under monotonicity, either $\Pr(B) = 0$ and LATE identifies $E(\Delta|D)$ or $\Pr(D) = 0$ and

LATE identifies $E(\Delta|B)$. 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 u. For $u \in A \bigcup C$, the denominator is zero and the parameter is not well defined. For $u \in B$, the parameter is $E(\Delta|U_D = u)$, for $u \in D$, the parameter is $E(\Delta|U_D = u)$. If we could restrict conditioning to $u \in B$ (or $u \in D$), we would obtain monotonicity within the restricted sample.

Now consider LIV. For simplicity, assume z is a scalar. Assume $\mu_D(z,u)$ is continuously differentiable in (z,u), with $\mu^j(z,u)$ denoting the partial derivative with respect to the jth argument. Assume that $\mu_D(z,U_D)$ 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 u points s.t. $\mu_D(z_0,u) = 0$. Let $j \in \mathcal{J} = \{1,...,L\}$ index the set of u evaluation points s.t. $\mu_D(z_0,u) = 0$, where L may be infinity, and thus write: $\mu_D(z_0,u_j) = 0$ for all $j \in \mathcal{J}$. Both the number of such evaluation points and the evaluation points themselves depends on the evaluation point, z_0, but we suppress this dependence for notational convenience.) One can show that

$$\frac{\partial}{\partial z}\Big[E(Y|Z = z_0)\Big] = \sum_{k=1}^{L} \frac{\mu^1(z_0,u_k)}{|\mu^2(z_0,u_k)|} E(\Delta|U_D = u_k)$$

and

$$\frac{\partial}{\partial z}[Pr(D=1|Z=z)] = \sum_{k=1}^{L} \frac{\mu^1(z,u_k)}{|\mu^2(z,u_k)|}.$$

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 $|\frac{\mu^1(z,u_k)}{\mu^2(z,u_k)}|$ does not vary with k. Using the fact that U_D is distributed unit uniform, we have

$$\begin{aligned}\Delta^{LIV}(z_0) = \Pr(\mu^1(z_0,U_D)>0|\mu(z_0,U_D)=0)E\Big(\Delta\Big|\mu_D(z_0,U_D)=0,\mu^1(z_0,U_D)>0\Big) \\ -\Pr(\mu^1(z_0,U_D)<0|\mu(z_0,U_D)=0)E\Big(\Delta\Big|\mu_D(z_0,U_D)=0,\mu^1(z_0,U_D)<0\Big).\end{aligned}$$

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 U_D and Z are separable in the treatment index while allowing them to be stochastically dependent:

$$\begin{aligned} D^* &= \mu_D(Z) - U_D \\ D &= 1 \text{ if } D^* \geq 0, \quad = 0 \text{ otherwise} \end{aligned}$$

with Z independent of (U_1, U_0), U_D distributed unit uniform, but allowing Z and U_D to be stochastically dependent. The analysis of Vytlacil (1999a) 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.[32]

We have

$$\Omega(z) = \{u : \mu(z) \leq u\}$$

and

$$P(z) \equiv \Pr(D = 1|Z = z) = \Pr(U \in \Omega(z)|Z = z).$$

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 U conditional on $Z = z$ need not equal the distribution of U 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.[33]

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 U_D

[32] To show that the monotonicity assumption implies a separable latent index model, one can follow the proofs of Vytlacil (1999a) with the only modification being replacing $P(z) = \Pr(D = 1|Z = z)$ 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.

[33] However, we again have equivalence between the alternative conditioning sets if we assume index sufficiency, i.e., that $F_{U|Z}(u|z) = F_{U|P(Z)}(u|P(z))$.

to be dependent. Now consider the TT parameter:

$$
\begin{aligned}
\Delta^{TT}(x,z,D=1) &= E(\Delta | X=x, Z=z, U_D \leq \mu_D(z)) \\
&= \tfrac{1}{P(z)} \int_0^{\mu_D(z)} E(\Delta | X=x, U_D=u) dF_{U|Z,X}(u|z,x) \\
&= \tfrac{1}{P(z)} \int_0^{\mu_D(z)} E(\Delta | X=x, U_D=u) \tfrac{f_{Z|U,X}(z|u,x)}{f_{Z|X}(z|x)} dF_U(u)
\end{aligned}
$$

where $f_{Z|X}$ and $f_{Z|U,X}$ denote the densities corresponding to $F_{Z|X}$ and $F_{Z|U,X}$ with respect to the appropriate dominating measure. We thus obtain

$$
\begin{aligned}
\Delta^{TT}(x,D=1) &= E(\Delta | X=x, U_D \leq \mu_D(Z)) \\
&= \frac{1}{\Pr(D=1|X=x)} \int \left[\int_0^{\mu_D(z)} E(\Delta | X=x, U_D=u) \frac{f_{Z|U,X}(z|u,x)}{f_{Z|X}(z|x)} dF_U(u) \right] dF_{Z|X}(z|x) \\
&= \frac{1}{\Pr(D=1|X=x)} \int \left[\int \mathbf{1}[u \leq \mu_D(z)] E(\Delta | X=x, U_D=u) \frac{f_{Z|U,X}(z|u,x)}{f_{Z|X}(z|x)} dF_{Z|X}(z|x) \right] dF_U(u) \\
&= \frac{1}{\Pr(D=1|X=x)} \int \left[\int \mathbf{1}[u \leq \mu_D(z)] E(\Delta | X=x, U_D=u) dF_{Z|U,X}(z|u,x) \right] dF_U(u) \\
&= \int_0^1 E(\Delta | X=x, U_D=u) g_x(u) du
\end{aligned}
$$

where again

$$
g_x(u) = \frac{\Pr(D=1|U_D=u, X=x)}{\Pr(D=1|X=x)}.
$$

We thus have that the definition of parameters and the relationships among the parameters that is developed in the text generalize naturally to the case where Z and U_D 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 U_D and Z to be stochastically dependent. We have

$$\begin{aligned}
E(Y|X &= x, Z = z) \\
&= P(z)\Big[E(Y_1|X = x, Z = z, D = 1)\Big] + (1 - P(z))\Big[E(Y_0|X = x, Z = z, D = 0)\Big] \\
&= \int_0^{\mu_D(z)} E(Y_1|X = x, U_D = u)dF_{U|X,Z}(u|x, z) + \int_{\mu_D(z)}^0 E(Y_0|X = x, U_D = u)dF_{U|X,Z}(u|x, z),
\end{aligned}$$

For simplicity, take the case where $\mu_D(z) > \mu_D(z')$. Then

$$\begin{aligned}
& E(Y|X = x, Z = z) - E(Y|X = x, Z = z') \\
= & \Big[\int_{\mu_D(z')}^{\mu_D(z)} E(Y_1|X = x, U_D = u)dF_{U|X,Z}(u|x, z) - \int_{\mu_D(z')}^{\mu_D(z)} E(Y_0|X = x, U_D = u)dF_{U|X,Z}(u|x, z')\Big] \\
& + \int_0^{\mu_D(z')} E(Y_1|X = x, U_D = u)\Big(dF_{U|X,Z}(u|x, z) - dF_{U|X,Z}(u|x, z')\Big) \\
& + \int_{\mu_D(z)}^1 E(Y_0|X = x, U_D = u)\Big(dF_{U|X,Z}(u|x, z) - dF_{U|X,Z}(u|x, z')\Big)
\end{aligned}$$

and thus

$$
\begin{aligned}
& \Delta^{LATE}(x,z,z') \\
= \quad & \delta_0(z)E(Y_1|X=x,Z=z,\mu_D(z') \leq U_D \leq \mu_D(z)) \\
& -\delta_0(z')E(Y_0|X=x,Z=z',\mu_D(z') \leq U_D \leq \mu_D(z)) \\
& +\left[\delta_1(z)E(Y_1|X=x,Z=z,U_D \leq \mu_D(z')) - \delta_1(z')E(Y_1|X=x,Z=z',U_D \leq \mu_D(z'))\right] \\
& +\left[\delta_2(z)E(Y_0|X=x,Z=z,U_D > \mu_D(z)) - \delta_2(z')E(Y_1|X=x,Z=z',U_D > \mu_D(z))\right]
\end{aligned}
$$

with

$$\delta_0(t) = \frac{\Pr(\mu_D(z') \leq U_D \leq \mu_D(z)|Z=t)}{\Pr(U_D \leq \mu_D(z)|Z=z,X=x) - \Pr(U_D \leq \mu_D(z')|Z=z',X=x)}$$

$$\delta_1(t) = \frac{\Pr(U_D \leq \mu_D(z')|Z=t)}{\Pr(U_D \leq \mu_D(z)|Z=z,X=x) - \Pr(U_D \leq \mu_D(z')|Z=z',X=x)}$$

$$\delta_2(t) = \frac{\Pr(U_D > \mu_D(z)|Z=t)}{\Pr(U_D \leq \mu_D(z)|Z=z,X=x) - \Pr(U_D \leq \mu_D(z')|Z=z',X=x)}$$

Note that $\delta_0(z) = \delta_0(z') = 1$ and the two terms in brackets are zero in the case where Z and U_D are independent. In the more general case, δ_0 may be bigger or smaller than 1, and the terms in brackets are of an 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

$f_{U|Z}(u|z)$ denote the density of U_D conditional on $Z = z$, and assume that this density is differentiable in z. Then, using equation (21), we have

$$\begin{aligned}\frac{\partial E(Y|X=x,Z=z)}{\partial z} &= E(\Delta|X=x,U_D=\mu_D(z))\mu_D'(z) \\ &\quad + \left[\int_0^{\mu_D(z)} E(Y_1|X=x,U_D=u)\frac{\partial f_{U|Z}(u|z)}{\partial z}du \right. \\ &\quad \left. + \int_{\mu_D(z)}^1 E(Y_0|X=x,U_D=u)\frac{\partial f_{U|Z}(u|z)}{\partial z}du\right],\end{aligned}$$

and

$$\frac{\partial \Pr(D=1|Z=z)}{\partial z} = \mu_D'(z) + \int_0^{\mu_D(z)} \frac{\partial f_{U|Z}(u|z)}{\partial z}du.$$

LIV is the ratio of the two terms. Thus, without the independence condition, the relationship between LIV and the MTE breaks down.

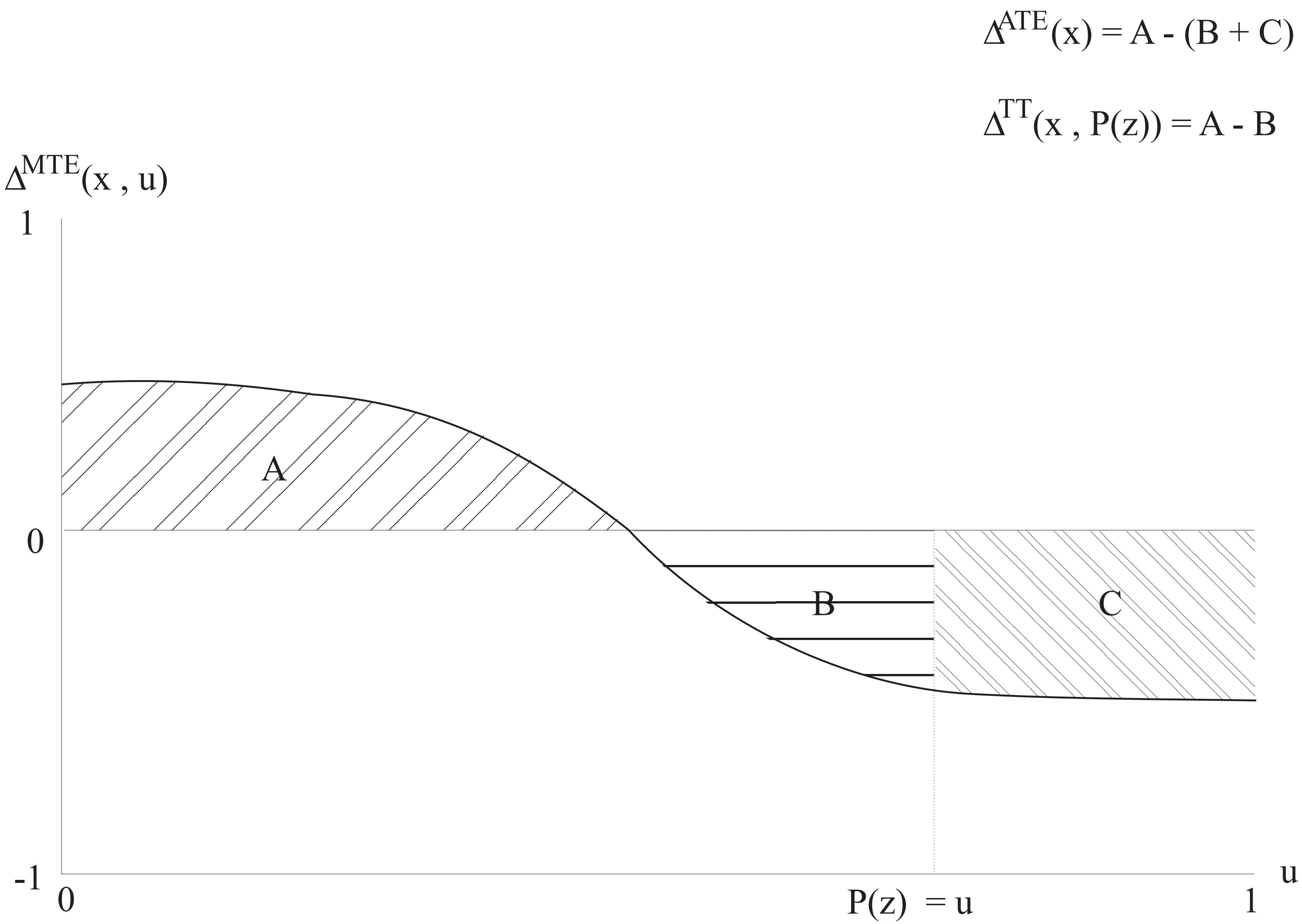

Figure 1: MTE integrates to ATE and TT under full support
(for dichotomous outcome)

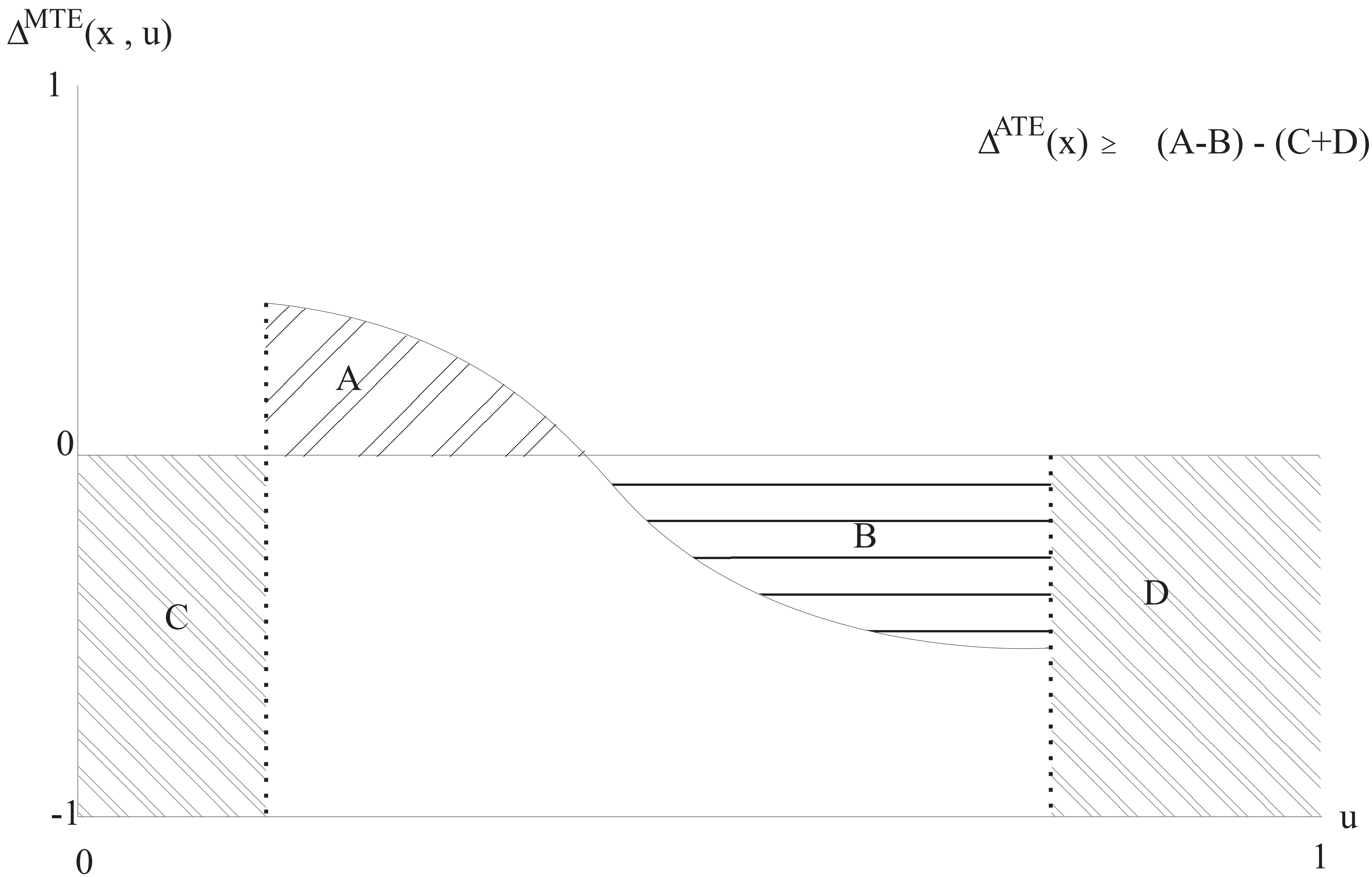

Figure 2: Lower bound on ATE under limited support
(for dichotomous outcome)

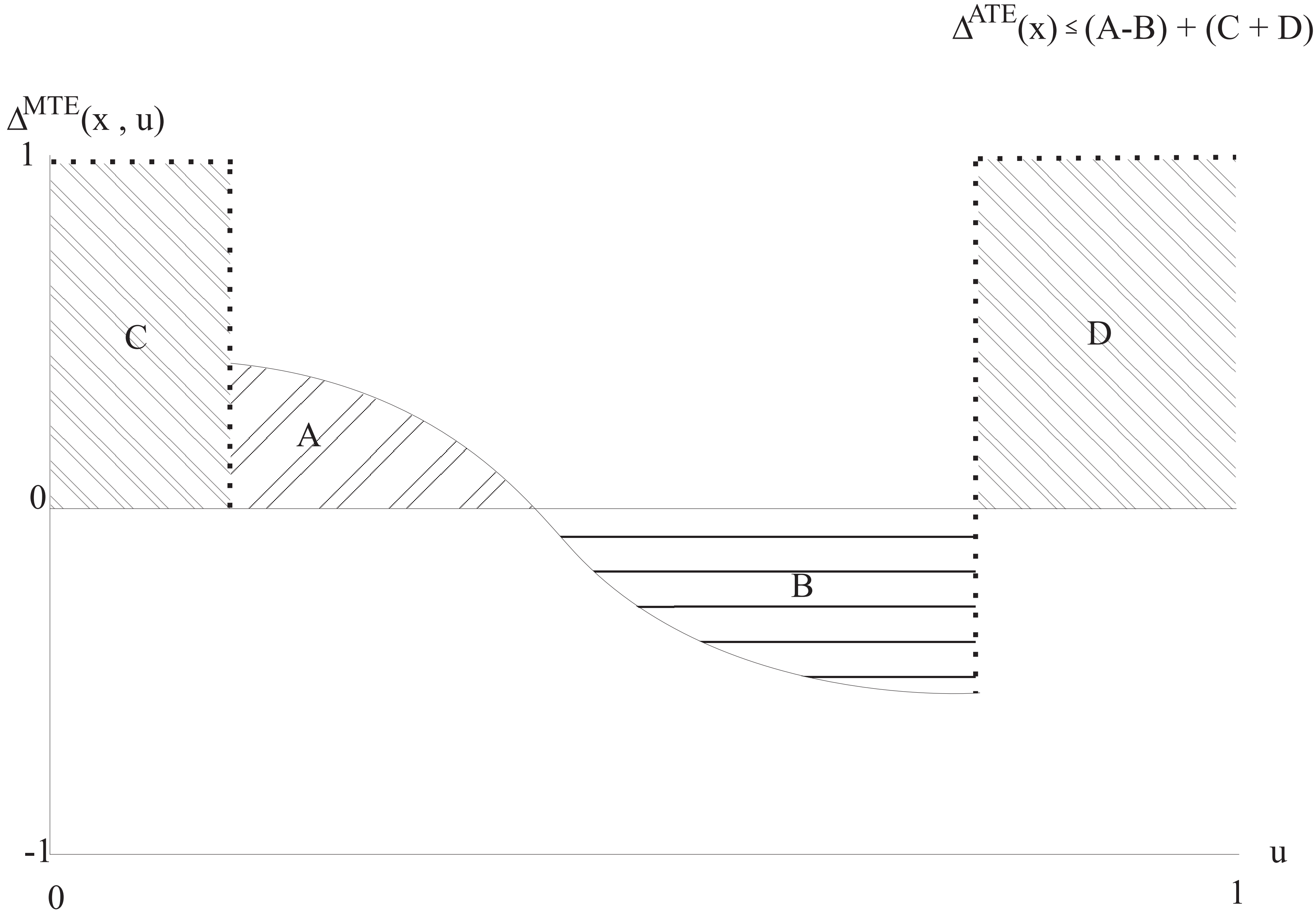

Figure 3: Upper bound on ATE under limited support
(for dichotomous outcome)