

Identification and Estimation of Treatment Effects in Sample Selection Models with a Monotone Instrument

Xin "Alex" Li
Department of Economics
University of North Carolina
Chapel Hill, NC 27599

January, 2005

Abstract

In the first half of this paper I use a bounding argument similar to that of Manski (1990, 1995), to clarify the identification power and limitations of the merits of the sample selection model with a monotone instrument, proposed by Heckman & Vytlacil (1999, 2000) and Vytlacil (2000). This type of model allows selection on un-observables and does not necessarily rely on an additive structure to recover the potential outcomes. Therefore it is theoretically and empirically attractive. I show that this model recovers the potential outcome distributions (and their functionals) of various "local" events, the events defined as groups of individuals characterized by the way they make the selection or participation decisions, but that it fails to do so when extended to the case with a polychotomous discrete treatment. Thus, it can identify various g -effects defined as a functional g of these conditional distributions under different support conditions of the propensity score. These effects include the well known average treatment effects (ATEs) and quantile treatment effects (QTEs). In case point identification fails one can derive the bounds for g -effects, with Heckman-Vytlacil (1999, 2000) bounds for ATE as a special case. Also, following Horowitz & Manski's (1995) argument, when using the parts of the sample space where propensity scores are close to the limits 0 and 1, the identification-on-limits results of smooth g -effects, including ATE and QTE, are derived.

In the second part, I propose methods to estimate average treatment effects (ATEs) in the sample selection model considered above. The building block of identification in this model is the expectation of observed outcomes conditioning on observed co-variates and the propensity score. The traditional control function approach uses an additive term to capture the influence of propensity score, which is applicable only in special cases of this model. Instead, I propose a generic two-stage method to estimate the conditional expectation term "locally" i.e. at a given value of the co-variates and the propensity score, in which any parametric and nonparametric estimation procedures can be used in both stages. I use this estimator to form estimates for different types of ATEs conditioning on given values of the co-variates. Second, using this estimator, I construct a matching estimator to approximate the ATE, specifically matching treated observations with high estimated propensity scores against untreated observations with low estimated propensity scores, and show it is \sqrt{n} convergent to a quantity approximating the ATE. A Monte Carlo experiment is conducted to assess its performance.

Key Word: Sample Selection, Selection on Unobservable, Treatment Effects, Identification, Bounds, Control function, Non-additive Model, Average Treatment Effects, Matching, Kernel Estimator.

1 Background

In studies of the effects of a binary 0-1 treatment, the observations are assumed to be draws from a set of (D, X, Y) , where $D \in \mathbf{A} = \{0, 1\}$ denotes the received treatment;

$X \in \mathbf{X} \subseteq R^m$ denotes the observed individual characteristics; and $Y \in \mathbf{Y} \subseteq R$ denotes the observed outcome. Usually, the researcher assumes that there exists a pair of potential (*ex ante*) outcomes (Y_0, Y_1) , under treatment 0 and 1, respectively. But only the outcome under the realized treatment can be observed as $Y = (1 - D)Y_0 + DY_1$. Also he/she assumes that the observed is generated by an underlying *ex ante* (unobserved) probability measure P on $(\mathbf{A} \times \mathbf{X} \times \mathbf{Y} \times \mathbf{Y}, \mathcal{B})$, and that the observed (or *ex post*) law will follow a probability measure P^* on $(\mathbf{A} \times \mathbf{X} \times \mathbf{Y}, \mathcal{B}^*)$, \mathcal{B} and \mathcal{B}^* being the correspondent Borel σ -fields. The observed P^* is consistent with P under the measurable map $f : (d, x, y_0, y_1) \mapsto (d, x, (1 - d)y_0 + dy_1)$. This means, for any event $B \in \mathcal{B}^*$, $P^*B = P(f^{-1}(B))$. The parameters of interest in the studies which need to be identified are, therefore, the *ex ante* law P , and/or some functionals of it, such as the conditional laws $P(Y_i|X)$, the conditional moments $E(g(Y_i)|X)$, and the conditional t -quantiles $Q_t(Y_i|X)$, for $i = 0, 1$.

There has been an extensive literature addressing the self-selection problem which arises when no random experiment with perfect compliance is available and the received treatment is the result of the self-selection of studied individuals. This means D is not necessarily independent of (X, Y_0, Y_1) . Rubin (1977) and a large body of following literature impose a selection-on-observables assumption, assuming D is independent of (Y_0, Y_1) conditional on X . With the help of this assumption, by stratifying the sample space according to X , one can look at each stratum at given $X = x$ and recover the conditional law $P(Y_i|X = x) = P(Y_i|X = x, D = i) = P^*(Y_i|X = x, D = i)$.¹ Many matching based estimators have been developed based on this conditional independence assumption. See Rubin (1977), Rosenbaum & Rubin (1983), Hahn (1998), Hirano, Imbens & Ridder (2002), and Firpo (2003). For a review of the literature see Imbens (2003).

Another line of research assumes a certain relationship between the *ex ante* outcomes (Y_0, Y_1) , such as sharing the same error or monotonicity, in order to derive bounds on, or to identify and to estimate functionals of, the *ex ante* law. See Doskum (1974), Lehmann (1974), Heckman, Smith & Clement (1997), Manski (1998), Chernozhukov & Hansen (2001), and Das (2000). Heckman (1976), Imbens & Newey (2002), and Heckman & Navarro-Lozano (2003) consider a control function approach.

On the other hand, Heckman (1990), and Heckman & Vytlacil (1999, 2000, 2001) model

¹ As far as an *ex post* event is concerned, one can use P instead of P^* because it is consistent with P .

the selection with a latent variable model with an instrument variable, without assuming any relationship between *ex ante* outcomes and the unobservables. Imbens & Angrist (1994), Abadie (2003), and Abadie, Angrist & Imbens (2002) use an instrumental variable framework, which is similar to the Heckman-type model and equivalent to it under certain conditions. (See Vytlacil 2002, also Chapter 2 of Li 2004.)

Each approach has its own merits and limitations. Also its validity depends on the validity of the imposed assumptions as to the real data generating *ex ante* law. However, the Heckman model/LATE approach has been widely used in many empirical studies. Thus it is of great value to clarify the parameters we can make inferences about in this framework. In particular, this paper investigates a variety of distributional effects that can be derived in this framework.

While most empirical studies focus on identifying and estimating the difference $E(Y_1|X) - E(Y_0|X)$, namely the average treatment effect (ATE), it is just one parameter that can be recovered. Imbens & Rubin (1997) consider the distributional effect in a binary instrument setting. Some recent papers (Abadie, Angrist & Imbens 2002, Chernozhukov & Hansen 2001) deal with the quantile treatment effect (QTE) $Q_t(Y_1|X) - Q_t(Y_0|X)$. In many cases, one evaluates a given policy intervention/treatment and is concerned with how the treatment affects the outcome distributions, not only the average but also over the whole spectrum, in some subgroups of the population. This paper shows that in the Heckman/LATE framework, under certain conditions one can identify any functional of the conditional laws $P(Y_i|X, E)$, E being some "local" events as defined in Section 4.² Thus this framework proves to be a good tool in evaluating the distributional effects of a self-selected treatment.

This paper is organized in this way. Sections 2-8 discuss the identification issues. Section 2 reviews Manski's bounding argument and relates it to the Heckman-Vytlacil model. Section 3 illustrates the identification in this model through bounding and differencing. Section 4 presents the point identification results. Section 5 provides the bounds for ΔD g -effects when they cannot be point identified. Section 6 analyzes the identification on the limits when the selection probability goes to its limits 0 and 1. Section 7 contends that the previous argument through bounding and differencing fails to achieve point identification in a model with a polychotomous treatment. Section 8 summarizes

² One example of "local" events E is the "complier" set $\{(\cdot)|D(Z = z_1) < D(Z = z_2), z_1 < z_2\}$ in LATE, the set of individuals who would choose $D = 0$ when assigned $Z = z_1$, and $D = 1$ when $Z = z_2$.

the identification results. Sections 9-13 considers the estimation of ATE. Section 9 reviews the identification results and related estimation problem. Section 10 proposes a two stage method to estimate the mean of observed outcomes conditioning on co-variables and the propensity score, by inverting the first stage estimate of the propensity score. This enables one to estimate/approximate the LATE, ATE on the treated, and ATE at a given value of $X = x$. Section 11 shows an asymptotically linear representation of the two-stage estimator in the previous section, \tilde{E} , in order to derive the distribution theory of an estimator pooling \tilde{E} within a part of the population. Section 12 gives the property of the pooling estimator, which is a valid approximation of the ATE within a part of the population. Section 13 provides a Monte Carlo simulation to evaluate the performance of the pooling ATE estimator. Section 14 concludes.

2 From Manski's Bounds to Heckman-Vytlacil Bounds

This section reviews the works of Manski (1990, 1995) and Heckman & Vytlacil (1999, 2000). Beginning with a worst-case-scenario without any assumption about the *ex ante* law, Manski (1990) bounds the *ex ante* conditional laws $P(Y_i|X)$ by the functionals of observed *ex post* law P^* , such that respectively for $i = 0, 1$:

$$\begin{aligned} P^*(Y \leq y|X, D = i)P^*(D = i|X) &\leq P(Y_i \leq y|X) \leq \\ &\leq P^*(Y \leq y|X, D = i)P^*(D = i|X) + P^*(D \neq i|X). \end{aligned}$$

Also in the literature, many assume the existence of at least one variable Z , called the excluded or instrument variable, which conditional on other variables in X ,³ may affect the selection of D , but not the outcomes Y_i . That is, to assume $P(Y_i|X, Z) = P(Y_i|X)$. Under this exclusion restriction, Manski (1990, 1995) modifies the bounds to:

$$\begin{aligned} \sup_z P^*(Y \leq y|X, D = i, Z = z)P^*(D = i|X, Z = z) &\leq P(Y_i \leq y|X) \leq \\ \leq \inf_z [P^*(Y \leq y|X, D = i, Z = z)P^*(D = i|X, Z = z) &+ P^*(D \neq i|X, Z = z)]. \end{aligned}$$

Suppose the instrument Z does affect the conditional treatment or participation probability $P(D = 1|X, Z)$, the well known "propensity score", that is, the propensity score varies when the value of the instrument changes. Thus the bounds could become tighter. The same argument also leads to the well-known point identification "on the limits":

³ Following the traditional notation, in this paper X denotes all the variables other than Z .

When for given X , one can observe a sequence of propensity scores arbitrarily close to 0 and 1, respectively, the *ex ante* $P(Y_i|X)$ can be point identified as the upper and lower bounds converge.

The sample selection model with a monotone instrument proposed by Heckman-Vytlacil makes further assumption about the relationship between D and Z . Vytlacil (2002) contends that under certain conditions, the Local Average Treatment Effect (LATE) approach is equivalent to a latent variable selection model. In this approach, using the propensity score $\Pi = P(D = 1|X, Z)$ as the instrument, as Li (2004), Chapter 2 points out, one can view the choice D when $\Pi = \pi$ as a random function $D(\pi)$ of the value π , (i.e. as a random element in the functional space $\mathbf{D}[-\varepsilon, 1 + \varepsilon]$), and the *ex ante* law P as a probability measure on the space $(\mathbf{D} \times [0, 1] \times \mathbf{X} \times \mathbf{Y} \times \mathbf{Y})$ of (d, π, x, y_0, y_1) . Also assuming D is almost surely monotone with respect to π , and is independent of the r.v. Π , one can use a model $D(\pi) = 1(\pi \geq \xi)$ to characterize the selection scheme, π being the value of Π and ξ being an r.v. supported on $[0, 1]$. Thus the *ex ante* law P is assumed to be a probability measure on the space $(\Xi \times [0, 1] \times \mathbf{X} \times \mathbf{Y} \times \mathbf{Y})$ ⁴ which satisfies $\Pi \perp (\xi, Y_0, Y_1)$ conditional on X , and the observed *ex post* law P^* follows the model (M1):

$$\begin{cases} X, \Pi; \\ D = 1(\Pi \geq \xi); \\ Y = (1 - D)Y_0 + DY_1. \end{cases}$$

Further let ξ have an absolutely continuous (marginal) law so that it follows a uniform law.⁵ This model does not assume any parametric relationship between (Y_0, Y_1) and X . Later on, all the analysis is done conditional on X . I just simplify the notation, dropping the "conditional on X " term for all conditional laws and for the propensity scores.

As $D(\pi)$ is monotone with respect to π , and (ξ, Y_1, Y_0) is independent of Π , one has

$$P(Y_1, D = 1|\Pi = \pi) = P(Y_1, \xi \leq \Pi|\Pi = \pi) = P(Y_1, \xi \leq \pi).$$

It is monotone with respect to π . Let $\bar{\pi}$ be the largest observed propensity score. Manski's bounding argument then leads to the (modified) Heckman-Vytlacil bounds (or H-V

⁴ The space of ξ , Ξ is topologically equivalent to the subset of \mathbf{D} in which the random function $D(\pi)$ resides.

⁵ ξ has a uniform law because here propensity scores are used as instruments, therefore $P(\xi \leq \pi) \equiv \pi$.

bounds) for $P(Y_1)$:⁶

$$P^*(Y \leq y | D = 1, \Pi = \bar{\pi})\bar{\pi} \leq P(Y_1 \leq y) \leq 1 - \bar{\pi} + P^*(Y \leq y | D = 1, \Pi = \bar{\pi})\bar{\pi}.$$

Following a similar argument, let $\underline{\pi}$ be the smallest observed propensity score, one can derive similar H-V bounds for $P(Y_0)$:

$$P^*(Y \leq y | D = 0, \Pi = \underline{\pi})(1 - \underline{\pi}) \leq P(Y_0 \leq y) \leq \underline{\pi} + P^*(Y \leq y | D = 0, \Pi = \underline{\pi})(1 - \underline{\pi}).$$

To point identify $(P(Y_0), P(Y_1))$, and then $E(Y_1 - Y_0)$, one needs $\bar{\pi} \uparrow 1$ and $\underline{\pi} \downarrow 0$.^{7 8}

3 Graphic Interpretation of Identification in the Heckman -Vytlacil Model

As seen in Section 2, the exclusion restriction along with a support condition for the propensity will provide the Manski's bounds, which are identical to Heckman-Vytlacil bounds if indeed the model M1 is valid. Thus a Heckman-Vytlacil model (M1) does provide H-V bounds as a special case of Manski's bounds but does not provide a much sharper tool to point identify $P(Y_i)$ than the Manski methods. This section shows the major merit of a Heckman-Vytlacil selection model is that it enables a "stratification" of the sample space on unobservables, which leads to the identification of various "local" events.

Figure 1 illustrates the basic concept of a model under the selection rule $D = 1(\Pi \geq \xi)$. Consider the case for the treated outcomes Y_1 . The model resorts to stratification of the sample space by (Π, ξ) . Let us put Π on the horizontal axis, and ξ on the vertical one.⁹ Because of the exclusion restriction, each vertical "slice" at $\Pi = \pi$ should reveal the same information about Y_1 as the whole population does. On each horizontal "slice" at $\xi = \xi_0$,

⁶ These bounds can be obtained from Heckman & Vytlacil (2000) by calculating the conditional expectations of $1(Y_1 \in (-\infty, y])$, which is a measurable function of Y_1 .

⁷ For two laws P_1 and P_2 , if $\forall y$ such that $P_1(-\infty, y] \leq P_2(-\infty, y]$, then P_1 (1st order) stochastically dominates P_2 . Later on, to simplify the notation, I will write it as $P_1 \succeq P_2$, instead of the inequality in the probability.

⁸ All the P^* here can be non-parametrically estimated in a sample when the propensity score Π is known. When Π is not known, one can follow a two-step procedure, first to obtain an estimate of Π , then use it to obtain an estimate of P^* . See Li 2004, Chapter 4.

⁹ It is not necessary that ξ has an absolutely continuous law. But assuming absolute continuity will make the following calculation much less complicated.

the conditional law $P(Y_1|\xi = \xi_0)$ may vary across the value of ξ_0 , because selection on unobservables is allowed. All the $P(Y_1|\xi = \xi_0), \xi_0 \in [0, 1]$, integrate to $P(Y_1)$.

$$\int_0^1 P(Y_1|\xi = \xi_0)d\xi_0 = P(Y_1). \quad (1)$$

But the selection rule dictates that only for the lower-right triangle part with $\Pi \geq \xi$ can one observe Y_1 . On the part of the slice $\{\Pi = \pi\}$ below the diagonal line $\{\Pi = \xi\}$, the conditional law of Y_1 is not necessarily the same as the *ex ante* $P(Y_1)$. But following the independence $\Pi \perp (\xi, Y_1)$, $P^*(Y|D = 1, \Pi = \pi) = P(Y_1|\Pi \geq \xi, \Pi = \pi) = P(Y_1|\xi \leq \pi)$ is identified, and $P(Y_1|\Pi < \xi, \Pi = \pi) = P(Y_1|\xi > \pi)$ is unknown. Let us denote ¹⁰

$$\begin{aligned} P_1^*(\pi) &\triangleq P^*(Y|D = 1, \Pi = \pi) = P(Y_1|\xi \leq \pi) \\ \text{and } P_1^N(\pi) &\triangleq P(Y_1|\Pi < \xi, \Pi = \pi) = P(Y_1|\xi > \pi). \end{aligned} \quad (2)$$

Following (1), $P(Y_1)$ is a mixture of $P_1^*(\pi)$ and $P_1^N(\pi)$:

$$P(Y_1) = \pi P_1^*(\pi) + (1 - \pi)P_1^N(\pi). \quad (3)$$

The larger the π is, the closer $P_1^*(\pi)$ is to $P(Y_1)$. Only for the longest observable slice at $\pi = 1$, one can surely have $P_1^*(\pi = 1) = P(Y_1)$. Thus one has the following result:

Proposition 3.1. Let " \succeq " denote the relationship of the first order stochastic dominance. Let P^∞ denote the law with unit mass at ∞ , and $P_{-\infty}$ denote the law with unit mass at $-\infty$. Given the fact that $P^\infty \succeq P_1^N(\pi) \succeq P_{-\infty}$, one has:

$$\pi P_1^*(\pi) + (1 - \pi)P^\infty \succeq P(Y_1) \succeq \pi P_1^*(\pi) + (1 - \pi)P_{-\infty}.$$

PROOF: The result follows from the argument (3) and the fact $P^\infty \succeq P_1^N(\pi) \succeq P_{-\infty}$. Similarly, $(1 - \pi)P_0^*(\pi) + \pi P^\infty \succeq P(Y_0) \succeq (1 - \pi)P_0^*(\pi) + \pi P_{-\infty}$.

Figure 2 shows that using the model (M1), it is possible to recover the information about Y_1 on the horizontal "slice" at $\xi = \xi_0$. Represented by below-the-diagonal-line sections of the two vertical "slices" at $\Pi = \pi_l, \pi_u$, $P_1^*(\pi_l) = P^*(Y|D = 1, \Pi = \pi_l)$ and $P_1^*(\pi_u) = P^*(Y|D = 1, \Pi = \pi_u)$ are known. It is intuitive that by differencing these two, one can get a conditional law on the horizontal slice, which I would name as the "Local Event". Indeed, following the independence that $\Pi \perp (\xi, Y_1)$, one has

¹⁰ Similarly we have $P_0^*(\pi) \triangleq P^*(Y|D = 0, \Pi = \pi) = P(Y_0|\Pi < \xi, \Pi = \pi) = P(Y_0|\xi > \pi)$ identified, and $P_0^N(\pi) \triangleq P(Y_0|\Pi \geq \xi, \Pi = \pi) = P(Y_0|\xi \leq \pi)$ unknown.

$\pi_l P_1^*(\pi_l) = \pi_l P(Y(1)|\xi \leq \pi_l) = P(Y(1), \xi \leq \pi_l)$ and $\pi_u P_1^*(\pi_u) = \pi_u P(Y(1)|\xi \leq \pi_u) = P(Y(1), \xi \leq \pi_u)$. This leads to the differencing:

$$P(Y_1, \pi_l < \xi \leq \pi_u) = \pi_u P_1^*(\pi_u) - \pi_l P_1^*(\pi_l). \quad (4)$$

This analysis reveals the sharpest bounds that Proposition 3.1 can provide:

Proposition 3.2.

a) $\forall \pi_l < \pi_u, \pi_l P_1^*(\pi_l) + (1 - \pi_l)P^\infty \succeq \pi_u P_1^*(\pi_u) + (1 - \pi_u)P^\infty$ and $\pi_u P_1^*(\pi_u) + (1 - \pi_u)P_{-\infty} \succeq \pi_l P_1^*(\pi_l) + (1 - \pi_l)P_{-\infty}$;

b) let $\bar{\pi}$ be the largest observed propensity score, $\bar{\pi} P_1^*(\bar{\pi}) + (1 - \bar{\pi})P^\infty \succeq P(Y_1) \succeq \bar{\pi} P_1^*(\bar{\pi}) + (1 - \bar{\pi})P_{-\infty}$ gives the sharpest dominance for $P(Y_1)$.

PROOF: Following the argument (4) $\pi_u P_1^*(\pi_u) = \pi_l P_1^*(\pi_l) + P(Y_1, \pi_l < \xi \leq \pi_u)$, therefore one has $(\pi_l/\pi_u)P_1^*(\pi_l) + [(1 - \pi_l)/\pi_u]P^\infty \succeq P_1^*(\pi_u)$ and $P_1^*(\pi_u) \succeq (\pi_l/\pi_u)P_1^*(\pi_l) + [(1 - \pi_l)/\pi_u]P_{-\infty}$. Thus a) is true. b) follows from a) and Proposition 3.1. ¹¹

REMARK 3.1. In a more general setting, one can put Π on the horizontal, and random function $D(\pi)$ on the vertical axis. This may be somehow improper as the space \mathbf{D} is not necessarily topologically equivalent to an interval. The longest observed vertical slice (i.e. the largest $P(Y_1, D = 1|\Pi = \pi)$) reveals the Manski's bounds. But without assuming the monotonicity that slice is not necessarily at the largest propensity score.

REMARK 3.2. The fact $0 \leq P_1^N \leq 1$ leads to the modified Heckman-Vytlacil bounds in Section 2.

Also $P(Y_1|\pi_l < \xi \leq \pi_u) = P(Y_1, \pi_l < \xi \leq \pi_u)/(\pi_u - \pi_l)$ is identified. ¹² Noting the fact that $\pi_l < \xi \leq \pi_u \Leftrightarrow D(\pi_l) < D(\pi_u)$, one recovers the information about Y_1 on the horizontal "belt", the "Local Event $\{D_l < D_u\}$ ". Actually, on several types of horizontal belts/slices one can recover the conditional law of Y_1 . Following Heckman & Vytlacil's (1999) terminology, let us call them as:

LOCAL EVENT: $A = \{\pi_l < \xi \leq \pi_u\}$. In Figure 2, it is represented by a horizontal belt such that $\pi_l < \xi \leq \pi_u$.

LIV EVENT: $B = \{\xi = \pi_u\}$, which is the case when π_l and π_u converge. It is represented by a horizontal slice such that $\xi = \pi_u$.

¹¹ Similarly, let $\underline{\pi}$ be the smallest observed propensity score, $(1 - \underline{\pi})P_0^*(\underline{\pi}) + \underline{\pi}P^\infty \succeq P(Y_0) \succeq (1 - \underline{\pi})P_0^*(\underline{\pi}) + \underline{\pi}P_{-\infty}$ gives the sharpest dominance for $P(Y_0)$.

¹² This result also follows (1).

TREATED EVENT: $T = \{\xi \leq \pi_u\}$, which is the case of event A when the lower limit π_l becomes $\underline{\pi} \downarrow 0$,¹³ and is represented by the horizontal belt that $\xi \leq \pi_u$.

THE WHOLE SPACE: This is the special case when lower limit π_l becomes $\underline{\pi} \downarrow 0$, and the upper limit π_u becomes $\bar{\pi} \uparrow 1$.¹⁴

The model M1 provides the means to recover the conditional distributions and functionals of them on these horizontal strata, the events conditional on the unobserved ξ which determines the latent decision rules. Let g be a functional $g : \Psi \mapsto R$, Ψ being the class of distributions. In practice, a functional like this can be the expectation, t -quantile, variance, or inter quantile distance. Borrowing from Heckman & Vytlačil's (1999) terminology, let us call:

LIV g -EFFECT: $g(P(Y_1|B)) - g(P(Y_0|B))$;

LOCAL g -EFFECT: $g(P(Y_1|A)) - g(P(Y_0|A))$;

g -EFFECT ON THE TREATED: $g(P(Y_1|T)) - g(P(Y_0|T))$;

g -EFFECT: $g(P(Y_1)) - g(P(Y_0))$.

When $g(P)$ is the expectation $\int h dP$ for a integrable function h , these effects are the well-known Average Treatment Effects (ATE) conditional on the given "local" events. When $g(P)$ is the t -quantile $Q_t(P)$, these effects can be named as Quantile Treatment Effects (QTE).

4 g -Effects for "Local" Events in Model M1

Let us follow the notation (2) such that $P_i^*(\pi) \triangleq P^*(Y|D = i, \Pi = \pi), i = 0, 1$. This section shows that following the characterization of Section 3, one can identify the following parameters:

REMARK 4.1. All the P^* terms (and any integral over them) can be estimated directly or following a two-step procedure exploiting the special property of the monotone instrument, to estimate Π first then to obtain a final estimate of P^* . Thus if the functional g takes an expectation form, the estimation of these varieties of g -effects will be straight-

¹³ Actually, when $\underline{\pi} \downarrow 0$ the limit event of A becomes $\{0 < \xi \leq \pi_u\}$. Assume the absolute continuity of ξ , it is different from the event T for only a 0-probability set. Further, if $P_i^*(\pi)$ is continuous around $\pi = 0$, there will be no problem to undertake the inference conditional on $\{\underline{\pi} < \xi \leq \pi_u\}$.

¹⁴ Assume the absolute continuity of ξ , and let $P_i^*(\pi)$ be continuous around $\pi = 0, 1$, the inference conditional on $\{\underline{\pi} < \xi \leq \bar{\pi}\}$.

forward, following the identification formulae. The two step procedure is described in Li (2004) Chapter 4.

Proposition 4.1. (Identification of LIV g -effects)

Under the model M1, for given $0 < \pi_u < 1$ where $P_1^*(\pi)$ and $P_0^*(\pi)$ are differentiable with respect to π :

a) With $P_1^*(\pi)$ observed for a neighborhood of π_u , $P(Y_1|B)$ is identified as

$$P_1^*(\pi_u) + \pi_u \frac{\partial P_1^*(\pi_u)}{\partial \pi};$$

b) With $P_0^*(\pi)$ observed for a neighborhood of π_u , $P(Y_0|B)$ is identified as

$$P_0^*(\pi_u) - (1 - \pi_u) \frac{\partial P_0^*(\pi_u)}{\partial \pi};$$

c) Consequently, $g(P(Y_1|B)) - g(P(Y_0|B))$ is identified;

d) Specifically $E(Y_1 - Y_0|B)$ and $Q_t(P(Y_1|B)) - Q_t(P(Y_0|B))$ are identified.

PROOF: See the calculation in Appendix 2.

REMARK 4.2. While the LIV average effects (i.e. when g takes an expectation form) can be readily estimated following the above identification results, the estimation of the LIV quantile effect may not be so direct. One possible way will be estimate an Local quantile effect around π_u . Using an analysis similar to the one in section 6,¹⁵ the obtained Local QTE will converge to the LIV QTE.

Proposition 4.2. (Identification of Local g -effects)

Under the model M1, for given pair of $\pi_l < \pi_u$:

a) With $P_1^*(\pi)$ observed for $\{\pi_l, \pi_u\}$, $P(Y_1|A)$ is identified as

$$P(Y_1|A) = \frac{\pi_u P_1^*(\pi_u) - \pi_l P_1^*(\pi_l)}{\pi_u - \pi_l}.$$

It can also be viewed as the average of $P(Y_1|B)$ over $\pi \in (\pi_l, \pi_u]$;

b) With $P_0^*(\pi)$ observed for $\{\pi_l, \pi_u\}$, $P(Y_0|A)$ is identified as

$$P(Y_0|A) = \frac{(1 - \pi_l)P_0^*(\pi_l) - (1 - \pi_u)P_0^*(\pi_u)}{\pi_u - \pi_l}.$$

It can also be viewed as the average of $P(Y_0|B)$ over $\pi \in (\pi_l, \pi_u]$;

c) Consequently, $g(P(Y_1|A)) - g(P(Y_0|A))$ is identified;

d) Specifically $E(Y_1 - Y_0|A)$ and $Q_t(P(Y_1|A)) - Q_t(P(Y_0|A))$ are identified.

¹⁵ Section 6 contains an analysis at $\pi = 0, 1$, here one needs a similar analysis at π_u .

PROOF: See the calculation in Appendix 1.

REMARK 4.3. $E(Y_1 - Y_0|A)$ can be related to the LATE in Imbens & Angrist (1994). And $Q_t(P(Y_1|A)) - Q_t(P(Y_0|A))$ can be compared with the LQTE in Abadie, Angrist & Imbens (2002). In Section 3 and in Proposition 4.2, one can see that $P(Y_i|A)$ is identified by differencing two conditional distributions. In the binary instrument setting, this fact has been stated by Imbens & Rubin (1997) and is the fundamental reason that the estimation of LATE/LQTE could involve a weighted regression with negative weights. (See Imbens & Rubin 1997, Abadie 2003, Abadie, Angrist & Imbens 2002)

Corollary 4.2a. LATE

$$P(Y_1|A) - P(Y_0|A) = \frac{\pi_u P_1^*(\pi_u) + (1 - \pi_u) P_0^*(\pi_u) - \pi_l P_1^*(\pi_l) - (1 - \pi_l) P_0^*(\pi_l)}{\pi_u - \pi_l}.$$

Consequently, provided the following integrals exist for a measurable function h ,

$$E(h(Y_1) - h(Y_0)|A) = \frac{E(h(Y)|\Pi = \pi_u) - E(h(Y)|\Pi = \pi_l)}{\pi_u - \pi_l}.$$

PROOF: The first equation follows directly from Proposition 4.2. The second one follows by integrating over the first equation. This is the same as Theorem 1 of Imbens & Angrist (1994).

Corollary 4.2b. LQTE

In the binary-instrument setting of Abadie (2003) and Abadie, Angrist & Imbens (2002), one has:

$$\begin{aligned} P(Y_1|A) &= E(1(Y_1 \leq y) | D(Z=1) > D(Z=0)) \\ &= \frac{1}{P(D(Z=1) > D(Z=0))} E\left[\left(1 - \frac{D(1-Z)}{P^*(Z=0)} - \frac{(1-D)Z}{P^*(Z=1)}\right) 1(Y_1 \leq y)\right]. \end{aligned}$$

For $P(Y_0|A)$ one has a similar result.

PROOF: See the calculation in Appendix 3.

REMARK 4.4. This result agrees with Abadie, Angrist & Imbens (2002) and Abadie (2003). Proposition 4.2 suggests that one assign observations $(\pi_u, D = 1)$ with weight $\pi_u/(\pi_u - \pi_l)$ and observations $(\pi_l, D = 1)$ with weight $-\pi_l/(\pi_u - \pi_l)$, in order to estimate $P(Y_1|A)$. Similarly, in order to estimate $P(Y_0|A)$, the weights should be $(1 - \pi_l)/(\pi_u - \pi_l)$ for $(\pi_l, D = 0)$ and $-(1 - \pi_u)/(\pi_u - \pi_l)$ for $(\pi_u, D = 0)$.

REMARK 4.5. The g -effect on the treated and g -effect can be viewed as special cases of

the Local g -effect. The former has $(\pi_l = \underline{\pi} \downarrow 0, \pi_u)$, the latter has $(\pi_l = \underline{\pi} \downarrow 0, \pi_u = \bar{\pi} \uparrow 1)$.

Proposition 4.3. (Identification of g -effects on the treated)

Under the model M1, assume $P_i^*(\pi)$ is continuous with respect to π at $\pi = 0$, for given π_u :

- a) With $P_1^*(\pi)$ observed for π_u , $P(Y_1|T)$ is identified as $P(Y_1|T) = P_1^*(\pi_u)$. It can also be viewed as the average of $P(Y_1|B)$ over $\pi \in (0, \pi_u]$;
- b) With $P_0^*(\pi)$ observed for π_u and for $\underline{\pi}$ arbitrarily close to 0, $P(Y_0|T)$ is identified as

$$P(Y_0|T) = \lim_{\underline{\pi} \downarrow 0} \frac{P_0^*(\underline{\pi}) - (1 - \pi_u)P_0^*(\pi_u)}{\pi_u - \underline{\pi}}.$$

It can also be viewed as the average of $P(Y_0|B)$ over $\pi \in (0, \pi_u]$;

- c) Consequently, $g(P(Y_1|T)) - g(P(Y_0|T))$ is identified;
- d) Specifically $E(Y_1 - Y_0|T)$ and $Q_t(P(Y_1|T)) - Q_t(P(Y_0|T))$ are identified.

PROOF: It follows Proposition 4.2, with the fact that $\lim_{\underline{\pi} \downarrow 0} \underline{\pi} P_1^*(\underline{\pi}) = 0$.

Proposition 4.4. (Identification of g -effects)

Under the model M1, assume $P_i^*(\pi)$ is continuous with respect to π at $\pi = 0, 1$:

- a) With $P_1^*(\pi)$ observed for $\bar{\pi}$ arbitrarily close to 1, $P(Y_1)$ is identified as $P(Y_1) = \lim_{\bar{\pi} \uparrow 1} P_1^*(\bar{\pi})$. It is also the average of $P(Y_1|B)$ over $\pi \in (0, 1]$;
- b) With $P_0^*(\pi)$ observed for $\underline{\pi}$ arbitrarily close to 0, $P(Y_0)$ is identified as $P(Y_0) = \lim_{\underline{\pi} \downarrow 0} P_0^*(\underline{\pi})$. It is also the average of $P(Y_0|B)$ over $\pi \in (0, 1]$;
- c) Consequently, $g(P(Y_1)) - g(P(Y_0))$ is identified;
- d) Specifically $E(Y_1 - Y_0)$ and $Q_t(P(Y_1)) - Q_t(P(Y_0))$ are identified.

PROOF: It follows Proposition 4.2, with the fact that $\lim_{\underline{\pi} \downarrow 0} \underline{\pi} P_1^*(\underline{\pi}) = 0$ and that $\lim_{\bar{\pi} \uparrow 1} (1 - \bar{\pi}) P_0^*(\bar{\pi}) = 0$.

REMARK 4.6. It is necessary that in the model M1, one needs $\bar{\pi} \uparrow 1$ and $\underline{\pi} \downarrow 0$ to identify the g -effects.

REMARK 4.7. Following the definition in Section 3, the g -effects and the g -effects on the treated can be estimated as special cases of the Local g -effects.

5 Bounds for ΔD g -Effects

When one does not have $\bar{\pi} \uparrow 1$ nor $\underline{\pi} \downarrow 0$, the g -effects can not be point identified. But it is still possible to derive bounds for these effects. Proposition 3.1 states that $\bar{\pi}P_1^*(\bar{\pi}) + (1 - \bar{\pi})P^\infty$ stochastically dominates $P(Y_1)$, which in turn dominates $\bar{\pi}P_1^*(\bar{\pi}) + (1 - \bar{\pi})P_{-\infty}$. And Similarly $(1 - \underline{\pi})P_0^*(\underline{\pi}) + \underline{\pi}P^\infty$ stochastically dominates $P(Y_0)$, which in turn dominates $(1 - \underline{\pi})P_0^*(\underline{\pi}) + \underline{\pi}P_{-\infty}$.

Horowitz & Manski (1995) shows that one can derive the bounds for a g -effect as long as the functional g preserves stochastic dominance. That is, the functional g satisfies:

$$g(P^A) \leq g(P^B), \text{ if } P^A \succeq P^B, \dots \text{Type I;}$$

or

$$g(P^A) \geq g(P^B), \text{ if } P^A \succeq P^B, \dots \text{Type II.}$$

One example of Type I functionals is the value of the c.d.f. $P(Y \leq a)$ at a given constant a . The examples of Type II functionals are the expectations and quantiles. Following the terminology of Manski (1998), the g -effects for these functionals are branded as ΔD g -effects. The bounds for ΔD g -effects in the model M1 are derived in the following proposition.

Proposition 5.1.

In model (M1), let $\bar{\pi}$ and $\underline{\pi}$ be the largest and smallest observed propensity scores respectively, for Type I g ,

$$g[\bar{\pi}P_1^*(\bar{\pi}) + (1 - \bar{\pi})P^\infty] \leq g[P(Y_1)] \leq g[\bar{\pi}P_1^*(\bar{\pi}) + (1 - \bar{\pi})P_{-\infty}];$$

$$g[(1 - \underline{\pi})P_0^*(\underline{\pi}) + \underline{\pi}P^\infty] \leq g[P(Y_0)] \leq g[(1 - \underline{\pi})P_0^*(\underline{\pi}) + \underline{\pi}P_{-\infty}].$$

For Type II g , just change the signs " \leq " to " \geq ". These bounds are sharp.

PROOF: This proposition directly follows Proposition 3.2 and Proposition 2.4 of Horowitz & Manski (1995). Thus the proof is omitted here.

REMARK 5.1. As the value of the distribution $P(Y \leq a)$ is Type I functional, one can use the above result to get the modified H-V bounds in Section 2. Both the mean and the t -quantile are Type II functionals. Thus one has to use the " \geq " results for type II. Furthermore, if the support of Y is bounded on $[L, U]$, one can simply change $(P_{-\infty}, P^\infty)$ to (P_L, P^U) , P_L and P^U being distributions concentrated on singletons $\{L\}$

and $\{U\}$ respectively. Thus Proposition 5.1 implies the Heckman & Vytlacil (1999, 2000) bounds for ATE.

6 Identification of Smooth g -Effects on the Limit

On the other hand, extrapolation and estimation-on-limit methods use only part of the data with propensity scores above a cut-off point $\bar{\pi}$ for the treated $D = 1$ group and/or below a $\underline{\pi}$ for the untreated $D = 0$ group. These methods may also let $\bar{\pi} \uparrow 1$ and $\underline{\pi} \downarrow 0$ when the sample size grows. When $\bar{\pi} \uparrow 1$ and $\underline{\pi} \downarrow 0$, one needs an identification-on-limits analysis to show the behavior of a smooth g -effect on the limits. The argument (3) in Section 3 shows that $P(Y_1)$ is a $(\bar{\pi}, 1 - \bar{\pi})$ mixture of $P_1^*(\bar{\pi})$ and an unknown distribution $P_1^N(\bar{\pi})$, such that $P(Y_1) = \bar{\pi}P_1^*(\bar{\pi}) + (1 - \bar{\pi})P_1^N(\bar{\pi})$. Therefore, $P_1^*(\bar{\pi}) = P(Y_1) - [(1 - \bar{\pi})/\bar{\pi}]P(Y_1) + [(1 - \bar{\pi})/\bar{\pi}]P_1^N(\bar{\pi})$. Let $\lambda = [(1 - \bar{\pi})/\bar{\pi}]$. $\bar{\pi} \uparrow 1$ means $\lambda \downarrow 0$ and $P_1^*(\bar{\pi}) \xrightarrow{W} P(Y_1)$. For a continuous g , $g(P_1^*(\bar{\pi})) \rightarrow g(P(Y_1))$.

Proposition 6.1. If $g : \Psi \mapsto R$ is continuous with respect to the weak topology of distributions (probability measures), $g(P_1^*(\bar{\pi})) \rightarrow g(P(Y_1))$ when $\bar{\pi} \uparrow 1$. Similarly $g(P_0^*(\underline{\pi})) \rightarrow g(P(Y_0))$ when $\underline{\pi} \downarrow 0$.

REMARK 6.1. This proposition validates the inference based on $g(P_1^*(\bar{\pi}))$ and $g(P_0^*(\underline{\pi}))$, used as the base for as extrapolation in several non/semi-parametric methods. The mean and t -quantile (with continuous p.d.f at Q_t) are continuous functionals.

When g is suitably differentiable, the argument in Proposition 5 of Horowitz & Manski (1995) can be applied, which provides a linear approximation of $g(P_1^*(\bar{\pi})) - g(P(Y_1))$. Let $P_1^N(\bar{\pi}) \in \Psi_1^N(\bar{\pi})$ be a proper class of these unknown distributions, and $P(Y_1) \in \Psi_1$ be a class of the possible true distribution. $\Psi_1^N(\bar{\pi}) - \Psi_1$ is the class of signed measure $P_1^N(\bar{\pi}) - P(Y_1)$. A proper type of differentiability for g is that at $P(Y_1)$, g is Hadamard-differentiable tangentially to $\Psi_1^N(\bar{\pi}) - \Psi_1$. That is, there exists a map $g' : (\Psi_1^N(\bar{\pi}) - \Psi_1) \mapsto R$, as $\lambda \downarrow 0$:

$$\sup_{P_1^N(\bar{\pi}) \in \Psi_1^N(\bar{\pi})} |g[P(Y_1) + (P_1^N(\bar{\pi}) - P(Y_1))\lambda] - g[P(Y_1)] - g'[P_1^N(\bar{\pi}) - P(Y_1)]\lambda| = o(\lambda)$$

such that $o(\lambda)$ term may depend on $P(Y_1)$ but is uniform for $P_1^N(\bar{\pi}) \in \Psi_1^N(\bar{\pi})$.

Proposition 6.2. Suppose that g is differentiable at $P(Y_1)$ as defined above, as $\bar{\pi} \uparrow 1$ and $\lambda \downarrow 0$, one has: (provided the supremum and infimum exist.)

$$\begin{aligned} g[P_1^*(\bar{\pi})] + \lambda \inf_{P_1^N(\bar{\pi}) \in \Psi_1^N(\bar{\pi})} g'[P_1^N(\bar{\pi}) - P(Y_1) + o(\lambda)] &\leq g[P(Y_1)] \leq \\ &\leq g[P_1^*(\bar{\pi})] + \lambda \sup_{P_1^N(\bar{\pi}) \in \Psi_1^N(\bar{\pi})} g'[P_1^N(\bar{\pi}) - P(Y_1)] + o(\lambda). \end{aligned}$$

PROOF: Following the mixture argument (3) in Section 3, this proposition is a special case of Proposition 5 of Horowitz & Manski (1995). The same proof will go through. Also for $P(Y_0)$, a similar analysis holds when $\underline{\pi} \downarrow 0$.

REMARK 6.2. This proposition guarantees that when dealing with a differentiable g , the extrapolation or estimation-on-limit such as in Andrews & Schafgans (1998) will have an asymptotically linearly (with respect to the employed cut-off propensity scores) shrinking bias, when the upper and lower cut-off points go to 1 and 0 respectively.

REMARK 6.3. When g takes an expectation form $g(P) = \int h dP$, it is linear with respect to P and the derivative $g'[P_1^N(\bar{\pi}) - P(Y_1)] = \int h d(P_1^N(\bar{\pi}) - P(Y_1))$. When one imposes the assumption that $h(Y)$ has a bounded support $[L, U]$ on the class $\Psi_1^N(\bar{\pi}) - \Psi_1$, g' is bounded and $g[P(Y_1) + \lambda(P_1^N(\bar{\pi}) - P(Y_1))] - g[P(Y_1)] = \lambda g'[P_1^N(\bar{\pi}) - P(Y_1)]$. In this case, one can even drop the $o(\lambda)$ term in Proposition 6.2, to get exact bounds. This result implies H-V bounds, which are linearly shrinking in $\bar{\pi}$ as $\bar{\pi} \uparrow 1$ and in $\underline{\pi}$ as $\underline{\pi} \downarrow 0$.

REMARK 6.4. When g is the t -quantile for a distribution which is continuous differentiable in a neighborhood $F^{-1}[t - \varepsilon, t + \varepsilon]$, g is Hadamard-differentiable tangentially to the set of distributions continuous on $F^{-1}[t - \varepsilon, t + \varepsilon]$. (See Lemma 9.23, van der Vaart & Wellner 2000) That is, g is differentiable with respect to a continuous perturbation of the distribution. The derivative map is $g' : \alpha \mapsto \alpha(Q_t(Y_i))/f_i(Q_t(Y_i))$, f_i being the p.d.f. of *ex ante* $P(Y_i)$ and Q_t being the true t -quantile value. With the assumptions of a bounded support $[L, U]$ and a bounded positive p.d.f for the *ex ante* $P(Y_i)$ in a neighborhood of $Q_t(Y_i)$, one can get an approximately linearly (in $\bar{\pi}$ and $\underline{\pi}$) shrinking bounds for the quantile treatment effects.

7 Failure of Point Identification under a Polychotomous Discrete Treatment

In some empirical investigations, the treatment D is not binary, but takes value on an ordered discrete set $\{s_1, \dots, s_{k+1}\}$, $k > 1$. Li (2004) Chapter 2 shows in this case under certain conditions that an instrument Z monotone with respect to D implies an ordered latent variable model. One can always transform the instrument to let the support be $[0, 1]$. The "generalized" model (M2) in this case will be

$$\begin{cases} X, Z; \\ 1(D = s_i) = 1(Z \in [\xi_{i-1}, \xi_i)), \text{ for } i = 1, \dots, k+1; \\ Y = \sum_i 1(D = s_i)Y(s_i). \end{cases}$$

where $\xi_0 = 0$, $\xi_{k+1} = 1$, and $0 \leq \xi_1 \leq \dots \leq \xi_k \leq 1$ are k ordered random variables supported on $[0, 1]$. $Y(s_i)$ is the *ex ante* outcomes under treatment s_i . In this case, one can no longer see the instrument Z as the propensity score, but as a single index that explains the selection behavior.

Following the same bounding argument as in Section 3, one can only know the *ex post* law $P^*(Y(s_i)|D = s_i)$ and bound the *ex ante* law $P(Y(s_i))$ as

Proposition 7.1.

$$\begin{aligned} \inf_z [p_i^*(z)P^*(Y(s_i)|D(z) = s_i) + (1 - p_i^*(z))P^\infty] &\succeq P(Y(s_i)) \succeq \\ &\succeq \sup_z [p_i^*(z)P^*(Y(s_i)|D(z) = s_i) + (1 - p_i^*(z))P_{-\infty}] \end{aligned}$$

where $p_i^*(z) = P^*(D(z) = s_i|Z = z)$ is the probability that the selected treatment D falls into the s_i category when given an instrument value $Z = z$. Bounds for g -effects can be readily derived following the same line as in Section 5.

Unfortunately, except for the groups receiving the top and bottom treatment, i.e. $D = s_1$ and $D = s_{k+1}$, the probability $p_i^*(z)$ is not necessarily monotone with respect to z .¹⁶ Following an argument similar to Proposition 3.2, only for the *ex ante* laws of the top and bottom treatment groups with $i = 1$, or $k + 1$, the bounds in Proposition 7.1 simplifies to the H-V bounds using the largest probability p_i^* . But this simplification does not necessarily hold for any other groups.

¹⁶ To see this, one can use a special case assuming (ξ_1, \dots, ξ_k) has an absolutely continuous law. When $z \downarrow 0$, $p_1^*(z) \uparrow 1$ and $\forall i > 1, p_i^*(z) \downarrow 0$. On the other hand, if $z \uparrow 1$, $p_{k+1}^*(z) \uparrow 1$ and $\forall i < k + 1, p_i^*(z) \downarrow 0$. Therefore, for any $1 < i < k + 1$, $p_i^*(z)$ is not monotone.

The point identification of effects on "local" events will fail in the polychotomous treatment model M2. (For a presentation of the failure of identification in LATE context with a binary instrument, see Imbens & Rubin 1997.) As a contrast, please recall that in the 0-1 treatment case that the set $\{\xi \leq \pi_l\}$ is a subset of $\{\xi \leq \pi_u\}$ if $\pi_l < \pi_u$, such that one can recover the *ex ante* law conditional on the set $\{\pi_l < \xi \leq \pi_u\}$ as $P(Y_i | \pi_l < \xi \leq \pi_u) = P(Y_i, \pi_l < \xi \leq \pi_u) / (\pi_u - \pi_l)$, $i = 0, 1$ by differencing the *ex post* laws. But one cannot do the same in this polychotomous treatment case. Of course one may still use a graphic illustration similar to the first one in Section 3, to represent the observed and unobserved parts for the top and bottom treatment groups with $D = s_1$, or s_{k+1} . This leads to the fact that if $z_l < z_u$, the set $\{z_u < \xi_1\}$ is a subset of $\{z_l < \xi_1\}$, and that the set $\{\xi_k \leq z_l\}$ is a subset of $\{\xi_k \leq z_u\}$. Therefore using the differencing argument in Section 3, one can identify the *ex ante* law $P(Y(s_1))$ and $P(Y(s_{k+1}))$ on the event $\{z_l < \xi_1 \leq z_u\}$ and $\{z_l < \xi_k \leq z_u\}$, respectively.

In order to show how the identification fails, consider a "local" event consisting of the "complier" individuals, those choosing $D = s_{i-1}$ if $Z = z_l$ and choosing $D = s_i$ if $Z = z_u$, with $1 < i < k + 1$.^{17 18}

Following a similar classification as in Section 3, this "local" event can be represented by the set $\{\xi_{i-2} \leq z_l < \xi_{i-1} \leq z_u < \xi_i\}$. Figure 3 shows the situation of a middle group $D = s_i$ such that $1 < i < k + 1$. It is obvious that the set $\{\xi_{i-1} \leq z_l < \xi_i\}$ is not necessarily within $\{\xi_{i-1} \leq z_u < \xi_i\}$ with probability 1.¹⁹ Similarly, nor is $\{\xi_{i-2} \leq z_u < \xi_{i-1}\}$ within $\{\xi_{i-2} \leq z_l < \xi_{i-1}\}$. A Proposition 3.2-type dominance fails to hold for middle treatment groups. Because there is no dominance between sets $\{\xi_{i-1} \leq z_l < \xi_i\}$ and $\{\xi_{i-1} \leq z_u < \xi_i\}$, the differencing method cannot even yield the *ex ante* law on $\{z_l < \xi_{i-1} \leq z_u < \xi_i\}$,²⁰ nor point identify the *ex ante* laws for the compliers. Also because there is no Proposition 3.2-type dominance for middle treatment groups, one

¹⁷ One can always split all $k + 1$ treatment groups into two groups with treatments $(s_1, \dots, s_i - 1)$ and (s_i, \dots, s_{k+1}) respectively, and let Y_a and Y_b be the mixture of *ex ante* $Y(s)$ of the two group. By reducing the polychotomous treatment model to a 0-1 treatment model, one can still utilize the arguments in previous sections.

¹⁸ There may be complier individuals changing their choices with increment more than 1. But it is simplest and without losing generality to consider compliers changing their choices between two consecutive treatment s_{i-1} and s_i .

¹⁹ This will be achieved if ξ_i is supported above z_u , such as the case for the top group $i = k + 1$ with $\xi_{k+1} = 1$, or if the selection of whether $D \leq s_i$ or $D > s_i$ is totally determined by the value of the instrument, as in the case of a random trial with perfect compliance.

²⁰ Nor can it yield the *ex ante* law on $\{\xi_{i-2} \leq z_l < \xi_{i-1} \leq z_u\}$.

cannot simplify the bounds in Proposition 7.1 except for the top and bottom groups.

REMARK 7.1. A sufficient condition for the set $\{\xi_{i-1} \leq z_l < \xi_i\}$ being within $\{\xi_{i-1} \leq z_u < \xi_i\}$ is ξ_i being (almost surely) a constant above z_u . Also the set $\{\xi_{i-2} \leq z_u < \xi_{i-1}\}$ being within $\{\xi_{i-2} \leq z_l < \xi_{i-1}\}$ can be achieved by ξ_{i-2} being a constant below z_l . This means the instrument Z by itself can separate all the individuals into three groups with treatments $(1, \dots, i-2)$, $(i-1, i)$, and $(i+1, \dots, k+1)$, such as the case of a random trial with perfect compliance separating individuals into these three groups. If this is true, one can point identify the *ex ante* laws $P(Y(s_{i-1}))$ and $P(Y(s_i))$ on the set $\{\xi_{i-2} \leq z_l < \xi_{i-1} \leq z_u < \xi_i\}$ by differencing the *ex post* laws. The 0-1 treatment case I discuss in the early sections is just a special example, with all the individuals having probability 1 being in the group with treatments either 0 or 1 (let $i = 1$).

8 Summary of Identification Results

This part of the paper summarizes the previous identification studies on the average treatment effects and bounds for them in the context of sample selection models with a monotone instrument. It also shows that the true merit of a sample selection model/LATE approach is to enable a "stratification" of the sample space by both the propensity score and the unobservables affecting the self-selection. When the *ex ante* law on each stratum can be recovered from *ex post* quantities, functionals of the conditional laws on some "Local" events, including Average Treatment Effects and Quantile Treatment Effects, can be identified, depending on the support of the propensity scores. When the support of the propensity scores does not span the whole 0-1 interval, bounds for the g -effects are derived. They can be related to the Manski and Heckman-Vytlacil bounds in the literature. Also an identification-on-limits analysis is conducted for smooth functionals, such as the expectations and quantiles, to provide a range of biases for extrapolation methods when the cut-off points of the propensity score converges to 0 and 1. Finally, when the model is extended to a polychotomous discrete treatment case, it shows the failure of point identification even for local events. But one can still put bounds on such effects. This suggests that in order to point identify the effects, stronger assumptions are needed.

9 Estimation Problems Rising from the Identification Results

Consider the sample selection model M1 (equivalent to LATE under certain conditions):

$$\begin{cases} X, \Pi; \\ D = 1(\Pi \geq \xi); \\ Y = (1 - D)Y_0 + DY_1. \end{cases}$$

It is an idealistic model. In reality, although the propensity score Π depends only on (X, Z) through a deterministic function $\Pi = E(D|X, Z)$,²¹ the function usually is unknown to the researcher, and has to be estimated. Given the identification results, the key items one needs to estimate

$$P^*(Y_1|X = x, \Pi = \pi, D = 1);$$

and

$$P^*(Y_0|X = x, \Pi = \pi, D = 0).$$

For the estimation of ATEs, one needs to estimate the expectations:

$$E(Y_1|X = x, \Pi = \pi, D = 1)$$

and

$$E(Y_0|X = x, \Pi = \pi, D = 0)$$

as integrals over the P^* 's.

When the estimates \tilde{E} of these terms are known, one can estimate the LATE as:

$$\frac{1}{\pi_u - \pi_l} [\pi_u \tilde{E}(Y_1|X = x, \Pi = \pi_u, D = 1) + (1 - \pi_u) \tilde{E}(Y_0|X = x, \Pi = \pi_u, D = 0) - \pi_l \tilde{E}(Y_1|X = x, \Pi = \pi_l, D = 1) - (1 - \pi_l) \tilde{E}(Y_0|X = x, \Pi = \pi_l, D = 0)];$$

for two propensity score values $\pi_l < \pi_u$.

Also one can estimate the LIV/MTE effect as

$$\frac{\partial [\pi \tilde{E}(Y_1|X = x, \Pi = \pi, D = 1) + (1 - \pi) \tilde{E}(Y_0|X = x, \Pi = \pi, D = 0)]}{\partial \pi};$$

and the ATE as

$$\lim_{\pi_u \uparrow 1} \tilde{E}(Y_1|X = x, \Pi = \pi_u, D = 1) - \lim_{\pi_l \downarrow 0} \tilde{E}(Y_0|X = x, \Pi = \pi_l, D = 0);$$

²¹ To see this, please note the propensity score is the probability of being treated within the population of observationally identical individuals, not for a specific type of individual such as a "complier".

or approximate it by

$$\tilde{E}(Y_1|X = x, \Pi = \pi, D = 1) - \tilde{E}(Y_0|X = x, \Pi = 1 - \pi, D = 0) \text{ for } \pi \text{ close to } 1.$$

The approximation bias should shrink linearly when π goes to 1, when $E(Y_i|X = x, \Pi = \pi, D = i0, Y_1)$ is suitably differentiable with respect to π , and \tilde{E} terms consistently estimates E terms.

Moreover, these effects at $X = x$ can be integrated into unconditional effects over the whole support of X . (or effects conditional on $X \in a \text{ given set } A$) Specifically, because

$$P(Y_i \leq y | \pi_l < \xi \leq \pi_u) = P(Y_i \leq y, \pi_l < \xi \leq \pi_u) / P(\pi_l < \xi \leq \pi_u),$$

and the numerator on the RHS is $\int P(Y_i \leq y, \pi_l < \xi \leq \pi_u | X = x) dP(x)$, and the denominator is $\int P(\pi_l < \xi \leq \pi_u | X = x) dP(x)$, one has

$$\begin{aligned} P(Y_i \leq y | \pi_l < \xi \leq \pi_u) &= \frac{\int P(Y_i \leq y, \pi_l < \xi \leq \pi_u | X = x) dP(x)}{\int P(\pi_l < \xi \leq \pi_u | X = x) dP(x)} \\ &= \frac{\int P(Y_i \leq y | \pi_l < \xi \leq \pi_u, X = x) (\pi_u - \pi_l) dP(x)}{\int (\pi_u - \pi_l) dP(x)} \end{aligned}$$

Please note that I do not cancel out the $\pi_u - \pi_l$ term because it may not be a constant but depends on the value x . Therefore by Fubini's Theorem, the unconditional LATE will be:

$$E(Y_1 - Y_0 | \pi_l < \xi \leq \pi_u) = \frac{\int E(Y_1 - Y_0 | \pi_l < \xi \leq \pi_u, X = x) (\pi_u - \pi_l) dP(x)}{\int (\pi_u - \pi_l) dP(x)}.$$

When one evaluates it at constant π_u and π_l , the unconditional LATE simplifies to be the average of LATE at $X = x$,

$$\int E(Y_1 - Y_0 | \pi_l < \xi \leq \pi_u, X = x) dP(x)$$

When the two propensity score values are not constant, such as commonly the case where

$$\pi_u = P(D = 1 | X = x, Z = z_u) \text{ and } \pi_l = P(D = 1 | X = x, Z = z_l),$$

the unconditional LATE can be identified as

$$\begin{aligned} &\frac{1}{\int (\pi_u - \pi_l) dP(x)} \int [\pi_u E(Y_1 | X = x, Z = z_u, D = 1) + (1 - \pi_u) E(Y_0 | X = x, Z = z_u, D = 0) \\ &\quad - \pi_l E(Y_1 | X = x, Z = z_l, D = 1) - (1 - \pi_l) E(Y_0 | X = x, Z = z_l, D = 0)] dP(x). \end{aligned}$$

By the same argument, the unconditional ATE is always the average of ATE at $X = x$:

$$E(Y_1 - Y_0) = \int E(Y_1 - Y_0 | X = x) dP(x).$$

The approximation argument still holds. One would simply change E term to \tilde{E} terms, and the integrals to empirical integrals (summations) to get meaningful estimators. Thus the estimation procedure based on the estimated $\tilde{E}(Y_i|X = x, \Pi = \pi, D = i)$ is quite flexible which suits many situations different values of π_l and π_u .

There are several ways to estimate $E(Y_i|X = x, \Pi = \pi, D = i)$. The traditional control function approach use to correct the selection bias ²² assumes that the observed $E(Y_i|X = x, \Pi = \pi, D = i)$ equals the unknown $E(Y_i|X = x)$ plus control terms $\psi^i(\pi)$. Therefore one can recover the unknown $E(Y_i|X = x)$ through the estimation of $E(Y_i|X = x, \Pi = \pi, D = i)$ and $\psi^i(\pi)$. The control function approach is valid within the model M1, if co-variables X are independent to unknown selection thresholds ξ and error terms $Y_i - E(Y_i|X)$. To see this, please note

$$\begin{aligned} E(Y_1|X, \Pi = \pi, D = 1) &= E((Y_1 - E(Y_1|X)) + E(Y_1|X)|X, \Pi = \pi, \xi \leq \pi) \\ &= E(Y_1|X) + E(Y_1 - E(Y_1|X)|X, \Pi = \pi, \xi \leq \pi). \end{aligned}$$

Since (X, Π) is independent to $(Y_1 - E(Y_1|X), \xi)$, one has

$$P(Y_1 - E(Y_1|X)|X, \Pi = \pi, \xi \leq \pi) = \frac{1}{\pi} P(Y_1 - E(Y_1|X), \xi \leq \pi)$$

which only depends on the value of π . Thus the second expectation term will only depends on π too. That is,

$$E(Y_1|X, \Pi = \pi, D = 1) = E(Y_1|X) + \psi^1(\pi).$$

Similar argument holds for Y_0 . When under a joint normality assumption as in Heckman, Tobias & Vytlacil (2001, 2003), one can use a traditional Heckit control function as the inverse Mill's ration. One can use an unspecified function as a more general approach when the joint normality does not hold.

One drawback of using a control function in the model M1 as we can see in the derivation, is that it may require stronger independence condition that rules out any conditional heteroscedasticity. By accepting this restriction, one can essentially extrapolate the $E(Y_1|X, \Pi = \pi, D = 1)$ value from where we can observe at a given π to the whole spectrum of Π through the control function. This extrapolation may lead to biases of unknown magnitude if the stronger independence does not hold. Thus it is of great interest to study the estimation of $E(Y_1|X, \Pi = \pi, D = 1)$ when no additive control term is available.

²² It does not necessarily require the model M1 framework. See Heckman (1976), Imbens & Newey (2002)

In derivation of the properties of estimators, I assume that the outcomes (Y_0, Y_1) have bounded support ²³ and the data (Y, D, X, Z) of n observations (n_1 treated, and n_0 untreated observations) are generated from the population through i.i.d. sampling under model (M1).

10 Estimation of $E(Y_i|X = x, \Pi = \pi, D = i)$, $i = 0, 1$

According to the identification results, the most important term in the estimation of different types of ATEs is the term $E(Y_i|X = x, \Pi = \pi, D = i)$, for both $i = 0, 1$. ²⁴ In case that the propensity score Π is observed, one can directly estimate this term using a parametric specification, a fully nonparametric method, or a semiparametric procedure under some restriction. In most cases, Π has to be estimated first. Thus a two-stage procedure is needed. It is well known that a simple 2SLS type procedure regressing Y on estimated $\hat{\Pi}$ does not work in non-parametric situation. ²⁵ An interesting way to look at the estimation of $E(Y_i|X = x, \Pi = \pi, D = i)$ when Π is unknown, is to think that the instrument Z is a proxy of Π through an unknown deterministic propensity score function. This is different from, but can be related to the error-in-variable type models.

Here I propose a two-stage method exploiting the assumption that the propensity score is monotone with respect to the instrument. For a part of the population, this assumption can be strengthened to a strict monotonicity. Working within this part of the population, for the conditional mean for the observed Y_0 given $D = 0$, $E(Y_0|X = x, Z = z, D = 0)$, one can first obtain estimates: ²⁶

$$\hat{E}(Y_0|X = x, Z = z, D = 0)$$

and for the propensity score $\Pi(X = x, Z = z) = E(D|X = x, Z = z)$, one has

$$\hat{\Pi}(X = x, Z = z) = \hat{E}(D|X = x, Z = z).$$

²³ Thus the conditional expectation is differentiable with bounded derivatives.

²⁴ For distributional effects, this term becomes $E(1(Y_i \leq y)|X = x, \Pi = \pi, D = i)$. One can simply replace the outcome Y_i with the indicator $1(Y_i \leq y)$ to evaluate the distributional effects.

²⁵ Heckman et al (1998) discuss a two-stage procedure, first obtaining an estimate $\hat{E}(Y|X = x, \Pi = \pi)$ and then plugging in a consistent estimate $\hat{\Pi}(x, z)$ for $\Pi(x, z)$ at a given point (x, z) . This procedure cannot be used here, unless there is a verification sample providing known Π for the first stage estimation.

²⁶ This section is concentrated on the estimation for observations with $D = 0$, and one can do the same procedure for the group with $D = 1$.

Because Π is strictly monotone with respect to Z , one can invert the estimated propensity score function with respect to z to get

$$\hat{z} = \hat{\Pi}^{-1}(X = x, \hat{\Pi} = \pi)$$

Plugging it into the conditional mean, one gets an estimator

$$\tilde{E}(Y_0|X = x, \Pi = \pi, D = 0) = \hat{E}(Y_0|X = x, Z = \hat{z}, D = 0).$$

To simplify the notation, I let $E_0(x, \pi)$, $E_0(x, z)$ and $\Pi(x, z)$ stand for $E(Y_0|X = x, \Pi = \pi, D = 0)$, $E(Y_0|X = x, Z = z, D = 0)$ and $\Pi(X = x, Z = z)$ respectively. Also let $E_1(x, \pi)$ and $E_1(x, z)$ stand for $E(Y_1|X = x, \Pi = \pi, D = 1)$ and $E(Y_1|X = x, Z = z, D = 1)$ respectively. Similar notation is used for their estimators. Suppose that $\hat{E}_0(x, z)$ and $\hat{\Pi}(x, z)$ are continuously differentiable²⁷ with respect to z , let $\pi^* = \hat{\Pi}(x, \hat{z}) = \Pi(x, z^*)$, one has:

$$\hat{\Pi}(x, \hat{z}) - \hat{\Pi}(x, z^*) = \frac{\partial \hat{\Pi}(x, z^*)}{\partial z}(\hat{z} - z^*) + o(|\hat{z} - z^*|) = \Pi(x, z^*) - \hat{\Pi}(x, z^*)$$

Thus

$$\hat{z} - z^* = \frac{-\partial z}{\partial \hat{\Pi}(x, z^*)}[\hat{\Pi}(x, z^*) - \Pi(x, z^*) + o(|\hat{z} - z^*|)]$$

Also note

$$\hat{E}_0(x, \hat{z}) = \hat{E}_0(x, z^*) + \frac{\partial \hat{E}_0(x, z^*)}{\partial z}(\hat{z} - z^*) + o(|\hat{z} - z^*|).$$

Then one has

$$\begin{aligned} \tilde{E}_0(x, \pi^*) - E_0(x, \pi^*) &= \hat{E}_0(x, \hat{z}) - E_0(x, z^*) = \\ &= [\hat{E}_0(x, z^*) - E_0(x, z^*)] + \\ &+ \frac{\partial \hat{E}_0(x, z^*)}{\partial z} \frac{-\partial z}{\partial \hat{\Pi}(x, z^*)} [\hat{\Pi}(x, z^*) - \Pi(x, z^*) + o(|\hat{z} - z^*|)] + o(|\hat{z} - z^*|) \end{aligned}$$

Suppose in a neighborhood of (x, z^*) that (**Assumptions A**)

$E_0(x, z)$ is continuously differentiable with uniformly bounded $\frac{\partial E_0(x, z)}{\partial z}$;

$\Pi(x, z)$ is continuously differentiable with $0 < d_{e1} \leq \frac{\partial z}{\partial \Pi(x, z)} \leq d_{e2} < \infty$ uniformly;

$\hat{\Pi}(x, z^*) - \Pi(x, z^*) = o_p(1)$;

$\sup_{x, z} \left| \frac{\partial \hat{E}_0(x, z)}{\partial z} - \frac{\partial E_0(x, z)}{\partial z} \right| = o_p(1)$;

$\sup_{x, z} \left| \frac{\partial z}{\partial \hat{\Pi}(x, z)} - \frac{\partial z}{\partial \Pi(x, z)} \right| = o_p(1)$.

²⁷ If they are twice differentiable, all the following $o(|z - z^*|)$ terms can be written as $O((z - z^*)^2)$, which is a result I will utilize in the next section.

Then $\hat{z} - z^* = O_p(\hat{\Pi}(x, z^*) - \Pi(x, z^*)) = o_p(1)$. In this case the asymptotic behavior of $\hat{E}_0(x, \pi^*) - E_0(x, \pi^*)$ is dominated by

$$[\hat{E}_0(x, z^*) - E_0(x, z^*)] - \frac{\partial \hat{E}_0(x, z^*)}{\partial z} \frac{\partial z}{\partial \hat{\Pi}(x, z^*)} [\hat{\Pi}(x, z^*) - \Pi(x, z^*)].$$

This is a general result, which can fit into different cases.

Case 1. The simple parametric case. When $\hat{E}_0(x, z^*)$ and $\hat{\Pi}(x, z^*)$ are regular parametric estimators under i.i.d. sampling satisfying Assumptions A, ($W_{j=1}^n$ being the data) which admit asymptotically linear forms:

$$\hat{E}_0(x, z^*) - E_0(x, z^*) = \frac{1}{n_0} \sum_{j \in J_0} s_e(W_j; x, z^*) + o_p(n_0^{-1/2});$$

$$\hat{\Pi}(x, z^*) - \Pi(x, z^*) = \frac{1}{n} \sum_{j=1}^n s_p(W_j; x, z^*) + o_p(n^{-1/2}),$$

where J_0 is the index set of untreated observations which is used to estimate $E_0(x, z^*)$. n_0 is the cardinality of it.²⁸ s_e and s_p are the influence functions for the two estimators. Assume $\lim_{n \rightarrow \infty} n/n_0 = a_0 < \infty$, $a_0 > 1$, Under i.i.d. sampling, one can expect $n/n_0 - a_0 = O_p(n^{-1/2})$. Therefore, the discrepancy between $\frac{1}{n_0} \sum_{j \in J_0} s_e$ and $\frac{1}{n} \sum_{j \in J_0} a_0 s_e = \frac{1}{n} \sum_{j=1}^n a_0 s_e (1 - D_j)$ is $o_p(n^{-1/2})$. One has

Proposition 10.1. Under Assumptions A and the above asymptotically linear representation for $\hat{E}_0(x, z^*)$ and $\hat{\Pi}(x, z^*)$, one has $\hat{E}_0(x, \pi^*) - E_0(x, \pi^*) = o_p(1)$ and

$$\sqrt{n}[\hat{E}_0(x, \pi^*) - E_0(x, \pi^*)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_1) \text{ where } V_1 = EAA^T$$

$$A = a_0 s_e(W_j; x, z^*) (1 - D_j) - \frac{\partial E_0(x, \pi^*)}{\partial \pi} s_p(W_j; x, z^*);$$

PROOF: This result directly follows from the linear representation, the chain rule, Slutsky's Theorem and a CLT.

Case 2. Local smoothing estimators. Following the intuition described above requires a consistent first stage estimator, one needs to eliminate any asymptotic bias in that stage if any smoothing is involved. Also a uniformly consistent (in the neighborhood of (x, z^*)) derivative estimator is needed. A smooth higher order kernel estimator will help to achieve this. Also a product kernel will simplify the calculation of the partial

²⁸ Similarly, let J_1 be the index set of treated observations, and n_1 is its cardinality.

derivative. ²⁹ In that neighborhood, I assume that (**Assumptions B**):

(x, z) has continuous Lebesgue density f_{XZ} bounded away from zero; f_{XZ} is r -smooth, d is the dimension of (x, z) , $r > \frac{3}{2}d + 2$; ³⁰ The same is true for $f_{XZ|D=1}$ and $f_{XZ|D=0}$; $E_0(x, z)$ and $\Pi(x, z)$ are r -smooth; $\partial\Pi/\partial z$ is uniformly bounded away from 0.

$K : R \mapsto R$, the 1-dimensional kernel function, is symmetric, supported on a compact set, integrated to 1, and has moments from 1st up to $(r - 1)$ -th order being zero and bounded r -th order moment; Let K_d be the product kernel and K_{d-1} be the product kernel for $d - 1$ dimensional x ; ³¹

K is at least $d + 2$ -smooth. This implies it is bounded;

h , the bandwidth, satisfies $nh^{d+2}/\log n \rightarrow \infty$ and $nh^{d+2r} \rightarrow 0$.

Thus one can use,

$$\hat{E}_0(x, z) = \frac{\sum_{j \in J_0} K_{d-1}\left(\frac{X_j - x}{h}\right) K\left(\frac{Z_j - z}{h}\right) Y_j}{\sum_{j \in J_0} K_{d-1}\left(\frac{X_j - x}{h}\right) K\left(\frac{Z_j - z}{h}\right)};$$

$$\hat{\Pi}(x, z) = \frac{\sum_{j=1}^n K_{d-1}\left(\frac{X_j - x}{h}\right) K\left(\frac{Z_j - z}{h}\right) D_j}{\sum_{j=1}^n K_{d-1}\left(\frac{X_j - x}{h}\right) K\left(\frac{Z_j - z}{h}\right)}.$$

Under these conditions the estimators have uniformly consistent derivatives. ³² There has been an extensive literature on the estimation of the derivatives. This uniform consistency result of the derivatives here can be deemed as a special 0-order polynomial case of Theorem 4 in Heckman et al (1998). Please note the uniform strong consistency of the estimates for the 1st order derivatives requires $nh^{d+2}/\log n \rightarrow \infty$. ³³ Also condition $nh^{d+2r} \rightarrow 0$ controls the bias and guarantees the estimates are consistent and asymptotically normal. (See Pagan and Ullah p110-113) Therefore the intuition for the plug-in estimator will work, and makes $\sqrt{nh^d}[\tilde{E}_0(x, \pi^*) - E_0(x, \pi^*)]$ having a normal law. As-

²⁹ A local polynomial estimator with a higher order kernel is also applicable.

³⁰ r -smoothness is more than required here, but will be useful in the derivation of a Heckman, Ichimura & Todd (1998) type linear representation in dealing with the average of this "local" estimator.

³¹ The choice of a compact-support kernel simplifies the requirements, and can be readily extended to the case for a Heckman, Ichimura & Todd (1998) type linear representation theory.

³² When the estimates \hat{E} and $\hat{\Pi}$ are set to be consistent and continuously differentiable, one may not need the uniform consistency of the derivatives to derive the following proposition, similar to the case discussed under Assumptions A. Thus a point-wise weak convergent of the derivatives requires $nh^{d+2} \rightarrow \infty$ only. But the uniform consistency of the derivatives is needed in order to derive a linear representation as in Heckman et al (1998).

³³ The condition $nh^{d+2}/\log n \rightarrow \infty$ guarantees a point-wise strong consistence of the estimates to their mean. One can use an Euclidean covering number argument similar to Theorem 1.37 in Pollard (1984), also Lemma 6 in Heckman et al (1998) to show uniform consistency, giving the boundedness of K and its derivatives within a bounded support. The proof is omitted here.

sume $\lim_{n \rightarrow \infty} (n/n_0) = a_0 < \infty$; $a_0 > 1$, then one has:

Proposition 10.2. Under Assumptions B,

$$\sqrt{nh^d}[\tilde{E}_0(x, \pi^*) - E_0(x, \pi^*)] = \sqrt{nh^d} \left[\frac{1}{n} \sum_{j \in J_0} a_0 A_{en} - \frac{\partial E_0(x, \pi^*)}{\partial \pi} \frac{1}{n} \sum_{j=1}^n A_{pn} \right] + o_p(1)$$

where $A_{en}(W_j; x, z^*) = (h^d f_{XZ|D=0}(x, z^*))^{-1} K_{d-1}(\frac{X_j - x}{h}) K(\frac{Z_j - z^*}{h})(Y_{0j} - E_0(X_j, Z_j))$, and

$A_{pn}(W_j; x, z^*) = (h^d f_{XZ}(x, z^*))^{-1} K_{d-1}(\frac{X_j - x}{h}) K(\frac{Z_j - z^*}{h})(D_j - \Pi(X_j, Z_j))$ are the influence functions of the kernel estimators. Therefore:

$$\sqrt{nh^d}[\tilde{E}_0(x, \pi^*) - E_0(x, \pi^*)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_2) \text{ where the variance is}$$

$$V_2 = f_{XZ}(x, z^*) \int K_d^2(u) du E \left[\frac{a_0(1-D)(Y - E_0(x, z))}{f_{XZ|D=0}(x, z^*)} - \frac{\partial E_0(x, \pi^*)}{\partial \pi} \frac{(D - \pi^*)}{f_{XZ}(x, z^*)} \right]^2.$$

PROOF: The first expression follows the above intuition for the two stage plug-in procedure and a standard result for the kernel estimator, noting the consistency of the first stage and of the derivative estimates. The normal law follows a CLT and the variance is derived using a dominated convergence theorem.

One can use the same estimation procedure in the treated group of observations, to obtain $\tilde{E}_1(x, \pi)$. With estimators $\tilde{E}_1(x, \pi)$ and $\tilde{E}_0(x, \pi)$ ready, at given values of $X = x$ one can directly utilize the identification results to obtain the estimate for the LATE as

$$\frac{1}{\pi_u - \pi_l} [\pi_u \tilde{E}_1(x, \pi_u) + (1 - \pi_u) \tilde{E}_0(x, \pi_u) - \pi_l \tilde{E}_1(x, \pi_l) - (1 - \pi_l) \tilde{E}_0(x, \pi_l)];$$

for two propensity score values $\pi_l < \pi_u$.

Also one can approximate the ATE as

$$\tilde{E}_1(x, \pi) - \tilde{E}_0(x, 1 - \pi) \text{ for } \pi \text{ close to } 1.$$

REMARK 3.1 In a large sample, because $\partial \hat{\Pi} / \partial z$ converges to $\partial \Pi / \partial z$, $\hat{\Pi}$ would be invertible with probability approaching 1 ($wp \rightarrow 1$). But in a finite sample, the invertibility is not guaranteed. Thus it is valuable to study shape restricted estimation procedures.

11 Asymptotically Linear Representation of \tilde{E} Estimator

Under the smoothness condition of $r > d$, Heckman, Ichimura & Todd (1998) derive a linear representation for the local polynomial estimator $\bar{E}_0(x, z)$ for interior points within the support of co-variates ³⁴ as:

$$\bar{E}_0(x, z) - E_0(x, z) = \frac{1}{n} \sum_{j \in J_0} \phi_{ne}(W_j; x, z) + b(x, z) + R(x, z)$$

where $\phi_{ne}(W_j; x, z)$ is the influence function with zero mean; b being a uniformly $O_p(h^r)$ bias term; and the remainder term satisfies $n^{-1/2} \sum_j R(X_j, Z_j) = o_p(1)$. ³⁵

First we derive an asymptotically linear representation similar to that in Heckman-Ichimura-Todd (1998), with a bias term being $O_p(h^r)$, for the plug-in estimator \tilde{E} using first stage kernel estimates \hat{E} and $\hat{\Pi}$. Here, because derivatives are involved, in order to use Ichimura's equicontinuity lemma (Lemma 3 in Heckman et al 1998), the smoothness requirement is a bit stricter than in Heckman et al (1998) as $r > \frac{3}{2}d + 2$. Consider an alternative assumption about the bandwidth h in Assumptions B:

Assumption C

$$nh^{d+2d+4}/\log n \rightarrow \infty \text{ and } nh^{2r} \rightarrow 0. \quad ^{36}$$

Under Assumption C and other conditions in Assumptions B, one can show that point-wise strong consistency of the derivatives of the kernel estimators: ³⁷

$$\Delta^\lambda \hat{f}_{XZ}(x, z) - \Delta^\lambda f_{XZ}(x, z) \xrightarrow{as} 0;$$

$$\Delta^\lambda \hat{E}_0(x, z) - \Delta^\lambda E_0(x, z) \xrightarrow{as} 0;$$

$$\Delta^\lambda \hat{\Pi}(x, z) - \Delta^\lambda \Pi(x, z) \xrightarrow{as} 0;$$

for all the partial derivatives Δ^λ up to order $d + 2$.

³⁴ They achieve this through a proper trimming function $1((x, z) \in S)$, S being a set of interior points.

³⁵ A similar expression holds for $\bar{\Pi}$.

³⁶ This implies a strong requirement for r to be $r > \frac{3}{2}d + 2$. The order of the kernel will make the bias term to be $O_p(h^r) = o_p(n^{-1/2})$, and therefore negligible.

³⁷ Heckman et al (1998) use local polynomial estimators and prove the uniform consistency of their 1st order derivatives. A reasonable conjecture is that their λ -order derivative will be consistent if $nh^{d+2\lambda}/\log n \rightarrow \infty$. Here I use the well known result of the consistency of the derivatives for kernel estimators.

where $\hat{E}_0(x, z)$ and $\hat{\Pi}(x, z)$ are kernel estimators defined as in the previous section. $\hat{f}_{XZ}(x, z)$ is the kernel estimator of $f_{XZ}(x, z)$.

Consider points in a bounded closed set S within the support of (x, z) , applying Theorem 1.37 of Pollard (1984), one can strengthen the consistency to be uniform over S for λ up to $d + 2$:³⁸

$$\begin{aligned} \sup_{(x,z) \in S} |\Delta^\lambda \hat{f}_{XZ}(x, z) - \Delta^\lambda f_{XZ}(x, z)| &\xrightarrow{as} 0; \\ \sup_{(x,z) \in S} |\Delta^\lambda \hat{E}_0(x, z) - \Delta^\lambda E_0(x, z)| &\xrightarrow{as} 0; \\ \sup_{(x,z) \in S} |\Delta^\lambda \hat{\Pi}(x, z) - \Delta^\lambda \Pi(x, z)| &\xrightarrow{as} 0. \end{aligned}$$

for λ up to $d + 2$.

Thus the functions $\hat{f}_{XZ}(x, z)$, $\hat{E}_0(x, z)$, $\hat{\Pi}(x, z)$, $\partial \hat{E}_0(x, z) / \partial z$ and $\partial z / \partial \hat{\Pi}(x, z)$ are at least $d + 1$ -smooth $wp \rightarrow 1$. This enables one to invoke the equicontinuity argument of Heckman et al (1998) to get:

Lemma 11.1. Supposes Assumptions B (except the bandwidth of h) and the revised bandwidth Assumption C hold for interior points (x, z) in a set $S = \{(x, z) | \Pi(x, z) \geq \pi_0, f_{XZ}(x, z) > f_0\}$. Let $w(t)$ is a smooth version of the trimming function $1(t \geq 0)$. It takes value on $[0, 1]$, and is strictly monotone and twice continuously differentiable.³⁹ π_0 is a given cut-off point for propensity scores. $f_0 > 0$ is a given cut-off value for the joint density. One has the asymptotically linear representation for:

$$\begin{aligned} \text{(a)} \quad & [\hat{E}_0(x, z) - E_0(x, z)]w[\hat{\Pi}(x, z) - \pi_0]w[\hat{f}_{XZ}(x, z) - f_0] = \\ & = \frac{1}{n_0} \sum_{j \in J_0} A_{en}(W_j; x, z)w_p w_f + \hat{b}_e + \hat{R}_e(x, z); \\ \text{(b)} \quad & [\hat{\Pi}(x, z) - \Pi(x, z)]w[\hat{\Pi}(x, z) - \pi_0]w[\hat{f}_{XZ}(x, z) - f_0] = \\ & = \frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; x, z)w_p w_f + \hat{b}_p + \hat{R}_p(x, z); \\ \text{(c)} \quad & [\tilde{E}_0(x, \pi) - E_0(x, \pi)]w[\hat{\Pi}(x, z) - \pi_0]w[\hat{f}_{XZ}(x, z) - f_0] = \end{aligned}$$

³⁸ One can see all the summation terms in $\Delta^\lambda \hat{E}_0(x, z)$ and $\Delta^\lambda \hat{\Pi}(x, z)$ as random processes indexed by the values of (x, z) . Let Ψ be the classes of all possible functions these summation terms could be. An argument similar to those in Lemma 5 of Heckman et al (1998) will show the uniform ε -covering number of Ψ depends only on the ε -covering number of the set S , which is in the polynomial order of $1/\varepsilon$. Thus one can apply Pollard's Theorem 1.37.

³⁹ A twice continuously differentiable trimming function can avoid the hassle caused by the derivative of an indicator function.

$$= \left[\frac{1}{n_0} \sum_{j \in J_0} A_{en}(W_j; x, z) - \frac{\partial E_0(x, \pi)}{\partial \pi} \frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; x, z) \right] w_p w_f + \hat{b}_{ep} + \hat{R}_{ep}(x, z).$$

where A_e and A_p are the correspondent influence functions defined as in the previous section. $w_p(x, z) = w[\Pi(x, z) - \pi_0]$ and $w_f(x, z) = w[f_{XZ}(x, z) - f_0]$. \hat{w}_p and \hat{w}_f are their estimators. When they are evaluated at (X_j, Z_j) , a subscript j is added. \hat{b} 's are uniformly $o_p(n^{-1/2})$ bias terms. For all remainder terms \hat{R} 's, $n^{-1/2} \sum_{j=1}^n \hat{R}(W_j) = o_p(1)$.

PROOF: See Appendix 4.

Also in order to approximate the ATE, one needs to compare $E(Y_1|X, \Pi, D = 1)$ with $E(Y_0|X, 1 - \Pi, D = 0)$. Therefore after obtaining $\tilde{E}_0(x, \pi)$, one needs to plug in the estimated $1 - \hat{\Pi}(x, z)$ into the π , to obtain $\tilde{E}_0(x, 1 - \hat{\Pi}(x, z))$. This plug-in estimator also admits an asymptotically linear representation.

Lemma 11.2. Let $\gamma : [0, 1] \mapsto [0, 1]$ be a twice continuously differentiable 1-1 map. Under Assumptions B and the revised bandwidth Assumption C, one has:

$$\begin{aligned} & [\tilde{E}_0(x, \gamma(\hat{\Pi}(x, z))) - E_0(x, \gamma(\Pi(x, z)))] w[\hat{\Pi}(x, z) - \pi_0] w[\hat{f}_{XZ}(x, z) - f_0] = \\ & = w_p w_f \left\{ \frac{1}{n_0} \sum_{j \in J_0} A_{en}(W_j; x, z^\gamma) - \Delta^{1(2)} E_0(x, \gamma(\pi)) \frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; x, z^\gamma) + \right. \\ & \quad \left. + \Delta^{1(2)} E_0(x, \gamma(\pi)) \frac{d\gamma(\pi)}{d\pi} \frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; x, z) \right\} + \hat{b}_{ep} + \hat{R}_{ep}(x, z). \end{aligned}$$

where $z^\gamma = \Pi^{-1}(x, \gamma(\Pi(x, z)))$, i.e. $\Pi(x, z^\gamma) = \gamma(\Pi(x, z))$. The differential operator $\Delta^{1(2)}$ means the 1st order partial differential with respect to the second argument. The bias and remainder terms are defined similarly as in Lemma 11.1. Specifically when $\gamma(t) = 1 - t$, let $\Pi(x, z^\gamma) = 1 - \Pi(x, z)$. One has:

$$\begin{aligned} & [\tilde{E}_0(x, 1 - \hat{\Pi}(x, z)) - E_0(x, 1 - \Pi(x, z))] w[\hat{\Pi}(x, z) - \pi_0] w[\hat{f}_{XZ}(x, z) - f_0] = \\ & = w_p w_f \left\{ \frac{1}{n_0} \sum_{j \in J_0} A_{en}(W_j; x, z^\gamma) - \Delta^{1(2)} E_0(x, 1 - \pi) \frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; x, z^\gamma) \right. \\ & \quad \left. - \Delta^{1(2)} E_0(x, 1 - \pi) \frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; x, z) \right\} + \hat{b}_{ep} + \hat{R}_{ep}(x, z). \end{aligned}$$

PROOF: See Appendix 5.

One can use this asymptotic linearity result to derive the distribution theory in the next section.

12 Approximation of the ATE

The identification results suggest that one can approximate the ATE at a given value $X = x$ where both a high and a low propensity score is achievable. Thus by matching observations in the treated group with high propensity score against observations in the untreated group with low propensity score, one can obtain an approximation of the ATE integrated over the matched part of the population. The approximation can be deemed as a nonparametric extension of the work of Andrews & Schafgans (1998). One can choose the observations in the treated group with propensity score higher than π_0 , a cut-off point, and untreated observations with propensity score lower than $1 - \pi_0$. Further, the approximation error will shrink linearly, when more data is available and one lets the cut-off point of the propensity score go to 1.

Let J_1 and J_0 be index sets of treated/untreated observations. An estimator to approximate ATE for the population with $\Pi \geq \pi_0$ is:

$$\frac{\sum_{k \in J_1} (Y_{1k} - E_0(X_k, 1 - \Pi_k)) 1(\Pi_k \geq \pi_0)}{\sum_{k \in J_1} 1(\Pi_k \geq \pi_0)}$$

where n_1 is the number of observations in the treated group $D = 1$. Let S be the previously defined subset of the population with $\Pi \geq \pi_0$. Let $\gamma(t) = 1 - t$ and $z^\gamma = \Pi^{-1}(x, 1 - \Pi(x, z))$, and assume (x, z^γ) is also an interior point. The proposed estimator is estimating

$$M_s = \int_S [E_1(x, \Pi(x, z)) - E_0(x, 1 - \Pi(x, z))] dP_{XZ|D=1}(x, z).$$

According to the identification results, M_s is a valid approximation of the ATE within the group of individuals having high propensity scores $\Pi \geq \pi_0$ and low propensity scores $\Pi \leq 1 - \pi_0$ under different instrument values.⁴⁰ But this estimator is not practical because it involves unknown quantities $E_0(x, \pi)$ and Π that need to be estimated. One has to plug in their estimators $\tilde{E}_0(x, \pi)$ and $\hat{\Pi}$. Also to ensure only the interior observations are used to achieve the asymptotic representation, one has to use some form of trimming. Thus a practical estimator is:

$$\hat{M} = \frac{\sum_{k \in J_1} [Y_{1k} - \tilde{E}_0(X_k, 1 - \hat{\Pi}(X_k, Z_k))] \hat{w}_{pk} \hat{w}_{fk}}{\sum_{k \in J_1} \hat{w}_{pk} \hat{w}_{fk}}$$

where $w(t)$ is the trimming function defined in the previous section. It is estimating

$$M_w = \frac{E_{D=1} \{ [E_1(x, \pi) - E_0(x, 1 - \pi)] w_p w_f \}}{E_{D=1}(w_p w_f)}$$

⁴⁰ On the other hand, the estimator $[\sum_{k \in J_0} (E(Y_1|X_k, 1 - \Pi_k, D = 1) - Y_{0k}) 1(\Pi_k \leq 1 - \pi_0)] / [\sum_{k \in J_0} 1(\Pi_k \leq 1 - \pi_0)]$ is also valid. See Remark 5.2.

which is exactly M_s if w_p takes the indicator form and if there is no trimming w_f . For this estimator, one has:

Theorem 12.1. Under Assumptions B and Assumption C about the bandwidth, let $\lim_{n \rightarrow \infty} (n/n_1) = a_1 < \infty$, $a_1 > 1$. Let us simplify the notation as $w'_{pk} = w'(\Pi(X_k, Z_k) - \pi_0)$ and $w'_{fk} = w'(f_{XZ}(X_k, Z_k) - f_0)$; $U_k = [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w]$; $\Delta_k = \Delta^{1(2)} E_0(X_k, 1 - \Pi_k)$; $w^\gamma_{pk} = w(\Pi(X_k, Z^\gamma_k) - \pi_0)$ and $w^\gamma_{fk} = w(f_{XZ}(X_k, Z^\gamma_k) - f_0)$; $\Delta^\gamma_k = \Delta^{1(2)} E_0(X_k, \Pi_k)$, where $\gamma(t) = 1 - t$ and $z^\gamma = \Pi^{-1}(x, 1 - \Pi(x, z))$. One has

$$\sqrt{n_1}(\hat{M} - M_w)E_{D=1}(w_p w_f) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V_3), \text{ where the variance}$$

$$V_3 = E(B^2 | k \in J_1) + E[(B_0 + B_1 + B_2)^2 | k \in J_1] + \frac{a_0}{a_1} E(C^2 | j \in J_0) + \frac{a_0}{a_1} E[(C_1 + C_2)^2 | j \in J_0];$$

where terms

$$B = [Y_{1k} - E_1(X_k, \Pi_k)] w_{pk} w_{fk};$$

$$B_0 = U_k w_{pk} w_{fk};$$

$$B_1 = \frac{\Pi_k - 1}{a_1 f_{XZ}(X_k, Z_k)} [\Delta^\gamma_k w^\gamma_{pk} w^\gamma_{fk} f_{XZ|D=1}(X_k, Z^\gamma_k) + \Delta_k w_{pk} w_{fk} f_{XZ|D=1}(X_k, Z_k)];$$

$$B_2 = \frac{1}{a_1} [U_k f_{XZ|D=1}(X_k, Z_k) (w'_{pk} w_{fk} \frac{1 - \Pi_k}{f_{XZ}(X_k, Z_k)} + w_{pk} w'_{fk}) - E_{D=1} U w_p w'_f f_{XZ}(x, z)];$$

$$C = [Y_{0j} - E_0(X_j, \Pi_j)] \frac{f_{XZ|D=1}(X_j, Z^\gamma_j)}{f_{XZ|D=0}(X_j, Z_j)} w_{pj}^\gamma w_{fj}^\gamma;$$

$$C_1 = \frac{\Pi_j}{a_0 f_{XZ}(X_j, Z_j)} [\Delta^\gamma_j w^\gamma_{pj} w^\gamma_{fj} f_{XZ|D=1}(X_j, Z^\gamma_j) + \Delta_j w_{pj} w_{fj} f_{XZ|D=1}(X_j, Z_j)];$$

$$C_2 = \frac{1}{a_0} [U_j f_{XZ|D=1}(X_j, Z_j) (w'_{pj} w_{fj} \frac{(-\Pi_j)}{f_{XZ}(X_j, Z_j)} + w_{pj} w'_{fj}) - E_{D=1} U w_p w'_f f_{XZ}(x, z)].$$

PROOF: See Appendix 6.

Here the term B comes from $Var(Y_1 | X, \Pi, D = 1)$, B_0 from $Var[E(Y_1 | X, \Pi) - E(Y_0 | X, 1 - \Pi) | D = 1]$, and C from $Var(Y_0 | X, 1 - \Pi, D = 0)$. Terms B_1 and C_1 come from the procedure of fitting the propensity score Π and plugging in $1 - \Pi$. If the propensity score does not affect E_0 , these terms are zero. Terms B_2 and C_2 arise because of the smooth trimming. When w' takes positive values over only a small support, these terms can be quite small.

REMARK 12.1. In deriving the above distribution theory, I assume that there exists always an interior Z^γ_k such that $\Pi(X_k, Z^\gamma_k) = 1 - \Pi(X_k, Z_k)$ for all (untrimmed) observations used in \hat{M} . When this is not possible, one can introduce a further trimming based

on $\hat{\Pi}^{-1}(x, 1 - \hat{\Pi}(x, z))$, requiring it to be larger than a certain z_0 . One can follow a similar proof as in Lemma 11.2 to show a linear representation for $\hat{\Pi}^{-1}(x, 1 - \hat{\Pi}(x, z)) - \Pi^{-1}(x, 1 - \Pi(x, z))$, and follow the same steps to show the asymptotics of the estimator with a further trimming $w[\hat{\Pi}^{-1}(x, 1 - \hat{\Pi}(x, z)) - z_0]$.

REMARK 12.2. One can also match untreated (i.e. $k \in J_0$) observations Y_{0k} against $E_1(X_k, 1 - \Pi_k)$, trimming for $\Pi_k \leq 1 - \pi_0$. The estimator will be

$$\frac{\sum_{k \in J_0} [\tilde{E}_1(X_k, 1 - \hat{\Pi}(X_k, Z_k)) - Y_{0k}] \hat{w}_{-pk} \hat{w}_{fk}}{\sum_{k \in J_0} \hat{w}_{-pk} \hat{w}_{fk}},$$

where $w_{-pk} = w[1 - \pi_0 - \Pi(X_k, Z_k)]$ and $\hat{w}_{-pk} = w[1 - \pi_0 - \hat{\Pi}(X_k, Z_k)]$. This estimator is estimating

$$\frac{E_{D=0}\{[E_1(x, 1 - \pi) - E_0(x, \pi)]w_{-p}w_f\}}{E_{D=0}(w_{-p}w_f)}.$$

One can also have an estimator pooling treated and untreated observations together:

$$\frac{\sum_{k=1}^n [\hat{Y}_{1k} - \hat{Y}_{0k}] [\hat{w}_{pk} D_k + \hat{w}_{-pk} (1 - D_k)] \hat{w}_{fk}}{\sum_{k=1}^n [\hat{w}_{pk} D_k + \hat{w}_{-pk} (1 - D_k)] \hat{w}_{fk}},$$

where

$$\hat{Y}_{1k} = \begin{cases} Y_{1k} & , \text{ if } D_k = 1; \\ \tilde{E}_1(X_k, 1 - \hat{\Pi}(X_k, Z_k)) & , \text{ if } D_k = 0. \end{cases}$$

and

$$\hat{Y}_{0k} = \begin{cases} \tilde{E}_0(X_k, 1 - \hat{\Pi}(X_k, Z_k)) & , \text{ if } D_k = 1; \\ Y_{0k} & , \text{ if } D_k = 0. \end{cases}$$

This estimator is estimating

$$\frac{E\{[E_1(x, \pi) - E_0(x, 1 - \pi)][w_p D + w_{-p}(1 - D)]w_f\}}{E\{[w_p D + w_{-p}(1 - D)]w_f\}}.$$

The asymptotic distribution theory can be derived in a similar method as that in Theorem 1.

REMARK 12.3. When the treatment takes more than two values on a discrete set, usually it is not possible to identify the conditional laws for "compliers". (See Imbens & Rubin 1997, Froelich 2002 in the LATE context; See Chapter 3 of Li (2004) for the case with an ordered latent variable model.) But even in this case, similar bounding argument as in Li (2004), Chapter 3 holds for the top and bottom treatment groups.⁴¹ one can match the observations in the top treatment group with high probability of being in that group, against the observations in the bottom treatment group with high probability of being

⁴¹ To derive the bounds for the top group, one needs to use a high probability of being in that group, instead of a high propensity score in the binary treatment case. Similarly, one uses a high probability of being in the bottom group, instead of a low propensity score, for bounds for the bottom group.

in it, to approximate the ATE of changing the treatment from the bottom to the top.

REMARK 12.4. In many studies of the LATE, the instrument is binary. Analogously, one can match the treated outcomes with untreated outcomes, but according to the instrument value instead of the propensity score. The similar matching estimator $\sum_{k \in J_1} [Y_{1k} - E(Y_0 | X_k, Z_k = 0, D_k = 0)] 1(Z_k = 1)$ can be used as an approximation for the ATE. Further in this case, one can point identify the LATE conditioning on $D(Z = 0) < D(Z = 1)$. Froelich (2002) studies a matching estimator for LATE, which matches observations with different instrument values. It is related but different from the estimator considered here which matches observations of different treatment statuses. His estimator can also be used as an approximation for the ATE.

REMARK 12.5 As long as one can prove the conjecture about the uniform consistency of derivatives of the local polynomial estimator, one can use local polynomial estimators for the first stage estimates \hat{E} and $\hat{\Pi}$. All the proofs will go through, when one changes the correspondent influence functions to the ones for local polynomial estimators derived in Heckman et al (1998).

REMARK 12.6. The variance term may be difficult to estimate in practice. Bootstrap methods may be a useful approach to assess it.

13 A Monte Carlo Evaluation of the ATE Estimator

This section presents a Monte Carlo simulation exercise to assess the performance of the proposed ATE estimator, using a model as follows:

$$\left\{ \begin{array}{l} X \sim N(0, 1); \\ Z = \ln(|e_0|); \\ \quad \text{where } e_0 \sim t_4; \\ \Pi = \Phi(0.3 + 0.1X + 1.2\arctan(Z)); \\ D = 1(\Pi \geq \xi); \\ \quad \text{where } \xi \sim U[0, 1]; \\ Y_1 = 2 + a_1 e^{(X/2)} + \varepsilon_1; \\ Y_0 = a_0 X^2/2 + \varepsilon_0; \\ \quad (e_1, \dots, e_4) \sim \text{independent } U[0, 1]; \\ \quad a_1 = (1/2.7)\Psi_3^{-1}[e_1 e_2/3 + (1 - e_1/3)\xi]; \\ \quad a_0 = (1/1.69)\Psi_2^{-1}[1 - e_1 e_3/1.5 - (1 - e_1/1.5)\xi]; \\ \quad \varepsilon_1 = \Phi^{-1}[e_1 e_4/1.5 + (1 - e_1/1.5)\xi]; \\ \quad \varepsilon_0 = \Phi^{-1}[1 - e_1 e_3/1.5 - (1 - e_1/1.5)\xi]; \\ Y = (1 - D)Y_0 + DY_1. \end{array} \right.$$

Where $N(0, 1)$ is the standard normal distribution with a c.d.f. Φ ; t_4 is the t -distribution with 4 degrees of freedom; $U[0, 1]$ is the uniform distribution on $[0, 1]$; Ψ_n is the c.d.f. of χ_n^2 , the χ^2 distribution with n degrees of freedom.

The setup is meant to have 1) a probit selection equation, (not exactly but close to a linear one); 2) Y_i related to X through random coefficients that are correlated with error terms ε and with the selection threshold ξ ; 3) error terms ε with marginal distributions similar to normal; 4) error terms ε correlated with ξ , but not following a joint normal distribution. ⁴²

I compare the estimator in the previous section with the following estimators of ATE:

1) a "wizard" estimator $\frac{1}{n} \sum_{j=1}^n [E(Y_1|X_j) - E(Y_0|X_j)]$ with known $E(Y_1|X)$ and $E(Y_0|X)$; All the variation of it is caused by the sampling variation of X .

2) another "wizard" estimator $\frac{1}{n} \sum_{j=1}^n (Y_1 - Y_0)$ with known *ex ante* Y_1 and Y_0 for the full sample (including counterfactual observations); Its variation comes from the sampling error according to the model.

3) OLS-based estimator $\frac{1}{n} \sum_{j=1}^n [\check{E}(Y_1|X_j) - \check{E}(Y_0|X_j)]$ where $\check{E}(Y_i|X) = X\check{\beta}_i$ is the OLS estimator in the observed sample $D = i$.

4) Control function estimators $\frac{1}{n} \sum_{j=1}^n [\check{\check{E}}(Y_1|X_j, \Pi = 1, D = 1) - \check{\check{E}}(Y_0|X_j, \Pi = 0, D = 0)]$ where $\check{\check{E}}(Y_1|X_j, \Pi = 1, D = 1) = X\check{\beta}_i + f_i(\pi)$ is estimated in the observed sample $D = i$ with control function f_i . I use a Heckit control function (inverse Mills ratio), a linear function in π , a quadratic function, and a cubic function as f .

The estimators are assessed in the range of X in $[-1.5, 1.5]$ because the kernel based

⁴² Therefore a Heckman control function approach might be invalid.

ATE estimator proposed in the previous section will be ill behaved at values where X is sparsely distributed. For the sake of simplicity, in the proposed kernel based estimator I use a normal kernel in its calculation, ignore w_f since I fix the range of X , and choose w_p as the indicator $1(\pi - 0.8 \geq 0)$.⁴³ This means I utilize treated observations with propensity scores above an upper cut-off of 0.80 and untreated observations with propensity scores below a lower cut-off 0.20. The number of repetitions is 1000. Each sample has 1000 observations. The results are reported in Table 1.

Table 1. Monte Carlo Simulation Results

| (Methods) | Wizard 1 | Wizard 2 | OLS | My Estimator |
|-----------------|----------|----------|-------|--------------|
| Mean bias | 0.000 | 0.000 | 0.088 | 0.050 |
| Median bias | 0.000 | 0.000 | 0.089 | 0.045 |
| Square Root MSE | 0.015 | 0.073 | 0.127 | 0.181 |
| MAD | 0.012 | 0.059 | 0.105 | 0.141 |

| (Methods) | Heckit Control | Linear Control | Quadratic Control | Cubic Control |
|-----------------|----------------|----------------|-------------------|---------------|
| Mean bias | -0.238 | -0.455 | -0.734 | -0.345 |
| Median bias | -0.239 | -0.451 | -0.717 | -0.360 |
| Square Root MSE | 0.303 | 0.496 | 0.812 | 0.626 |
| MAD | 0.253 | 0.456 | 0.736 | 0.502 |

In this simulation OLS performs surprisingly well, even with a systematic upward bias. The proposed kernel based estimator successfully reduces the bias,⁴⁴ but at the expense of larger variance because fewer observations are utilized. The control function methods essentially extrapolate the estimated control functions to where $\pi = 1$ for the treated, and $\pi = 0$ for the untreated. The sizable over-correction bias in this simulation suggests that one has to be cautious in extrapolating.

⁴³ The proof in the previous section requires a higher order kernel. In reality the choice of kernel makes little difference, and a simple kernel reduces computation costs. Also the proof requires a smooth truncation function w . But in a given finite sample, one can always use an indicator since there always is a smooth truncation giving exactly the same trimming as the indicator.

⁴⁴ This estimator is meant to be an approximation with a bias which will shrink when the upper and lower cut-offs go to 1 and 0 respectively.

14 Conclusion

This paper considers a sample selection model with a monotone instrument variable. I clarify the identification under the assumption framework of this model, showing that this model allows recovering of the potential outcome distributions (and their functionals) conditional on the selection probability, namely the propensity score, and on various "local" events depending on the propensity score. But that it fails to do so when extended to the case with a polychotomous discrete treatment. Thus, it can identify various g -effects defined as a functional g of these conditional distributions under different support conditions of the propensity score. In the estimation of the average treatment effects, (here g is the integral functional) I propose a two-stage procedure exploiting the special property of a monotone instrument, to estimate the treated/untreated *ex post* outcome expectation conditional on co-variates and the propensity score. The proposed procedure is general that one can adopt for parametric or nonparametric estimation. Second I show that using kernel estimators in both stages, the proposed estimator admits an asymptotically linear representation similar to that in Heckman et al (1998). Also I use this estimator in a matching scheme, matching the outcomes in the treated group with high propensity score against outcomes from the untreated group with low propensity score, to approximate the ATE in the population where it is near identified. Finally, a Monte Carlo simulation shows the proposed estimator performs reasonably well.

Reference

- Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, vol 113, 231-263.
- Abadie, A., J. Angrist and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, vol. 70(1), 91-117, 2002.
- Abadie, A., and G. Imbens (2001): "Simple and BiasCorrected Matching Estimators for Average Treatment Effects," Harvard working paper.
- Andrews, D. W. K., and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497-517.
- Chernozhukov, V., C. Hansen (2001): "An IV Model of Quantile Treatment Effect," MIT Working Paper.
- Das, M. (2000): "Instrumental Variables Estimation of Nonparametric Models with Discrete Endogenous Regressors," Columbia working paper.

- Doksum, K. (1974): "Empirical probability plots and statistical inference for nonlinear models in the two-sample case," *Ann. Statist.*, 2, 267-277.
- Firpo, S. (2003): "Efficient Semiparametric Estimation of Quantile Treatment Effects", January, 2003, UC-Berkeley Working Paper.
- Froelich, M. (2002): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," Univ. of St. Gallen working paper.
- Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315-331.
- Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475-492.
- Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings*, 80, 313-338.
- Heckman, J., H. Ichimura, and P. Todd (1998) "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261-294.
- Heckman, J., S. Navarro-Lozano (2003) "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models," NBER working paper.
- Heckman, J., J. Smith, and N. Clements (1997): "Making the Most out of Program Evaluations and Social Experiments Accounting for Heterogeneity in Program Impacts," *Review of Economic Studies*, 64(4), 487-535.
- Heckman, J., J. Tobias, and E. Vytlacil (2001): "Four Parameters of Interests in the Evaluation of Social Programs," *Southern Economic Journal*, 68(2), 210-223.
- Heckman, J., J. Tobias, and E. Vytlacil (2003): "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85(3), 748-755.
- Heckman, J., E. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, April 13, 1999, 96:4730-4734.
- Heckman, J., E. Vytlacil (2000): "Instrument Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," NBER Working Paper.
- Heckman, J., E. Vytlacil (2001): "Policy Relevant Treatment Effects," *American Economic Review, Papers and Proceedings*, May 2001, 91(2): 107-111.
- Hirano, K., G. Imbens, and G. Ridder (2002): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," forthcoming, *Econometrica*.
- Horowitz, J., and C. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63(2), 281-302.
- Imbens, G. W. (2003): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," NBER working paper.

- Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467-475.
- Imbens, G. W., and W. Newey (2002): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," NBER working paper.
- Imbens, G. W., and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, 64, 555-574.
- Lehmann, E. L. (1974): *Nonparametrics: statistical methods based on ranks*. Holden-Day Inc., San Francisco, Calif., With the special assistance of H. J. M. d'Abrera, Holden-Day Series in Probability and Statistics.
- Li, X. (2004): "Essays on Identification and Estimation in Sample Selection Models," UNC-Chapel Hill Dissertation.
- Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319-323.
- Manski, C. F. (1995): *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, Mass.
- Manski, C. F. (1998): "Monotone Treatment Response," *Econometrica*, 65, No.5, 1311-1334.
- Manski, C. F., and J. V. Pepper (2000): "Monotone Instrument Variables: with an Application to the Returns to Schooling," *Econometrica*, 68, 997-1010.
- Pagan A., A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge, UK.
- Pollard, D. (1984) *Convergence of Stochastic Processes*. Springer-Verlag, New York, INC.
- Rosenbaum, P., D. Rubin (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(2), 135-146.
- Rubin, D., (1977): "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2(1), 1-26.
- Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, INC.
- van der Vaart, A. W., and J. A. Wellner (2000): *Weak Convergence and Empirical Processes*, Springer-Verlag, New York, INC.
- Vytlacil, E. (2000): "Semiparametric Identification of the Average Treatment Effect in Non-separable Models," Mimeo, Stanford University.
- Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result" *Econometrica*, 70, 331-341.

Appendix 1

Proof of Proposition 4.2

Let us prove Proposition 4.2 first.

$$\begin{aligned}
P(Y_1|A) &\triangleq P(Y_1|\pi_l < \xi \leq \pi_u) \\
&= \frac{P(Y_1, \pi_l < \xi \leq \pi_u)}{P(\pi_l < \xi \leq \pi_u)} \\
&= \frac{P(Y_1, \xi \leq \pi_u) - P(Y_1, \xi \leq \pi_l)}{P(\xi \leq \pi_u) - P(\xi \leq \pi_l)} \\
&= \frac{P(Y_1|\xi \leq \pi_u, \Pi = \pi_u)P(\xi \leq \pi_u|\Pi = \pi_u) - P(Y_1|\xi \leq \pi_l, \Pi = \pi_l)P(\xi \leq \pi_u|\Pi = \pi_l)}{P(\xi \leq \pi_u|\Pi = \pi_u) - P(\xi \leq \pi_l|\Pi = \pi_l)} \\
&= \frac{\pi_u P_1^*(\pi_u) - \pi_l P_1^*(\pi_l)}{\pi_u - \pi_l}.
\end{aligned}$$

The first equation is by the Bayes' rule. The second is by definition. The third is by the Bayes' rule again and the fact that $\Pi \perp (\xi, Y_1)$. The last one is by definition. Similarly,

$$\begin{aligned}
P(Y_0|A) &\triangleq P(Y_0|\pi_l < \xi \leq \pi_u) \\
&= \frac{P(Y_0, \pi_l < \xi \leq \pi_u)}{P(\pi_l < \xi \leq \pi_u)} = \frac{P(Y_0, \xi > \pi_l) - P(Y_0, \xi > \pi_u)}{P(\xi \leq \pi_u) - P(\xi \leq \pi_l)} \\
&= \frac{P(Y_0|\xi > \pi_l, \Pi = \pi_l)P(\xi > \pi_l|\Pi = \pi_l) - P(Y_0|\xi > \pi_u, \Pi = \pi_u)P(\xi > \pi_u|\Pi = \pi_u)}{P(\xi \leq \pi_u|\Pi = \pi_u) - P(\xi \leq \pi_l|\Pi = \pi_l)} \\
&= \frac{(1 - \pi_l)P_0^*(\pi_l) - (1 - \pi_u)P_0^*(\pi_u)}{\pi_u - \pi_l}.
\end{aligned}$$

The integral statements on $P(Y_i|A)$ as the average of $P(Y_i|B)$ follow from the argument (1). QED

Appendix 2

Proof of Proposition 4.1

Let $\pi_l \uparrow \pi_u$. This leads to

$$\begin{aligned}
P(Y_1|B) &= \lim_{\pi_l \uparrow \pi_u} \frac{\pi_u P_1^*(\pi_u) - \pi_l P_1^*(\pi_l)}{\pi_u - \pi_l} \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} [\pi_u P_1^*(\pi_u) - (\pi_u - \delta) P_1^*(\pi_u - \delta)] \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \{ \pi_u P_1^*(\pi_u) - (\pi_u - \delta) P_1^*(\pi_u) + (\pi_u - \delta) [\delta \frac{\partial P_1^*(\pi_u)}{\partial \pi} + o(\delta)] \} \\
&= P_1^*(\pi_u) + \pi_u \frac{\partial P_1^*(\pi_u)}{\partial \pi}.
\end{aligned}$$

Actually the above argument only requires that the left derivative of $P_1^*(\pi_u)$ exists. The first term in the final line is the observed law $P^*(Y|D = 1, \Pi = \pi_u)$. The second term can be deemed as a selection bias correction term.⁴⁵ If the selection is based on the observables only, this term is 0. Similarly,

$$P(Y_0|B) = \lim_{\pi_l \uparrow \pi_u} \frac{(1 - \pi_l)P_0^*(\pi_l) - (1 - \pi_u)P_0^*(\pi_u)}{\pi_u - \pi_l}$$

⁴⁵ Vytlačil's (2000) identification is achieved at where the bias controlling terms for two treatment groups can cancel out each other.

$$\begin{aligned}
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} [(1 - \pi_u + \delta)P_0^*(\pi_u - \delta) - (1 - \pi_u)P_0^*(\pi_u)] \\
&= \lim_{\delta \downarrow 0} \frac{1}{\delta} \left\{ (1 - \pi_u + \delta)P_0^*(\pi_u) - (1 - \pi_u)P_0^*(\pi_u) + (1 - \pi_u + \delta) \left[-\delta \frac{\partial P_0^*(\pi_u)}{\partial \pi} + o(\delta) \right] \right\} \\
&= P_0^*(\pi_u) - (1 - \pi_u) \frac{\partial P_0^*(\pi_u)}{\partial \pi}.
\end{aligned}$$

Again one only needs the left derivative exists. The first term is the observed $P^*(Y|D=0, \Pi = \pi_u)$, and the second term is a selection bias correction term. QED

Assuming the law of ξ is absolutely continuous around 0 and 1, and $P_i^*(\pi)$ is continuously differentiable there, the same proof will go through by letting $\pi_u \downarrow \pi_l = 0$ and $\pi_l \uparrow \pi_u = 1$.

Appendix 3

Proof of Corollary 4.2b

Let us denote $\pi_l \triangleq P^*(D=1|Z=0)$ and $\pi_u \triangleq P^*(D=1|Z=1)$, and without losing generality assume $\pi_l < \pi_u$, *a.s.* Recall that $D(Z=1) > D(Z=0)$ means $\pi_l < \xi \leq \pi_u$. One has

$$\begin{aligned}
P(Y_1 | \pi_l < \xi \leq \pi_u) &= \frac{\pi_u P_1^*(\pi_u) - \pi_l P_1^*(\pi_l)}{\pi_u - \pi_l} \\
&= \frac{P^*(D=1|Z=1)P^*(Y_1|D=1, Z=1) - P^*(D=1|Z=0)P^*(Y_1|D=1, Z=0)}{P^*(D=1|Z=1) - P^*(D=1|Z=0)}.
\end{aligned}$$

All terms in this expression are from the observed *ex post* laws. The denominator is $P(D(Z=1) > D(Z=0))$ in Abadie, Angrist & Imbens' (2002) notation. Provide $P(Z=1) > 0$, the first term in the numerator is:

$$\begin{aligned}
P^*(D=1|Z=1)P^*(Y_1|D=1, Z=1) &= P(Y_1, D=1|Z=1) \\
&= P(Y_1|Z=1) - P(Y_1, D=0|Z=1) \\
&= P(Y_1) - \frac{P(Y_1, D=0, Z=1)}{P(Z=1)} \\
&= E \left\{ 1(Y_1 \leq y) - \frac{Z(1-D)1(Y_1 \leq y)}{P^*(Z=1)} \right\};
\end{aligned}$$

And the second term is

$$\begin{aligned}
P^*(D=1|Z=0)P^*(Y_1|D=1, Z=0) &= P(Y_1, D=1|Z=0) \\
&= \frac{P(Y_1, D=1, Z=0)}{P(Z=0)} \\
&= E \left\{ \frac{D(1-Z)1(Y_1 \leq y)}{P^*(Z=0)} \right\}.
\end{aligned}$$

Combining the results for the numerator and for the denominator, the result in the corollary follows. QED

Appendix 4

Proof of Lemma 11.1

Let us first simplify the notation as $\hat{w}_p = w[\hat{\Pi}(x, z) - \pi_0]$, and $\hat{w}_f = w[\hat{f}_{XZ}(x, z) - f_0]$.

For $[\hat{E}_0(x, z) - E_0(x, z)]\hat{w}_p\hat{w}_f$ and $[\hat{\Pi}(x, z) - \Pi(x, z)]\hat{w}_p\hat{w}_f$, one can directly apply Theorem 3 of Heckman et al (1998), with the following modifications:

a) By the boundedness of all the Y_i 's, the propensity score Π , the joint density f_{XZ} , and the Lipschitzness of w , there exists an envelope function that satisfies the Lindeberg condition,

condition ii in Ichimura's equicontinuity lemma, i.e. Lemma 3 in Heckman et al (1998);
b) w is a given Lipschitz continuous function. Functions $\hat{\Pi}$ and \hat{f}_{XZ} are at least $d + 1$ -smooth $w_p \rightarrow 1$. Thus the uniform covering number condition in Lemma 4 of Heckman et al (1998) holds;
c) Following the same argument as in Lemma 5 of Heckman et al (1998) by applying theorem 1.37 in Pollard (1984), one can show $\sup_{(x,z) \in S} |\hat{\Pi}(x,z) - \Pi(x,z)| \rightarrow 0$, and $\sup |\hat{f}_{XZ}(x,z) - f_{XZ}(x,z)| \rightarrow 0$ almost surely. Together with the Lipschitzness of w and boundedness of f_{XZ} , this fact guarantees $\hat{w}_p \hat{w}_f$ converges to $w_p w_f$ in \mathcal{L}^2 norm;

a), b) and c) ensure the application of Lemma 3 in Heckman et al. Following the same steps as those in the proof of Theorem 3 in Heckman et al (1998), the results follow by plugging the influence function of kernel estimators. Please note that the bias term will be $O_p(h^r)$, which is $o_p(n^{-1/2})$ and thus negligible.

As for $[\tilde{E}_0(x, \pi) - E_0(x, \pi)] \hat{w}_p \hat{w}_f$, Please note that the results in Li (2004), Section 3 of Chapter 4 that

$$\begin{aligned} \tilde{E}_0(x, \pi) - E_0(x, \pi) &= \hat{E}_0(x, z) - E_0(x, z) + \\ &+ \frac{\partial \hat{E}_0(x, z)}{\partial z} \frac{-\partial z}{\partial \hat{\Pi}(x, z)} [\hat{\Pi}(x, z) - \Pi(x, z) + o(|\hat{z} - z|)] + o(|\hat{z} - z|) \end{aligned}$$

where

$$\hat{z} = \hat{\Pi}^{-1}(x, \pi) = z + \frac{-\partial z}{\partial \hat{\Pi}(x, z)} [\hat{\Pi}(x, z) - \Pi(x, z) + o(|\hat{z} - z|)]$$

Also because under the condition of this lemma, $-\partial z / \partial \hat{\Pi}(x, z)$ is continuously differentiable, one can replace all $o(|\hat{z} - z|)$ terms by $O((\hat{z} - z)^2)$. Given that $-\partial z / \partial \hat{\Pi}$ is uniformly consistent, and that $\hat{\Pi}(x, z) - \Pi(x, z)$ is uniformly $O_p((nh^d)^{-1/2} + h^r)$, $\hat{z} - z$ has to be of the same order. Therefore, the $O((\hat{z} - z)^2)$ term is $O_p((nh^d)^{-1} + h^{2r})$ which is $o_p(n^{-1/2})$ and negligible. Note that under the conditions of this lemma, functions $\partial \hat{E}_0(x, z) / \partial z$ and $-\partial z / \partial \hat{\Pi}(x, z)$ are at least $d + 1$ -smooth $w_p \rightarrow 1$, and uniformly convergent to $\partial E_0(x, z) / \partial z$ and $-\partial z / \partial \Pi(x, z)$, respectively. One applies the same argument as for the cases of $[\hat{E}_0(x, z) - E_0(x, z)] \hat{w}_p \hat{w}_f$ and $[\hat{\Pi}(x, z) - \Pi(x, z)] \hat{w}_p \hat{w}_f$, and the result follows. QED

Appendix 5

Proof of Lemma 11.2

Consider $\{\tilde{E}_0[x, \gamma(\hat{\Pi}(x, z))] - E_0[x, \gamma(\Pi(x, z))]\} \hat{w}_p \hat{w}_f$. It is $\{\tilde{E}_0[x, \gamma(\hat{\Pi}(x, z))] - \tilde{E}_0[x, \gamma(\Pi(x, z))]\} \hat{w}_p \hat{w}_f + \{\tilde{E}_0[x, \gamma(\Pi(x, z))] - E_0[x, \gamma(\Pi(x, z))]\} \hat{w}_p \hat{w}_f$. The second term follows the asymptotic representation derived in Lemma 11.1, by putting in $\Pi(x, z^\gamma) = \gamma(\Pi(x, z))$. As for the first term, it is $\Delta^{1(2)} \tilde{E}_0(x, \gamma(\pi))(d\gamma/d\pi)(\hat{\Pi}(x, z) - \Pi(x, z)) + O((\hat{\Pi}(x, z) - \Pi(x, z))^2)$. By the same argument as in the proof of Lemma 11.1, $O((\hat{\Pi}(x, z) - \Pi(x, z))^2) = o_p(n^{-1/2})$ and is negligible. Also following the same argument as in the second part of the proof of Lemma 11.1, $\Delta^{1(2)} \tilde{E}_0(x, \gamma(\pi))$ is at least $d + 1$ -smooth and uniformly convergent to $\Delta^{1(2)} E_0(x, \gamma(\pi))$. Thus one can apply the same steps as in Theorem 3 of Heckman et al (1998) again. The result readily follows. QED

Appendix 6

Proof of Theorem 12.1

Let us simplify the notation as $\Pi_k = \Pi(X_k, Z_k)$, $\hat{\Pi}_k = \hat{\Pi}(X_k, Z_k)$, $f_k = f_{XZ}(X_k, Z_k)$, and $\hat{f}_k = \hat{f}_{XZ}(X_k, Z_k)$. Note J_1 and J_0 are index sets of treated/untreated observations. Using this one has

$$\sqrt{n_1}(\hat{M} - M_w) = \frac{1}{\sqrt{n_1}} \left\{ \sum_{k \in J_1} [Y_{1k} - \tilde{E}_0(X_k, 1 - \hat{\Pi}_k) - M_w] \hat{w}_{pk} \hat{w}_{fk} \right\} / \left(n_1^{-1} \sum_{k \in J_1} \hat{w}_{pk} \hat{w}_{fk} \right)$$

Consider the numerator as the sum of three terms:

$$\frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [Y_{1k} - E_1(X_k, \Pi_k)] \hat{w}_{pk} \hat{w}_{fk} + \quad (5)$$

$$+ \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w] \hat{w}_{pk} \hat{w}_{fk} + \quad (6)$$

$$+ \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [E_0(X_k, 1 - \Pi_k) - \tilde{E}_0(X_k, 1 - \hat{\Pi}_k)] \hat{w}_{pk} \hat{w}_{fk}. \quad (7)$$

a) For the term (5), following the equicontinuity Lemma 3 of Heckman et al (1998), the empirical process $\sqrt{n_1} \sum_{k \in J_1} [Y_{1k} - E_1(X_k, \Pi_k)] \psi_n(X_k, Z_k, Y_k)$ is equicontinuous, when indexed by functions ψ_n more than $d/2$ -smooth. This applies to \hat{w}_{pk} and \hat{w}_{fk} . With they converge to w_{pk} and w_{fk} uniformly, and with an envelope satisfying Lindeberg condition, one has:

$$\frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [Y_{1k} - E_1(X_k, \Pi_k)] \hat{w}_{pk} \hat{w}_{fk} = \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [Y_{1k} - E_1(X_k, \Pi_k)] w_{pk} w_{fk} + o_p(1).$$

b) For the term (6), note that $\hat{w}_{pk} \approx w_{pk} + w'(\Pi_k - \pi_0)(\hat{\Pi}_k - \Pi_k)$. The term dropped here is uniformly $o_p(n^{-1/2})$ as shown in the proof of Lemma 11.1. Note that $w'_{pk} = w'(\Pi_k - \pi_0)$ is bounded, plugging in an asymptotically linear representation for $[E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w] w'(\Pi_k - \pi_0)(\hat{\Pi}_k - \Pi_k) \hat{w}_{fk}$ similar to the one we derive in Lemma 11.1, one has:

$$\text{Term (2)} = o_p(1) + \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w] w_{pk} \hat{w}_{fk} + \quad (8)$$

$$+ \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w] w'_{pk} w_{fk} \left[\frac{1}{n} \sum_{j=1}^n A_{pn}(W_j; X_k, Z_k) \right]. \quad (9)$$

Then observe $\hat{w}_{fk} \approx w_{fk} + w'(f_k - f_0)(\hat{f}_k - f_k)$, dropping a uniformly $o_p(n^{-1/2})$ term. As one can write

$$\hat{f}_k - f_k \approx n^{-1} \sum_{j=1}^n [K_{d-1}(\frac{X_j - X_k}{h}) K(\frac{Z_j - Z_k}{h}) - E\{K_{d-1}(\frac{X_j - X_k}{h}) K(\frac{Z_j - Z_k}{h}) | X_k, Z_k\}]$$

dropping a uniformly $O(h^r)$ term. Let A_{fn} be the influence function in this expression, and plug this into the term (8), using an argument similar to that in the proof of Theorem 2 of Heckman et al (1998). Dropping an $o_p(1)$ term, this yields:

$$\text{Term (4)} \approx \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w] w'_{fk} w_{pk} \left[\frac{1}{n} \sum_{j=1}^n A_{fn}(W_j; X_k, Z_k) \right] \quad (10)$$

$$+ \frac{1}{\sqrt{n_1}} \sum_{k \in J_1} [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w] w_{pk} w_{fk} \quad (11)$$

Thus the term (6) is about the sum of terms (9), (10) and (11). Note $U_k = [E_1(X_k, \Pi_k) - E_0(X_k, 1 - \Pi_k) - M_w]$. Leaving term (11), the sum of (9) and (10) will be:

$$\begin{aligned} & \frac{1}{n\sqrt{n_1}} \sum_{k \in J_1} \sum_{j=1}^n U_k [w'_{pk} w_{fk} A_{pn}(W_j; X_k, Z_k) + w_{pk} w'_{fk} A_{fn}(W_j; X_k, Z_k)] = \\ & (\sqrt{n/n_1}) \frac{1}{n\sqrt{n}} \sum_{k=1}^n \sum_{j=1}^n D_k U_k [w'_{pk} w_{fk} A_{pn}(W_j; X_k, Z_k) + w_{pk} w'_{fk} A_{fn}(W_j; X_k, Z_k)]. \quad (12) \end{aligned}$$

The term (12) is $n\sqrt{1/n_1}$ times of an order 2 V-statistic. For a given k ,

$$E\{D_k U_k [w'_{pk} w_{fk} A_{pn}(W_j; X_k, Z_k) + w_{pk} w'_{fk} A_{fn}(W_j; X_k, Z_k)] | W_k\} = 0.$$

For a given j ,

$$\text{Var}\{D_k U_k [w'_{pk} w_{fk} A_{pn}(W_j; X_k, Z_k) + w_{pk} w'_{fk} A_{fn}(W_j; X_k, Z_k)] | W_j\} = o(n).$$

Then following Lemma 9 in Heckman et al (1998) for the order 2 U-statistic, this term is asymptotically equivalent to:

$$\begin{aligned} & (\sqrt{n/n_1}) n^{-1/2} a_1^{-1} \sum_{k=1}^n E\{U_l [w'_{pl} w_{fl} A_{pn}(W_k; X_l, Z_l) + w_{pl} w'_{fl} A_{fn}(W_k; X_l, Z_l)] | W_k, l \in J_1\} = \\ & = n_1^{-1/2} a_1^{-1} \sum_{k \in J_1} E\{U_l [w'_{pl} w_{fl} A_{pn}(W_k; X_l, Z_l) + w_{pl} w'_{fl} A_{fn}(W_k; X_l, Z_l)] | W_k, l \in J_1\} + \\ & + (\sqrt{n_0/n_1}) n_0^{-1/2} a_1^{-1} \sum_{k \in J_0} E\{U_l [w'_{pl} w_{fl} A_{pn}(W_k; X_l, Z_l) + w_{pl} w'_{fl} A_{fn}(W_k; X_l, Z_l)] | W_k, l \in J_1\}. \end{aligned}$$

To avoid confusion, the conditional expectation is written as the integral over the subscript $l \in J_1$.

c) One can apply the asymptotical representation result in Lemma 11.2 to the third term. The term (7) then is: (Note $\Delta_k = \Delta^{1(2)} E_0(X_k, 1 - \Pi_k)$)

$$n_1^{-1/2} \sum_{k \in J_1} w_{pk} w_{fk} \left\{ \frac{1}{n_0} \sum_{j \in J_0} A_{en}(W_j; X_k, Z_k^\gamma) - \Delta_k \frac{1}{n} \sum_{j=1}^n [A_{pn}(W_j; X_k, Z_k^\gamma) + A_{pn}(W_j; X_k, Z_k)] \right\}$$

$$\approx -(\sqrt{n/n_1}) n^{-3/2} \sum_{k=1}^n \sum_{j \in J_1} D_k w_{pk} w_{fk} \Delta_k [A_{pn}(W_j; X_k, Z_k^\gamma) + A_{pn}(W_j; X_k, Z_k)] + \quad (13)$$

$$+ n_1^{1/2} (n_0 n_1)^{-1} \sum_{k \in J_1} \sum_{j \in J_0} w_{pk} w_{fk} A_{en}(W_j; X_k, Z_k^\gamma). \quad (14)$$

dropping the negligible term. Here the term (13) is $n\sqrt{1/n_1}$ times an order 2 V-statistic, and the term (14) is $n_1^{1/2}$ times of an order 1 two sample U-statistic. Following Lemma 9 in Heckman et al (1998) for the order 2 U-statistic, the term (13) is asymptotically equivalent to:

$$\begin{aligned} & -(\sqrt{n/n_1}) n^{-1/2} a_1^{-1} \sum_{k=1}^n E\{w_{pl} w_{fl} \Delta_l [A_{pn}(W_k; X_l, Z_l^\gamma) + A_{pn}(W_k; X_l, Z_l)] | W_k, l \in J_1\} = \\ & = -n_1^{-1/2} a_1^{-1} \sum_{k \in J_1} E\{w_{pl} w_{fl} \Delta_l [A_{pn}(W_k; X_l, Z_l^\gamma) + A_{pn}(W_k; X_l, Z_l)] | W_k, l \in J_1\} + \\ & -(\sqrt{n_0/n_1}) n_0^{-1/2} a_1^{-1} \sum_{k \in J_0} E\{w_{pl} w_{fl} \Delta_l [A_{pn}(W_k; X_l, Z_l^\gamma) + A_{pn}(W_k; X_l, Z_l)] | W_k, l \in J_1\}. \end{aligned}$$

To avoid confusion, the conditional expectation is written as for the subscript $l \in J_1$. And following Lemma 10 in Heckman et al (1998) for a two sample U-statistic, the term (14) is equivalent to:

$$n_0^{-1/2} (a_0/a_1)^{1/2} \sum_{j \in J_0} E\{w_{pl} w_{fl} A_{en}(W_j; X_l, Z_l^\gamma) | W_j, l \in J_1\}.$$

d) Finally, the denominator is

$$n_1^{-1} \sum_{k \in J_1} \hat{w}_{pk} \hat{w}_{fk} = n_1^{-1} \sum_{k \in J_1} w_{pk} w_{fk} + o_p(1) = E_{D=1}(w_p w_f) + o_p(1).$$

Pooling a), b), c), and d) together, one has the variance of $\sqrt{n_1}(\hat{M} - M_w)E_{D=1}w_pw_f$ is determined by:

$$\begin{aligned}
& n_1^{-1/2} \sum_{k \in J_1} \{[Y_{1k} - E_1(X_k, \Pi_k)]w_{pk}w_{fk} + U_k w_{pk}w_{fk}\} + \\
& + n_1^{-1/2} a_1^{-1} \sum_{k \in J_1} E\{U_l[w'_{pl}w_{fl}A_{pn}(W_k; X_l, Z_l) + w_{pl}w'_{fl}A_{fn}(W_k; X_l, Z_l)]|W_k, l \in J_1\} + \\
& - n_1^{-1/2} a_1^{-1} \sum_{k \in J_1} E\{w_{pl}w_{fl}\Delta_l[A_{pn}(W_k; X_l, Z_l^\gamma) + A_{pn}(W_k; X_l, Z_l)]|W_k, l \in J_1\} + \\
& + n_0^{-1/2} (a_0/a_1)^{1/2} a_0^{-1} \sum_{j \in J_0} E\{U_l[w'_{pl}w_{fl}A_{pn}(W_j; X_l, Z_l) + w_{pl}w'_{fl}A_{fn}(W_j; X_l, Z_l)]|W_j, l \in J_1\} + \\
& - n_0^{-1/2} (a_0/a_1)^{1/2} a_0^{-1} \sum_{j \in J_0} E\{w_{pl}w_{fl}\Delta_l[A_{pn}(W_j; X_l, Z_l^\gamma) + A_{pn}(W_j; X_l, Z_l)]|W_j, l \in J_1\} + \\
& + n_0^{-1/2} (a_0/a_1)^{1/2} \sum_{j \in J_0} E\{w_{pl}w_{fl}A_{en}(W_j; X_l, Z_l^\gamma)|W_j, l \in J_1\}.
\end{aligned}$$

When $n \rightarrow \infty$ and $h \rightarrow 0$, the conditional expectation terms are equivalent to:

$$\begin{aligned}
& U_k[w'_{pk}w_{fk}(1 - \Pi_k) \frac{f_{XZ|D=1}(X_k, Z_k)}{f_{XZ}(X_k, Z_k)} + w_{pk}w'_{fk}f_{XZ|D=1}(X_k, Z_k)] - E_{D=1}Uw_pw'_f f_{XZ}(x, z); \\
& w_{pk}^{-\gamma^{-1}} w_{fk}^{-\gamma^{-1}} \Delta_k^{\gamma^{-1}} (1 - \Pi_k) \frac{f_{XZ|D=1}(X_k, Z_k^{\gamma^{-1}})}{f_{XZ}(X_k, Z_k)} + w_{pk}w_{fk}\Delta_k(1 - \Pi_k) \frac{f_{XZ|D=1}(X_k, Z_k)}{f_{XZ}(X_k, Z_k)}; \\
& U_j[w'_{pj}w_{fj}(-\Pi_j) \frac{f_{XZ|D=1}(X_j, Z_j)}{f_{XZ}(X_j, Z_j)} + w_{pj}w'_{fj}f_{XZ|D=1}(X_j, Z_j)] - E_{D=1}Uw_pw'_f f_{XZ}(x, z); \\
& w_{pj}^{-\gamma^{-1}} w_{fj}^{-\gamma^{-1}} \Delta_j^{\gamma^{-1}} (-\Pi_j) \frac{f_{XZ|D=1}(X_j, Z_j^{\gamma^{-1}})}{f_{XZ}(X_j, Z_j)} + w_{pj}w_{fj}\Delta_j(-\Pi_j) \frac{f_{XZ|D=1}(X_j, Z_j)}{f_{XZ}(X_j, Z_j)}; \\
& \text{and } w_{pj}^{-\gamma^{-1}} w_{fj}^{-\gamma^{-1}} (Y_{0j} - E_0(X_j, Z_j)) \frac{f_{XZ|D=1}(X_j, Z_j^{\gamma^{-1}})}{f_{XZ|D=0}(X_j, Z_j)}.
\end{aligned}$$

where $Z^{\gamma^{-1}}$ is defined as a value ζ such that $Z = \zeta^\gamma$, i.e. $\Pi(X, Z) = \gamma(\Pi(X, \zeta))$ and $\Pi(X, \zeta) = \gamma^{-1}(\Pi(X, Z))$. Note that when $\gamma(t) = 1 - t$, $\gamma^{-1} = \gamma$ and $Z^{\gamma^{-1}} = Z^\gamma$. Under this case $w_{pk}^{\gamma^{-1}} = w[\Pi(X_k, Z_k^{\gamma^{-1}}) - \pi_0] = w[1 - \Pi(X_k, Z_k) - \pi_0]$; $w_{fk}^{\gamma^{-1}} = w[f_{XZ}(X_k, Z_k^{\gamma^{-1}}) - f_0]$; and $\Delta_k^{\gamma^{-1}} = \Delta^{1(2)}E_0(X_k, 1 - \Pi(X_k, Z_k^{\gamma^{-1}})) = \Delta^{1(2)}E_0(X_k, \Pi(X_k, Z_k))$.

The result readily follows. QED