Dummy Endogenous Variables in Weakly Separable Models

Edward Vytlacil and Nese Yildiz*

April 3, 2004

Abstract

This paper considers the nonparametric identification and estimation of the average effect of a dummy endogenous variable in models where the error term is weakly but not additively separable from the regressors. The analysis includes the case of a dummy endogenous variable in a discrete choice model as a special case. This paper establishes conditions under which it is possible to identify and consistently estimate the average effect of the dummy endogenous variable without the use of large support conditions and without relying on parametric functional form or distributional assumptions. A root-N consistent and asymptotically normal estimator is developed for a special case of the model.

JEL Numbers: C50, H43
KEYWORDS: instrumental variables, sample selection models, social program evaluation

1 Introduction

This paper considers dummy endogenous variables in models where the error term is weakly but not additively separable from the regressors. The paper shows conditions for identification and estimation of the average effect of the dummy endogenous variable without imposing large support assumptions as are required by "identification-at-infinity" arguments, and without imposing parametric functional form or distributional assumptions.

An important special case of this analysis is to examine the effect of a dummy endogenous variable in a discrete choice model. For example, if a researcher wishes to examine the effect of a job training program on later employment, he or she might specify a probit equation for employment and include a dummy variable regressor for whether the individual received job training. One might expect that job training is endogenous, in particular, is correlated with

*Stanford University, Department of Economics. We would like to thank Jaap Abbring, Greg Brumfiel, Han Hong, Hide Ichimura, Jim Heckman, John Pepper, and Jim Powell for very helpful comments on this paper. This research was conducted while Edward Vytlacil was W. Glenn Campbell and Rita Ricardo-Campbell Hoover National Fellow. Correspondence: Landau Economics Building, 579 Serra Mall, Stanford CA 94305; Email: vytlacil@stanford.edu; Phone: 650-725-7836; Fax: 650-725-5702.

the error term in the employment decision rule. In the discrete choice model, the error term is not additively separable from the regressors and thus standard instrumental variable techniques are not valid even if one has a variable that is correlated with job training but not with the error term of the employment equation.[1] Following Heckman (1978), one can impose a system of equations for the joint determination of the endogenous variable (job training) and the outcome variable (later employment). Heckman (1978) imposes joint normality assumptions and develops the maximum likelihood estimator for the resulting model. The model has a form similar to a multivariate probit model, and is referred to as a "multivariate probit model with structural shift" by Heckman (1978).[2]

This raises the question of whether it is possible to identify and consistently estimate the effect of a dummy endogenous variable in weakly separable outcome equations such as discrete choice models without imposing parametric distributional assumptions. One approach is to follow the analysis of Heckman (1990a,b) to use "identification-at-infinity" arguments to identify and estimate the average effect of the dummy endogenous variable on the outcome of interest if large support conditions hold. In particular, this approach assumes that the propensity score has support equal to the full unit interval, where the propensity score is the probability of the dummy endogenous variable equaling one conditional on observed covariates.[3] The drawbacks of this method is that it requires very strong, large support conditions, and that estimation that directly follows the identification strategy involves estimation on "thin sets" and thus a slow rate of convergence.[4,5]

This paper shows that it is possible to identify and estimate the average effect of a dummy endogenous variable in a weakly separable outcome equation (a) without imposing large support conditions, and (b) without relying on parametric distributional or functional form assumptions. This result holds in a large class of weakly separable models referred to as "generalized regression" models by Han (1987), and includes both threshold crossing models as used in discrete choice analysis and transformation models such as the Box-Cox model and the proportional hazards

[1] See, e.g., the discussion in Heckman and Robb, 1985.

[2] A closely related model is the simultaneous probit model of Amemiya (1978), in which a probit model contains a continuous endogenous regressor. Later analysis of this model includes Rivers and Vuong (1988), and Newey (1986). See Blundell and Powell (2000) for analysis of a semiparametric version of this model. The assumptions and methods used by Blundell and Powell (2000) are not appropriate for the case of a dummy endogenous variable, and likewise the assumptions and methods imposed here are not appropriate for the case of a continuous endogenous variable.

[3] Heckman (1990a,b) assumed that the outcome equation is additively separable in the regressors and the error term, but his analysis extends immediately to the case without additive separability. See also Cameron and Heckman (1998) and Chen, Heckman and Vytlacil (1999) for identification-at-infinity arguments in the context of a system of discrete choice equations. Heckman and Vytlacil (1999,2001a) also further develop relevant identification-at-infinity arguments.

[4] See Andrews and Schafgans (1998), Schafgans (2000), and Schafgans and Zinde-Walsh (200) for results for the additively separable model.

[5] In the context of a non-additively separable model, Angrist (1991, 2001) proposes treating the outcome equation as a linear equation as an approximation or using instrumental variables to identify the "local average treatment effect" (LATE) as in Imbens and Angrist (1994). See Bhattacharya, et al., 1999, for a Monte Carlo analysis that also considers treating the outcome equation as a linear equation as an approximation.

model with unobserved heterogeneity. A root-N consistent and asymptotically normal estimator is developed for a special case of the model.

Other work that considers endogenous regressors in semiparametric or nonparametric models without additive separability includes Altonji and Matzkin (1998), Altonji and Ichimura (1998), Blundell and Powell (1999), and Imbens and Newey (2001).[6] Blundell and Powell (1999) and Imbens and Newey (2001) consider estimation of the average partial effect of a continuous endogenous regressor in non additively separable models, but their identification strategies are not appropriate for a discrete endogenous regressor as considered in this paper. Altonji and Ichimura (1998) consider estimation of the average derivatives of a general class of non additively separable outcome equations with tobit-type censoring of the outcome, but do not consider the effect of an endogenous binary regressor. Altonji and Matzkin (1997) allow for endogenous regressors in a panel data model with exchangeability. See Blundell and Powell (2000) for a survey of this literature.

2 Model:

For any random variable A, let a denote a realization of A, let F_A denote the distribution of A, and let Supp(A) denote the support of A. Let Y denote the outcome variable of interest and D denote the binary endogenous variable of interest. Following Heckman (1978), consider

$$\begin{aligned} Y^* &= X\beta + \alpha D + \epsilon \\ D^* &= Z\gamma + U \\ Y &= 1[Y^* \geq 0] \\ D &= 1[D^* \geq 0] \end{aligned}$$

where (X, Z) is an observed random vector, (ϵ, U) is an unobserved random vector, $1[\cdot]$ is the indicator function, $(X, Z) \perp\!\!\!\perp (\epsilon, U)$, and (ϵ, U) is normally distributed. Heckman (1978) refers to a model of this form as a multivariate probit model with a structural shift. In this model, the average effect of D on Y given covariates X is $F_\epsilon(X\beta + \alpha) - F_\epsilon(X\beta)$. Heckman (1978) develops the maximum likelihood estimator for the model.

This paper examines the more general model where one does not impose parametric distribution assumptions on the error terms, does not impose linear index assumptions, and is for a more general class of outcome equations that include the above threshold crossing model as a special

[6] Work on non additively separable models with exogenous regressors includes Matzkin (1991, 1992, 1993, 2003). There is also a large literature on identification and estimation of the slope parameters of binary choice models without parametric distributional assumptions and while relaxing the independence of the error terms and the regressors to a weaker condition such as median independence (see, e.g., Manski 1975, 1988). This literature recovers the slope parameters of the binary choice models but not the error distribution, and thus cannot answer questions related to the average effect of one of the regressors on the outcome variable.

case. In particular, we assume that Y and D are determined by:

$$Y = g(\nu(X, D), \epsilon) \quad (1)$$
$$D = 1[\vartheta(Z) - U \geq 0] \quad (2)$$

where $(X, Z) \in \Re^{K_X} \times \Re^{K_Z}$ is a random vector of other observed covariates, $(\epsilon, U) \in \Re^2$ are unobserved random variables, $g : \Re^2 \mapsto \Re$, and $\nu(\cdot, \cdot) : \Re^{K_X} \times \{0, 1\} \mapsto \Re$. We are assuming that ϵ is a scalar random variable for simplicity, the analysis can be directly extended to allow ϵ to be a random element.[7] We will assume that (X, Z) is exogenous, in particular, that $(X, Z) \perp\!\!\!\perp (\epsilon, U)$. This system of equations includes the classical case discussed above by taking $\vartheta(Z) = Z\delta$, $\nu(X, D) = X\beta + \alpha D$, $g(t, \epsilon) = 1[t + \epsilon \geq 0]$, and (ϵ, U) distributed joint normal. In the following analysis, the functions ν and g need not be known and no parametric distributional assumption will be imposed on (ϵ, U).

The form of the outcome equation for Y is referred to as a generalized regression model by Han (1987), who considered the estimation of $\nu(\cdot)$ when $\nu(\cdot)$ is known up to a finite dimensional parameter vector and all regressors are exogenous.[8] This form of the outcome equation for Y imposes that (X, D) is weakly separable from ϵ. This weak separability restriction will be critical to the following analysis, and makes the model more restrictive than the Roy-model/switching regression framework considered in Heckman (1990a,b). The purpose of this paper is to exploit this weak separability condition to by-pass the identification-at-infinity arguments for identification and estimation which are required for nonparametric switching regression models.[9] However, the results in this paper will directly extend to the switching regression model of $Y = g(\nu(X, D), \epsilon_D)$ with $\epsilon_D = D\epsilon_1 + (1 - D)\epsilon_0$, if one restricts ϵ_1 and ϵ_0 to have the same distribution conditional on U.

The model for D is a threshold-crossing model.[10] Here, $\vartheta(Z) - U$ is interpreted as net utility to the agent from choosing $D = 1$. Without loss of generality, assume that $U \sim \text{Unif}[0, 1]$ and $\vartheta(z) = P(z)$, where $P(z) = \Pr(D = 1|Z = z)$. $P(Z)$ is sometimes called the "propensity score", following Rosenbaum and Rubin (1983).

We will maintain the following assumptions:

(A-1) The distribution of U is absolutely continuous with respect to Lebesgue measure;

[7] See Altonji and Ichimura (1998) for related analysis that allows the error term to be a random element. We would like to thank Hide Ichimura for suggesting this point to us.

[8] See also Matzkin (1991, 2003), who considers estimation of $\nu(\cdot)$ when curvature restrictions but no parametric assumptions are imposed on $\nu(\cdot)$, and again all regressors are exogenous. Note that this paper differs from Han (1987) and Matzkin (1991, 2003) both by allowing for the dummy endogenous variable and by defining the object of interest to be the average effect of the endogenous variable and not recovery of the ν function.

[9] Heckman and Vytlacil (2001b) prove that the large support conditions imposed in identification-at-infinity arguments are necessary and sufficient for identification of the average treatment effect in general switching regression models.

[10] There is a larger class of latent index models that will have a representation of the form $D = \mathbf{1}[\vartheta(Z) - U \geq 0]$. For example, the seemingly more general model $D = \mathbf{1}[\theta(\vartheta(Z), U) \geq 0]$ with $\theta : \Re^2 \mapsto \Re^1$ will have a representation as $D = \mathbf{1}[\vartheta(Z) - U \geq 0]$ under mild regularity conditions (Vytlacil, 2003).

(A-2) (U, ϵ) is independent of (Z, X);

(A-3) $g(\nu(X,1),\epsilon)$ and $g(\nu(X,0),\epsilon)$ have finite first moments;

(A-4) $E(g(t,\epsilon)|U=u)$ is strictly increasing in t for a.e. u;

(A-5) There exist sets $\mathcal{S}^1_{X,Z}$ and $\mathcal{S}^0_{X,Z}$ with the following properties, where $I_j = \mathbf{1}[(X,Z) \in \mathcal{S}^j_{X,Z}]$,

(A-5-a) $\Pr[I_j = 1] > 0$, $j = 0, 1$.

(A-5-b) $\Pr[0 < P(Z) < 1 | I_j = 1] = 1$

(A-5-c) $P(Z)$ is nondegenerate conditional on $(X, I_j = 1)$, $j = 0, 1$.

(A-5-d) $\text{Supp}\big[(\nu(X,1), P(Z))\big|I_1 = 1\big] \subseteq \text{Supp}\big[(\nu(X,0), P(Z))\big]$,
$\text{Supp}\big[(\nu(X,0), P(Z))\big|I_0 = 1\big] \subseteq \text{Supp}\big[(\nu(X,1), P(Z))\big]$.

Assumption (A-1) is a regularity condition imposed to guarantee smoothness of the relevant conditional expectation functions. Assumption (A-2) is a critical independence condition, that the observed covariates (besides for the binary endogenous variable of interest) are independent of the unobserved covariates. Assumption (A-3) is a standard regularity condition required to have the parameter of interest be well defined. We will strengthen (A-3) for estimation.

Assumption (A-4) is a monotonicity condition.[11] It is important to note that (A-4) does not require g to be strictly increasing in t, it does not impose any form of monotonicity of g in ϵ, nor does it impose any form of monotonicity on the ν_1, ν_0 functions. One example of a model which will satisfy (A-4) is the transformation model, where $g(t_0, \epsilon) = r(t_0 + \epsilon)$, and r is a (possibly unknown) strictly increasing function. This model is referred to as a transformation model, and includes as special cases the Box-Cox model and the proportional hazards model with unobserved heterogeneity. Since r is strictly increasing, condition (A-4) is immediately satisfied. However, (A-4) also allows for cases where g is not strictly monotonic in t. An important special case is the threshold crossing models for a binary outcome variable, where $g(t,\epsilon) = \mathbf{1}(\epsilon \leq t)$ so that $E(g(t,\epsilon)|U=u) = \Pr(\epsilon \leq t|U=u)$. If $\text{Supp}(\epsilon, U) = \Re \times [0,1]$, then condition (A-4) is immediately satisfied, even though g itself is not strictly increasing.

Let $\mathcal{X}^j = \{x : \exists \ z \ \text{ with } \ (x,z) \in \mathcal{S}^j_{X,Z}\}$, $j = 0, 1$. The analysis will be done for $x \in \mathcal{X}^j$. Condition (A-5-a) guarantees that these sets have positive probability. Condition (A-5-b) guarantees there are both treated and untreated individuals with positive probability for (almost every) realization of Z within the set. Assumption (A-5-c) requires an exclusion restriction: there exists a variable that enters the decision rule for D but does not directly determine Y. Assumption (A-5-d) is a support condition, which will be discussed at length later in this paper. As will be shown in this paper, (A-5-d) has an empirical analog and it is possible to empirically determine these sets even though they are defined in terms of the ν function.

[11] The following analysis can be trivially extended to the case where $E(g(t,\epsilon)|U=u)$ is strictly decreasing in t for a.e. u.

Our goal is to identify and consistently estimate the average effect of D on Y. Using counterfactual notation, let

$$Y_d = g(\nu(X, d), \epsilon)$$

denote the outcome that would have been observed had an individual with observable vector X and unobservable ϵ been randomly assigned d. In this case, for any measurable set $\mathcal{A} \subseteq \text{Supp}(X)$, we can define the average outcome if all individuals with observed covariates $X \in \mathcal{A}$ had been randomly assigned $d = 1$,

$$E(Y_1|X \in \mathcal{A}) = E(g(\nu(X, 1), \epsilon)|X \in \mathcal{A}),$$

and the average outcome if all individuals with observed covariates X had been randomly assigned $d = 0$,[12]

$$E(Y_0|X \in \mathcal{A}) = E(g(\nu(X, 0), \epsilon)|X \in \mathcal{A}).$$

In this notation, the average effect of $D = 1$ versus $D = 0$ is[13]

$$E(Y_1 - Y_0|X \in \mathcal{A}) = E(g(\nu(X, 1), \epsilon) - g(\nu(X, 0), \epsilon)|X \in \mathcal{A}).$$

This paper will include identification and estimation results for $E(Y_0|X \in \mathcal{A})$, $E(Y_1|X \in \mathcal{A})$, and the average effect conditional on covariates, $E(Y_1 - Y_0|X \in \mathcal{A})$.

3 Identification Analysis

In this section we assume that the distribution of (Y, D, X, Z) is known and consider identification of the average effect of the dummy endogenous variable. In particular, we will show identification conditions given that one knows the following functions over the support of (X, Z),[14]

$$\begin{aligned} \Pr[D = 1|Z = z] &= P(z) \\ E(DY|X = x, Z = z) &= P(z)E(Y_1|D = 1, X = x, Z = z) \\ E((1 - D)Y|X = x, Z = z) &= (1 - P(z))E(Y_0|D = 0, X = x, Z = z). \end{aligned} \tag{3}$$

We wish to identify the average effect of D on Y given covariates, $E(Y_1 - Y_0|X = x)$, and thus need to identify $E(Y_1|X = x)$ and $E(Y_0|X = x)$. Using equation (1) and that Z is independent of

[12]Note that, since X is exogenous, the function $\phi(x, d) \equiv E(Y_d|X = x)$ corresponds to the average structural function as defined by Blundell and Powell (1999). From assumption (A-3), we have that $E(Y_1|X \in \mathcal{A})$ and $E(Y_0|X \in \mathcal{A})$ exist and are finite for every set $\mathcal{A}$ such that $\Pr[X \in A] > 0$.

[13]From assumption (A-3), it follows that $E(Y_1 - Y_0|X \in \mathcal{A})$ exists and is finite for every measurable set $\mathcal{A}$ such that $\Pr[X \in \mathcal{A}] > 0$.

[14]Throughout the identification section, a statement that a conditional expectation is identified or known is used as a shorthand for the more correct statement that the appropriate equivalence class of conditional expectation functions is known. For example, the statement that the function $P(z) = \Pr[D = 1|Z = z]$ is known is a shorthand for the statement that the F_Z-equivalence class, $[P] := \{q \in \mathcal{L}^1 : q = P \text{ a.e. } F_Z\}$, is known. In the estimation section, smoothness conditions will be imposed which will imply that the conditional expectations are unique subject to the smoothness conditions, but no such smoothness conditions are imposed here for identification.

ϵ conditional on X, we have that Y_1, Y_0 are mean independent of Z conditional on X, $E(Y_j|X) = E(Y_j|X,Z)$, $j = 0, 1$. Thus, applying the law of iterated expectations, we have that

$$E(Y_1|X=x) = P(z)E(Y_1|D=1, X=x, Z=z) + (1-P(z))E(Y_1|D=0, X=x, Z=z),$$

$$E(Y_0|X=x) = P(z)E(Y_0|D=1, X=x, Z=z) + (1-P(z))E(Y_0|D=0, X=x, Z=z).$$

From equation (3), we identify the first term of the first equation and the second term of the second equation but we do not immediately identify the other terms. Our analysis will use the model to identify these terms.

To see how the identification analysis will proceed, note that for any version of the conditional expectations that is consistent with our model of equations (1)-(2) and assumptions (A-1)-(A-4),

$$E(Y_1|X=x, Z=z, D=1) = E(g(\nu(x,1),\epsilon)|U \leq P(z)) \tag{4}$$

$$E(Y_0|X=x, Z=z, D=0) = E(g(\nu(x,0),\epsilon)|U > P(z)), \tag{5}$$

where we have substituted in the models for D and Y and are using the independence assumption (A-2). The problem is to identify

$$E(Y_0|X=x, Z=z, D=1) = E(g(\nu(x,0),\epsilon)|U \leq P(z)), \tag{6}$$

$$E(Y_1|X=x, Z=z, D=0) = E(g(\nu(x,1),\epsilon)|U > P(z)). \tag{7}$$

The central idea for the identification analysis is that if we can find shifts in X which directly compensate for a shift in D, then we can use information from equation (4) to fill in the missing information for equation (6), and from equation (5) to fill in the missing information for equation (7). In particular, if we identify (x, x_1) and (x_0, x) pairs such that $\nu(x,0) = \nu(x_1,1)$ and $\nu(x,1) = \nu(x_0,0)$, then evaluating equation (4) at x_1 tells us the answer for evaluating equation (6) at x, and evaluating equation (5) at x_0 tells us the answer for evaluating equation (7) at x. Because of selection (D being endogenous), we cannot immediately use the conditional expectations in the data to recover such pairs. However, given our model and assumptions, we can use the variation in the conditional expectations for changes in Z to identify such pairs. Given that equations (6) and (7) are identified by this procedure, then (a version of) $E(Y_0|X=x)$, $E(Y_1|X=x)$ and thus $E(Y_1 - Y_0|X=x)$ will be identified if the appropriate support condition holds.

For the identification analysis, it will be convenient to work with expectations conditional on $P(Z)$ instead of conditional on Z. Note that, given our assumptions, we have that any version of the conditional expectations that is consistent with our model of equations (1)-(2) and assumptions (A2) and (A3) will satisfy the following index sufficiency restriction,

$$\begin{array}{rcl} E(DY|X=x, Z=z) & = & E(DY|X=x, P(Z)=P(z)), \\ E((1-D)Y|X=x, Z=z) & = & E((1-D)Y|X=x, P(Z)=P(z)). \end{array} \tag{8}$$

Define

$$h_1(p_0,p_1,x) = \frac{1}{p_1-p_0}\Big[E\big(DY\big|X=x, P(Z)=p_1\big) - E\big(DY\big|X=x, P(Z)=p_0\big)\Big]$$

$$h_0(p_0,p_1,x) = -\frac{1}{p_1-p_0}\Big[E\big((1-D)Y\big|X=x, P(Z)=p_1\big) - E\big((1-D)Y\big|X=x, P(Z)=p_0\big)\Big].$$

One can easily show that

$$h_1(p_0, p_1, x) - h_0(p_0, p_1, x) = \frac{E(Y|X = x, P(Z) = p_1) - E(Y|X = x, P(Z) = p_0)}{p_1 - p_0}.$$

This expression is the probability limit of the Wald IV estimator with $P(Z)$ as the instrument shifting from $P(Z) = p_0$ to $P(Z) = p_1$.[15] h_1 and h_0 individually have the form of the probability limit of the Wald IV estimator applied to DY and $(1-D)Y$ separately. Evaluating $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)$ with $x_0 \neq x_1$, the difference has a form similar to the Wald IV estimator but shifting X and the instrument simultaneously. We will use the h_1, h_0 functions to uncover (x_0, x_1) pairs such that $\nu(x_1, 1) = \nu(x_0, 0)$.

Let $\text{sgn}(t)$ denote the sign function, defined as follows:

$$\text{sgn}[t] = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0. \end{cases}$$

We then have the following Lemma.

Lemma 3.1 *Assume that (D, Y) are generated according to equations (1)-(2). Assume conditions (A-1)-(A-5). Then*

$$sgn[h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)] = sgn[\nu(x_1, 1) - \nu(x_0, 0)].$$

Proof: See Appendix A.

We thus have that $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) = 0$ implies $\nu(x_1, 1) = \nu(x_0, 0)$. In other words, if $h_1(p_0, p_1, x_1) = h_0(p_0, p_1, x_0)$, then shifting X from x_0 to x_1 directly compensates for shifting D from 0 to 1. Note that if $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) = 0$ for some (p_0, p_1) evaluation points, then $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) = 0$ for all p_0, p_1 evaluation points. Let

$$\begin{array}{rcl} h_1^{-1}h_0(x_0) & = & \{x \in \text{Supp}(X) \ : \ h_1(p_0, p_1, x) = h_0(p_0, p_1, x_0) \ \text{ for some } \ p_0, p_1\} \\ h_0^{-1}h_1(x_1) & = & \{x \in \text{Supp}(X) \ : \ h_1(p_0, p_1, x_1) = h_0(p_0, p_1, x) \ \text{ for some } \ p_0, p_1\}. \end{array} \tag{9}$$

From Lemma 3.1, we have that

$$x \in h_1^{-1}h_0(x_0) \Rightarrow \nu(x, 1) = \nu(x_0, 0)$$
$$x \in h_0^{-1}h_1(x_1) \Rightarrow \nu(x_1, 1) = \nu(x, 0).$$

There is a support condition required in order to be able to find such pairs – one needs to find enough variation in X to compensate for a shift in D. Recall that $\mathcal{X}^j = \{x \ : \ \exists \ z \text{ with } (x, z) \ \in \ \mathcal{S}^j_{X,Z}\}$, $j = 0, 1$. From assumption (A-5-d), we have that, for any $x \in \mathcal{X}^1$, there is enough variation in X to compensate for a shift from $D = 1$ to $D = 0$. Likewise, for any $x \in \mathcal{X}^0$, there is enough variation in X to compensate for a shift from $D = 0$ to $D = 1$. In particular, we have that $h_1^{-1}h_0(x_0)$ is nonempty for $x_0 \in \mathcal{X}^0$, and $h_0^{-1}h_1(x_1)$ is nonempty for $x_1 \in \mathcal{X}^1$. We have the following theorem.

[15] This is the form used by Heckman and Vytlacil (1999, 2001a) for the LATE parameter, building on Imbens and Angrist (1994).

Theorem 3.1 *Assume that (D, Y) are generated according to equations (1)-(2). Assume conditions (A-1)-(A-5). Assume that the distribution of (D, Y, X, Z) is known.*

1. *For any $\mathcal{A} \subset \mathcal{X}^0$, $E(Y_0|X \in \mathcal{A})$ is identified and given by*

$$E(Y_0|X \in \mathcal{A}) = \int \left[\int \Big(E(DY|X \in h_1^{-1}h_0(x), P = p) + E((1-D)Y|X = x, P = p) \Big) dG_{P|X}(p|x) \right] dF_{X|\mathcal{A}}(x)$$

where $F_{X|\mathcal{A}}$ is the distribution function of X conditional on $X \in \mathcal{A}$, and $G_{P|X}$ is any distribution function that is absolutely continuous with respect to the distribution of $P(Z)$ conditional on X.

2. *For any $\mathcal{A} \subset \mathcal{X}^1$, $E(Y_1|X \in \mathcal{A})$ is identified and given by*

$$E(Y_1|X \in \mathcal{A}) = \int \left[\int \Big(E(DY|X = x, P = p) + E((1-D)Y|X \in h_0^{-1}h_1(x), P = p) \Big) dG_{P|X}(p|x) \right] dF_{X|\mathcal{A}}(x)$$

where $F_{X|\mathcal{A}}$ is the distribution function of X conditional on $X \in \mathcal{A}$, and $G_{P|X}$ is any distribution function that is absolutely continuous with respect to the distribution of $P(Z)$ conditional on X.

3. *For any $\mathcal{A} \in \mathcal{X}^0 \bigcap \mathcal{X}^1$, $E(Y_1 - Y_0|X \in \mathcal{A})$ is identified and given by*

$$E(Y_1 - Y_0|X \in \mathcal{A}) = \int \left[\int \Big(E(DY|X = x, P = p) + E((1-D)Y|X \in h_0^{-1}h_1(x), P = p) - E(DY|X \in h_1^{-1}h_0(x), P = p) - E((1-D)Y|X = x, P = p) \Big) dG_{P|X}(p|x) \right] dF_{X|X\in\mathcal{A}}(x)$$

where $F_{X|\mathcal{A}}$ is the distribution function of X conditional on $X \in \mathcal{A}$, and $G_{P|X}$ is any distribution function that is absolutely continuous with respect to the distribution of $P(Z)$ conditional on X.

Proof: See Appendix A.

The requirement that $\mathcal{A} \subseteq \mathcal{X}^j$ involves two types of support conditions. One is that there is sufficient variation in $P(Z)$ conditional on X for $X \in \mathcal{A}$. This requires that there be an exclusion restriction, a variable in Z that is not contained in X. The second, less standard type of support condition is that it is possible to find variation in X that compensates for a change from $D = 0$ to $D = 1$. This support condition is likely to fail near the boundaries of the support of X, as illustrated by the following example.

Illustrative Example: To illustrate the conditions of Theorem 1, take the special case of a threshold-crossing model with linear indices. In particular, assume that the true data generating process is:

$$Y = \mathbf{1}(\epsilon \leq X\beta + \delta D),$$
$$D = \mathbf{1}(V \leq Z\gamma)$$

with (ϵ, V) independent of (X, Z), having a distribution which is absolutely continuous with respect to Lebesgue measure on $\Re^2$, and having support $\Re^2$. We can map the equation for D into the form of equation 2 by taking $U = F_V(V)$. We thus have

$$\begin{aligned} E(DY|X = x, P = p) &= \Pr(V \leq F_V^{-1}(p), \epsilon \leq x\beta + \delta), \\ E((1-D)Y|X = x, P = p) &= \Pr(V > F_V^{-1}(p), \epsilon \leq x\beta), \end{aligned}$$

and thus

$$\begin{aligned} h_1(p_0, p_1, x) &= \Pr(F_V^{-1}(p_0) < V \leq F_V^{-1}(p_1), \epsilon \leq x\beta + \delta), \\ h_0(p_0, p_1, x) &= \Pr(F_V^{-1}(p_0) < V \leq F_V^{-1}(p_1), \epsilon \leq x\beta). \end{aligned}$$

Suppose that (X, Z) has support equal to the cross product of the support of X and the support of Z, $\text{Supp}(X, Z) = \text{Supp}(X) \times \text{Supp}(Z)$. For simplicity, suppose that the support of $X\beta$ is an interval, $\text{Supp}(X\beta) = [t_L, t_U]$. Then

$$\begin{aligned} h_1^{-1}h_0(x_0) &= \{x \in \text{Supp}(X) : (x_0 - x)\beta = \delta\} \\ h_0^{-1}h_1(x_1) &= \{x \in \text{Supp}(X) : (x - x_1)\beta = \delta\}, \end{aligned}$$

and

$$\begin{aligned} \mathcal{X}^1 &= \{x \in \text{Supp}(X) : x\beta \in [t_L - \delta, t_U - \delta]\} \\ \mathcal{X}^0 &= \{x \in \text{Supp}(X) : x\beta \in [t_L + \delta, t_U + \delta]\}. \end{aligned}$$

Thus, if $\delta \geq 0$,

$$\mathcal{X}^1 \bigcap \mathcal{X}^0 = \{x \in \text{Supp}(X) : x\beta \in [t_L + \delta, t_U - \delta]\}$$

and if $\delta \leq 0$,

$$\mathcal{X}^1 \bigcap \mathcal{X}^0 = \{x \in \text{Supp}(X) : x\beta \in [t_L - \delta, t_U + \delta]\}.$$

In this example, $E(Y_1 - Y_0|X = x)$ is identified for all $x \in \text{Supp}(X)$ if $\text{Supp}(X\beta)$ is unbounded. If the support of $X\beta$ is bounded, then $E(Y_1 - Y_0|X = x)$ is identified for some x values if $t_U - t_L > 2\delta$. It will not be identified for x values such that $x\beta$ is within δ of the limits of the support of $X\beta$.

We conclude the section by considering the testable restrictions imposed by the model. The assumption of a selection model imposes testable restrictions. Heckman and Vytlacil (2001a) consider a model which includes the model of the present paper as a special case, and derive two testable restrictions of the model.

Testable Restriction (1): Index sufficiency,

$$\begin{aligned}\Pr(DY \in \mathcal{A}|X=x, Z=z) &= \Pr(DY \in \mathcal{A}|X=x, P(Z)=P(z)),\\ \Pr((1-D)Y \in \mathcal{A}|X=x, Z=z) &= \Pr((1-D)Y \in \mathcal{A}|X=x, P(Z)=P(z)).\end{aligned}$$

Testable Restriction (2): If $\Pr[Y_1 \geq y_x^1|X=x]=1$, $\Pr[Y_0 \geq y_x^0|X=x]=1$, then $E[(Y_0 - y_x^0)(1-D) \mid X, P(Z)=p]$ is decreasing in p and $E[(Y_1 - y_x^1)D \mid X, P(Z)=p]$ is increasing in p.

The model of this paper implies additional testable restrictions. Under conditions (A-1)-(A-5), we have

Testable Restriction (3):

$$\begin{aligned}&\left|\int\int \big(h_1(p_0,p_1,x_1) - h_0(p_0,p_1,x_0)\big) dG(p_0|x_0,x_1) dG(p_1|x_0,x_1)\right| \\ &\qquad = \int\int \Big(\Big|h_1(p_0,p_1,x_1) - h_0(p_0,p_1,x_0)\Big|\Big) dG(p_0|x_0,x_1) dG(p_1|x_0,x_1)\end{aligned}$$

where $G(\cdot|x_0,x_1)$ is any distribution function that is absolutely continuous with respect to both the distribution of $P(Z)$ conditional on $X=x_1$ and the distribution of $P(Z)$ conditional on $X=x_0$.

Testable Restriction (4): Define $\mathcal{U}(x), \mathcal{L}(x), B^U(x), B^L(x)$ as in the statement of Theorem **??**. Let $\mathcal{A}$ denote the set of x values such that both $\mathcal{U}(x)$ and $\mathcal{L}(x)$ are nonempty. Then

$$\inf_{x\in\mathcal{A}} |B^U(x) - B^L(x)| \geq 0.$$

Testable Restriction (3) follows directly from Lemma 3.1, while Testable Restriction (4) follows directly from Theorem **??**.

4 Estimation

For simplicity, the estimation analysis will proceed under the assumption that Z contains a continuous element not contained in X. Recall that the identification analysis of the previous section does not require this assumption, and note that the following estimation strategy can be adapted for the case where Z contains only discrete elements. For ease of exposition, we only

consider estimation of $E(Y_0)$. However, estimation of $E(Y_1)$ is completely symmetric, which in turn implies an estimator for the average effect.

Given that Z contains a continuous element, and given smoothness conditions on $P(Z)$ and $E(Y|X, P(Z), D)$ as functions of Z, we can work with the derivative form of the h_1 and h_0 functions. In particular, let

$$\begin{aligned} h_1(x,p) &= \frac{\partial}{\partial p} E\big(DY\big|X = x, P(Z) = p\big) \\ h_0(x,p) &= -\frac{\partial}{\partial p} E\big((1-D)Y\big|X = x, P(Z) = p\big) \end{aligned}$$

and

$$q(t_1, t_2) = E(Y|D = 1, h_1(X, P(Z)) = t_1, P(Z) = t_2).$$

Define $h_1^{-1}(t_1; t_2) = \{x : h_1(x, t_2) = t_1\}$, $h_0^{-1}(t_1; t_2) = \{x : h_0(x, t_2) = t_1\}$. For a given t_1, t_2, $x_1 \in h_1^{-1}(t_1; t_2)$ and $x_0 \in h_0^{-1}(t_1; t_2)$ implies that $x_1 \in h_1^{-1}h_0(x_0)$ where $h_1^{-1}h_0(\cdot)$ was defined in equation (9). From the identification analysis of the previous section, we have that

$$\begin{aligned} q(t_1, t_2) &= E(Y_1|D = 1, X \in h_1^{-1}(t_1; t_2), P(Z) = t_2) \\ &= E(Y_0|D = 1, X \in h_0^{-1}(t_1; t_2), P(Z) = t_2). \end{aligned}$$

Assume that the support of $(h_1(X, P(Z)), P(Z))$ contains the support of $(h_0(X, P(Z)), P(Z))$ so that we can evaluate $q(t_1, t_2)$ at all (t_1, t_2) evaluation points in the support of $(h_0(X, P(Z)), P(Z))$. Let $P_i = P(Z_i)$, $h_{ji} = h_j(X_i, P_i)$, and assume that $\{(X_i, Z_i, D_i, Y_i) : i = 1, ..., N\}$ is an i.i.d sample. The identification analysis then suggests the following infeasible estimator of $E(Y_0)$,

$$\hat{\Delta} = \frac{1}{N} \sum_i \Big[(1 - D_i)Y_i + D_i q(h_{0i}, P_i)\Big].$$

Theorem 4.1 *Assume conditions (A-1)-(A-5). Assume that $\{X_i, Z_i, D_i, Y_i : i = 1, ..., N\}$ is i.i.d, that Y_0 has a strictly positive, finite second moment, and that the support of $(h_1(X, P(Z)), P(Z))$ contains the support of $(h_0(X, P(Z)), P(Z))$. Then*

$$\sqrt{N}\left(\frac{\hat{\Delta} - \Delta}{\sqrt{V}}\right) \xrightarrow{d} N(0, 1),$$

where

$$\begin{aligned} \Delta &= E(Y_0) \\ V &= Var\left[E\left(Y_0\middle|X, P, D\right)\right] + E\left[(1 - P)\, Var\left(Y_0\middle|X, P, D = 0\right)\right]. \end{aligned}$$

Proof: From Theorem 3.1, we have that $q(h_{0i}, P_i) = E(Y_0|D = 1, X = X_i, P(Z) = P_i)$. The theorem then follows from applying the Central Limit Theorem for i.i.d. data with a strictly positive, finite second moment.

The estimator has the form of an imputation based estimator, with the value of Y_0 for those with $D = 1$ being imputed. The form is reminiscent of a matching estimator (see, e.g., Heckman, Ichimura, and Todd, 1998, and Hahn,1998). However, the underlying assumptions of the matching estimator is different from those assumptions imposed here, and the form of the imputation is quite different as a result. If $D_i = 1$, then the matching estimator uses $E(Y_0|D = 0, X = X_i)$ to impute Y_{0i}. The missing Y_{0i} information for $D_i = 1$ observations is filled in using $Y_{0i'}$ data from $D_{i'} = 0$ observations that have (approximately) the same value of X. In contrast, the estimator proposed here fills in the missing Y_{0i} information for $D_i = 1$ observations using $Y_{1i'}$ information from $D_{i'} = 1$ observations that have different values of X, with the different value of X chosen in a way to compensate for the effect of D. These very different imputation procedures are driven by the difference in the underlying assumptions.

The above estimator would be feasible if the functions $P(\cdot)$, $h_1(\cdot,\cdot)$, $h_0(\cdot,\cdot)$, and $E(Y|D = 1, h_1(X, P(Z)) = \cdot, P(Z) = \cdot)$ were known. They are not known, which suggests using a two step semiparametric estimator where these unknown functions are replaced by consistent, nonparametric estimates. In addition, trimming is needed in practice for two reasons. First, to get uniformly consistent estimates for P, h_0 and h_1 functions, we have to trim out those observations of (X_i, Z_i) for which the value of the density $f_{X,Z}$ is low. Second, we have assumed thus far that the support of $(h_1(X, P(Z)), P(Z))$ contains the support of $(h_0(X_i, P(Z_i)), P(Z_i))$, but this is not a realistic assumption. Thus, we need to trim out those observations for which $f_{h_1,P}$ evaluated at $(h_0(X_i, P(Z_i)), P(Z_i))$ is low. Let the two trimming functions be denoted by $I_{1i} = 1\{f_{X,Z}(X_i, Z_i) \geq q_{01}\}$ and $I_{2i} = 1\{f_{h_1,P}(h_{0i}, P_i) \geq q_{02}\}$, where $q_{01}, q_{02} > 0$. These trimming functions are not known and must be estimated because the corresponding densities are not known. Thus, consider

$$\tilde{\Delta} = \frac{\frac{1}{N}\sum_i \Big[(1 - D_i)Y_i + D_i\hat{q}(\hat{h}_{0i}, \hat{P}_i)\Big]\hat{I}_{1i}\hat{I}_{2i}}{\frac{1}{N}\sum_i \hat{I}_{1i}\hat{I}_{2i}},$$

where

$$\begin{aligned}
P(z) &= E(D|Z = z) \\
\hat{P}(z) &= \hat{E}(D|Z = z) \\
h_0(x, P(z)) &= \frac{\partial}{\partial P}E\big[-(1 - D)Y|X = x, P(Z) = P(z)\big] \\
\hat{h}_0(x, \hat{P}(z)) &= \frac{\partial}{\partial P}\hat{E}\big[-(1 - D)Y|X = x, \hat{P}(Z) = \hat{P}(z)\big] \\
h_1(x, P(z)) &= \frac{\partial}{\partial P}E\big[DY|X = x, P(Z) = P(z)\big] \\
\hat{h}_1(x, \hat{P}(z)) &= \frac{\partial}{\partial P}\hat{E}\big[DY|X = x, \hat{P}(Z) = \hat{P}(z)\big] \\
q(h_0(x, P(z)), P(z)) &= E(Y|D = 1, h_1(X, P(Z)) = h_0(x, P(z)), P(Z) = P(z)) \\
\hat{q}(\hat{h}_0(x, \hat{P}(z)), \hat{P}(z)) &= \hat{E}(Y|D = 1, h_1(X, P(Z)) = \hat{h}_0(x, \hat{P}(z)), \hat{P}(Z) = \hat{P}(z))
\end{aligned}$$

and $\hat{P}_i = \hat{P}(Z_i)$ with $\hat{P}(\cdot)$ a consistent nonparametric estimator of $P(\cdot)$, and so forth, and $\hat{I}_{1i} = 1\{\hat{f}_{X,Z}(X_i, Z_i) \geq q_{01}\}$, $\hat{I}_{2i} = 1\{\hat{f}_{\hat{h}_1,\hat{P}}(\hat{h}_{0i}, \hat{P}_i) \geq q_{02}\}$.

We now develop the asymptotic distribution of this estimator when local polynomial regression estimators are used in a first step for these unknown conditional expectations functions. We need to impose some regularity conditions to carry out the estimation. To state these regularity conditions suppose $\{\tilde{h}_{N1}\}$, $\tilde{K}_1$, $\{\tilde{h}_{N2}\}$ and $\tilde{K}_2$ denote the bandwidth parameter sequence and kernel function used to estimate $f_{X,Z}$ and $f_{h_1,P}$, respectively. Similarly, let $\{h_{NP}\}$, $\{h_{Nh}\}$ and $\{h_{Nq}\}$ and K^P, K^h and K^q denote the bandwidth sequences and kernel functions used in estimating P, h_0, h_1 and q, respectively.[16] We will call a function p-smooth if it is $p+1$ times continuously differentiable and its $p+1^{st}$ derivative is Holder continuous with parameter $0 < a \leq 1$[17].

Assumption 4.1 *$\{D_i, Y_i, X_i, Z_i\}$ are i.i.d., (X_i, Z_i) takes values in $\mathbb{R}^{d_X} \times \mathbb{R}^{d_Z} = \mathbb{R}^d$, and $var(Y_i) < \infty$*

Assumptions related to the estimation of $f_{X,Z}$ and $f_{h_1,P}$:

Assumption 4.2

(a) *$f_{X,Z}$ and $f_{h_1,P}$ are both uniformly continuous and have uniformly continuous first derivatives.*

(b) *$f_{X,Z}$, P, h_0, h_1 and $f_{h_1,P}$ are all $\tilde{p}$-smooth with $\tilde{p} > d$.*

(c) *Let $q_{01} > 0$ and $q_{02} > 0$ be such that*

- *(i)* *q_{01} has a neighborhood U such that $f_{X,Z}(X,Z)$ has a continuous Lebesgue density that is strictly positive on U. Moreover for each $(x,z) \in f^{-1}_{X,Z}(U)$, $||Df_{X,Z}(x,z)|| > 0$.*
- *(ii)* *q_{02} has a neighborhood V such that $f_{h_1,P}(h_1(X,P(Z)),P(Z))$ has a continuous Lebesgue density that is strictly positive on V. Moreover for each $(x,z) \in f^{-1}_{X,Z}(U)$, $||Df_{h_1,P}(h_0(x,P(z))|| > 0$.*

(d)
- *(i)* *For each $z \in \text{supp}(Z)$ such that there exists an $x \in \text{supp}(X)$ with $(x,z) \in f^1_{X,Z}(U)$, $||DP(z)|| > 0$.*
- *(ii)* *For each $(x,z) \in f^{-1}_{X,Z}(U)$, $||D_x h_1(x,P(z))|| > 0$, and $||D_P h_1(x,P(z))|| > 0$.*

(e) *$\tilde{K}_1$ and $\tilde{K}_2$ satisfy Condition (C). $\tilde{K}_2$ is Lipschitz. Moreover, $\tilde{K}_1'$, and $\tilde{K}_2'$ satisfy parts (a), (b) and (d) of Condition (C), where*

Definition 4.1 Condition (C):

[16] We can use the same kernel function and bandwidth sequence in the estimation of h_0 and h_1.

[17] We use the same definition as in Heckman, Ichimura and Todd (1998). Namely, we say a function ϱ is Holder continuous at $X = x_0$ with constant $0 < a \leq 1$ if $|\varrho(x,t) - \varrho(x_0,t)| \leq C||x - x_0||^a$ for some $C > 0$ for all x and t in the domain of the function $\varrho(\cdot,\cdot)$. We assume that Holder continuity holds uniformly over t whenever there is an additional argument.

(a) $\tilde{K}$ *is uniformly continuous (with modulus of continuity* $w_{\tilde{K}}$*) and of bounded variation* $V(\tilde{K})$

(b) $\int |\tilde{K}(x)|\mathrm{d}x < \infty$ *and* $\tilde{K}(x) \to 0$ *as* $||x|| \to \infty$

(c) $\int \tilde{K}(x)\mathrm{d}x = 1$

(d) $\int \sqrt{||(x \log ||x||)||}|\mathrm{d}\tilde{K}(x)| < \infty$

(f) (i) $\tilde{h}_{N1} \to 0$, $\frac{\log N}{N\tilde{h}_{N1}^{d+1}} \to 0$.

(ii) $\tilde{h}_{N2} \to 0$, $\frac{\log N}{N\tilde{h}_{N2}^{3}} \to 0$, *and* $N\tilde{h}_{N2}^{12} \to c \in (0, \infty]$.

Assumptions related to the estimation of $E(D|Z)$**:**

Assumption 4.3

(a) *Bandwidth sequence* $\{h_{NP}\}$ *satisfies* $h_{NP} \to 0$, $Nh_{NP}^{d_z}/\log N \to \infty$, *and* $Nh_{NP}^{2\overline{p}_P} \to c_P \in (0, \infty)$, *where* $d_z < \overline{p}_P \leq \tilde{p}$ [18]

(b) *Kernel function* K^P *is symmetric, supported on a compact set, and is Lipschitz continuous. Also it has moments of order* $p_P + 1$ *through* $\overline{p}_P - 1$ *that are equal to 0.*

Assumptions related to the estimation of $E(DY|P(Z), X)$ **and** $E(-(1-D)Y|P(Z), X)$**:**

Assumption 4.4

(a) $\{h_{Nh}\}$ *satisfies* $Nh_{Nh}^{d_x+2}/\log N \to \infty$ *and* $Nh_{Nh}^{2(\overline{p}_h-1)} \to c_h < \infty$ *for some* $c_h \geq 0$, *and* $\overline{p}_h > d_x + 2$.

(b) *Kernel function* $K^h(\cdot)$ *is 1-smooth, symmetric and supported on a compact set. It has moments of order* $p + 1$ *through* $\overline{p}_h - 1$ *that are equal to zero.*

Assumptions related to the estimation of $E(Y|D = 1, h_1(X, P(Z)), P(Z)), P(Z))$**:**

Assumption 4.5

(a) $\{h_{Nq}\}$ *satisfies* $Nh_{Nq}^2/\log N \to \infty$ *and* $Nh_{Nq}^{2\overline{p}_q} \to c_q < \infty$ *for some* $c_q \geq 0$ *and* $\overline{p}_q > 2$.

(b) *Kernel function* $K^q(\cdot)$ *is 1-smooth, symmetric and supported on a compact set. It has moments of order* $p + 1$ *through* $\overline{p}_q - 1$ *that are equal to zero.*

(c) *There exists* $\eta \in \mathbb{R}_+$ *such that* $P\big(P(Z) > \eta\big) = 1$.

[18] In principle, we could choose $\overline{p}_P = \tilde{p}$. But to control the bias of our local polynomial estimator, certain moments of the kernel function we use must be zero, and using $\tilde{p}$ requires more moments of this function to be 0.

These assumptions impose restrictions on the kernels and on the rates at which the bandwidths go to zero for the local polynomial regressions. They also impose smoothness conditions on the unknown densities and conditional expectation functions. These types of restrictions are standard for nonparametric estimation. Note that the smoothness conditions on $f_{h_1,P}(\cdot,\cdot)$ given in Assumption 4.2(b) might appear to be stronger than are needed, but this level of smoothness is required to guarantee that the composite function $f_{h_1,P}(h_1(\cdot,P(\cdot)),P(\cdot))$ is $\tilde{p}_1$-smooth as a function of (x,z). The least standard of these regularity conditions is assumption (4.2), which restricts the behavior of the unknown densities and conditional expectation functions in a neighborhood of the trimming levels. These restrictions will be used when studying the asymptotic properties of our trimming functions. Given these assumptions, we have the following results:

(i) $\hat{P}(z)$ is asymptotically linear with trimming:

$$\left[\hat{P}(z)-P(z)\right]\hat{I}_1(x,z)=\frac{1}{N}\sum_{i=1}^{N}\psi_{NP}(D_i,X_i,Z_i;x,z)+\hat{b}_P(x,z)+\hat{R}_P(x,z)$$

where $N^{-1/2}\sum_{i=1}^{N}\hat{R}_P(X_i,Z_i)=o_p(1)$, $plim_{N\to\infty}N^{-1/2}\sum_{i=1}^{N}\hat{b}_P(X_i,Z_i)=b_P<\infty$, $E[\psi_{NP}(D_i,X_i,Z_i;X,Z)|X,Z]=0$.

(ii) $\hat{h}_0(x,\hat{P}(z))$ is asymptotically linear with trimming:

$$\begin{aligned}\left[\hat{h}_0(x,\hat{P}(z))-h_0(x,P(z))\right]\hat{I}_1(x,z) &= N^{-1}\textstyle\sum_{j=1}^{N}\left[\psi_{Nh_0}(-(1-D_j)Y_j,X_j,P(Z_j);x,z)+\frac{\partial h_0(x,P(z))}{\partial p}\psi_{NP}(D_j,X_j,Z_j;x,z)\right]\\ &+ \hat{b}_{\hat{h}_0}(x,z)+\hat{R}_{\hat{h}_0}(x,z)\end{aligned}$$

with $\text{plim}_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\hat{b}_{\hat{h}_0}(X_j,Z_j)=b_{h_0}+b_{h_0P}<\infty$, $\text{plim}_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\hat{R}_{\hat{h}_0}(X_j,Z_j)=0$, $E[\psi_{Nh_0}(-(1-D_j)Y_j,X_j,P(Z_j);X,Z)|X,Z]=0$.

(iii) $\frac{1}{\sqrt{N}}\sum_j D_j\left[\hat{q}(\hat{h}_{0j},\hat{P}_j)-q(h_{0j},P_j)\right]\hat{I}_{1j}\hat{I}_{2j}$ is asymptotically equivalent to

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}E\left[\frac{D_i}{P(Z_j)}\psi_{Nq}(Y_i,h_{1i},P_i;X,Z,P,h_0)I_2|X,Z,P,h_0\right]+b_q$$

with $E\left[\frac{D_i}{P(Z_j)}\psi_{Nq}(Y_i,h_{1i},P_i;X,Z,P,h_0)I_2|X,Z,P,h_0\right]=0$ and $b_q<\infty$.

(iv) $\sqrt{N}[\tilde{\Delta}-E(Y_0|A_1\cap A_2)]$ is asymptotically equivalent to

$$\left[\frac{1}{N}\sum_i I_{1i}I_{2i}\right]^{-1}\times\left(\frac{1}{\sqrt{N}}\sum_i E\left[D_jI_{2j}\psi(D_i,Y_i,X_i,Z_i;X_j,Z_j)\middle|Y_i,D_i,X_i,Z_i\right]+b\right.$$
$$\left.+\frac{1}{\sqrt{N}}\sum_i\left[(1-D_i)Y_i+D_iq(h_{0i},P_i)-E(Y_0|A_1\cap A_2)\right]I_{1i}I_{2i}\right)$$

where

$$A_1 = \{(x,z) \in \text{supp}(X,Z) : f_{X,Z}(x,z) \geq q_{01}\},$$
$$A_2 = \{(x,z) \in \text{supp}(X,Z) : f_{h_1,P}(h_0(x,P(z)),P(z)) \geq q_{02}\},$$
$$b = b_{qP} + b_{qh_0} + b_{qh_0P} + b_q,$$

and

$$\psi(D_i,Y_i,X_i,Z_i;X_j,Z_j) = \frac{\partial q}{\partial P}(h_{0j},P_j)\psi_{NP}(D_i,Y_i,X_i,Z_i;X_j,Z_j)$$
$$+ \frac{\partial q}{\partial h_1}(h_{0j},P_j)\psi_{Nh_0P}(D_i,Y_i,X_i,Z_i;X_j,Z_j) + \frac{1}{P(Z_j)}\psi_{Nq}(Y_i,h_{1i},P_i;X_j,Z_j,P_j,h_{0j}),$$

with

$$\psi_{Nh_0P}(D_j,Y_j,X_j,Z_j;x,z) := \psi_{Nh_0}(-(1-D_j)Y_j,P(Z_j),X_j;P(z),x,z)$$
$$+ \frac{\partial h_0(P(z),x)}{\partial p}\psi_{NP}(D_j,X_j,Z_j;x,z).$$

The key result is (iv). The result shows that estimation of the trimming function is asymptotically negligible for our estimator. An application of the central limit theorem to the result immediately implies that the estimator is $\sqrt{N}$-normal with bias b which arises due to the nonparametric estimation of the underlying conditional expectation functions. Note that $\tilde{\Delta} \xrightarrow{p} E(Y_0|A_1 \cap A_2)$, i.e., the estimator is consistent. The asymptotic distribution of $\sqrt{N}[\tilde{\Delta} - E(Y_0|A_1 \cap A_2)]$ has variance equal to the variance of the sum of two terms. The first term is $E\left[D_j I_{2j}\psi(D_i,Y_i,X_i,Z_i;X_j,Z_j)\middle|Y_i,D_i,X_i,Z_i\right]$. This term enters the asymptotic variance due to the nonparametric estimation of the unknown conditional expectation functions. The second term is $\left[(1-D_i)Y_i + D_i q(h_{0i},P_i) - E(Y_0|A_1 \cap A_2)\right] I_{1i}I_{2i}$. This second term arises from estimation noise that would be present even in the infeasible estimator if we knew the conditional expectation functions. The main argument and the analysis behind to achieve the above results are presented in Appendix B, while additional results are presented in Appendix C that is available upon request.

To highlight the major steps in the derivations, consider

$$
\begin{aligned}
\sqrt{N}(\tilde{\Delta} - E(Y_0|A_1 \cap A_2)) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\frac{\big[(1-D_i)Y_i + D_i \hat{q}(\hat{h}_{0i}, \hat{P}_i)\big] \hat{I}_{1i} \hat{I}_{2i}}{N^{-1} \sum_{i=1}^{N} \hat{I}_{1i} \hat{I}_{2i}} - E(Y_0|(X,Z) \in A_1 \cap A_2) \right] \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\big[(1-D_i)Y_i + D_i \hat{q}(\hat{h}_{0i}, \hat{P}_i) - E(Y_0|(X,Z) \in A_1 \cap A_2)\big] \hat{I}_{1i} \hat{I}_{2i}}{N^{-1} \sum_{i=1}^{N} \hat{I}_{1i} \hat{I}_{2i}} \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\big[(1-D_i)Y_i + D_i q(h_{0i}, P_i) - E(Y_0|(X,Z) \in A_1 \cap A_2)\big] \hat{I}_{1i} \hat{I}_{2i}}{N^{-1} \sum_{i=1}^{N} \hat{I}_{1i} \hat{I}_{2i}} \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{D_i \big[\hat{q}(\hat{h}_{0i}, \hat{P}_i) - q(h_{0i}, P_i)\big] \hat{I}_{1i} \hat{I}_{2i}}{N^{-1} \sum_{i=1}^{N} \hat{I}_{1i} \hat{I}_{2i}}
\end{aligned}
$$

We study the asymptotic behavior of $N^{-1} \sum_{i=1}^{N} \hat{I}_{1i} \hat{I}_{2i}$, $N^{-1/2} \sum_{i=1}^{N} \big[(1-D_i)Y_i + D_i q(h_{0i}, P_i) - E(Y_0|(X,Z) \in A_1 \cap A_2)\big] \hat{I}_{1i} \hat{I}_{2i}$, and $N^{-1/2} \sum_{i=1}^{N} D_i \big[\hat{q}(\hat{h}_{0i}, \hat{P}_i) - q(h_{0i}, P_i)\big] \hat{I}_{1i} \hat{I}_{2i}$, separately. Using our regularity conditions of Assumption 4.2, we are able to show that $N^{-1} \sum_{i=1}^{N} \hat{I}_{1i} \hat{I}_{2i} \xrightarrow{p} E(I_1 I_2)$. Thus, estimation of the trimming function does not affect the asymptotic distribution of our estimator. For the $N^{-1/2} \sum_{i=1}^{N} \big[(1-D_i)Y_i + D_i q(h_{0i}, P_i) - E(Y_0|(X,Z) \in A_1 \cap A_2)\big] \hat{I}_{1i} \hat{I}_{2i}$ term, we are able to show that the estimation of $\hat{I}_{1i} \hat{I}_{2i}$ does not effect the limiting distribution of this term, so that the term is asymptotically equivalent to $N^{-1/2} \sum_{i=1}^{N} \big[(1-D_i)Y_i + D_i q(h_{0i}, P_i) - E(Y_0|(X,Z) \in A_1 \cap A_2)\big] I_{1i} I_{2i}$. Finally, consider the $N^{-1/2} \sum_{i=1}^{N} D_i \big[\hat{q}(\hat{h}_{0i}, \hat{P}_i) - q(h_{0i}, P_i)\big] \hat{I}_{1i} \hat{I}_{2i}$ term. An application of the mean value theorem to this term reveals that its asymptotic behavior is largely determined by the asymptotic behavior of $\hat{P}(z)$, $\hat{h}_0(x, \hat{P}(z))$ and $\hat{q}(h_0(x, P(z)), P(z))$. The asymptotic properties of $\hat{P}(z)$ can be obtained by applying Theorem 3 of Heckman, Ichimura and Todd (1998). Analyzing the asymptotic behavior of $\hat{h}_0(x, \hat{P}(z))$ requires simple modifications of Theorems 3 and 4 of Heckman, Ichimura and Todd (1998). The modifications are needed because $h_0(X, P(Z))$ itself is not a conditional expectation, but it is the derivative of one. Analyzing the asymptotic properties of $\hat{q}(h_0(x, P(z)), P(z))$ is also slightly different because this is an estimator for the expectation of Y given $D = 1$, $h_1(X, P(Z))$ and $P(Z)$ evaluated at the value the random vector $\big(h_0(X, P(Z)), P(Z)\big)$ takes (and $D = 1$). The details of our trimming function and how these three estimators behave asymptotically are given in the Appendix.

5 Conclusion

This paper has shown identification and a consistent estimator of the average effect of a dummy endogenous variable in a nonparametric, weakly separable model. These results are promising for identification more generally in models with dummy endogenous variables. For example, the results can easily be extended to identification and estimation of the structural parameters of semiparametric models with dummy endogenous variables. As another example, the analysis of this paper can be immediately applied to identify state dependence in panel data models with

binary outcomes as long as there is a time-varying continuous regressor and the lagged dependent variables do not have random coefficients associated with them.

References

[1] Altonji, J. and H. Ichimura, 1998, "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables," unpublished manuscript, Northwestern University and University College London.

[2] Altonji, J. and R. Matzkin, 1998, "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," unpublished manuscript, Northwestern University.

[3] Amemiya, T., 1978, "The Estimation of a Simultaneous Equation Generalized Probit Model," *Econometrica*, **46**, 1193-1205.

[4] Andrews, D. and M. Schafgans, 1998, "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, **65**, 497-517.

[5] Angrist, J., 1991, "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," NBER Technical Working Paper No. 115.

[6] _____, 2001, "Estimation of Limited-Dependent Variable Models with Binary Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics.*

[7] Bhattacharya, J., D. McCaffrey, and D. Goldman, 1999, "Estimating Probit Models with Endogenous Covariates," unpublished working paper, RAND.

[8] Blundell, R. and J. Powell, 1999, "Endogeneity in Single Index Models," unpublished working paper, University College London and UC-Berkeley.

[9] Blundell, R. and J. Powell, 2000, "Endogeneity in Nonparametric and Semiparametric Regression Models," unpublished working paper presented at the World Conference of the Econometric Society.

[10] Cameron, S. and J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias", *Journal of Political Economy* 106:2, 262-333.

[11] Chen, X., J. Heckman, and E. Vytlacil, 1999, "Identification and $\sqrt{N}$ Estimation of Semiparametric Panel Data Models with Binary Dependent Variables and a Latent Factor," unpublished working paper, University of Chicago.

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

[13] Han, A. K., 1987, "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics,* **35**, 303-316.

[14] Heckman, J., 1978, "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica* , **46**, 931-959.

[15] _____, 1990a, "Varieties of Selection Bias," *American Economic Review*, **80**, 313-318.

[16] _____, 1990b, "Alternative Approaches to the Evaluation of Social Programs," Barcelona Lecture, World Conference of the Econometric Society.

[17] _____, 1997, "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, **32**, 441-462.

[18] Heckman, J., H. Ichimura, and P. Todd, 1998, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies,*, **65**, 261-294.

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

[20] Heckman, J., and E. Vytlacil, 1999, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences,* **96**, 4730-4734.

[21] _____, 2000, "The Relationship Between Treatment Parameters within a Latent Variable Framework," *Economics Letters,* January 2000, 66(1): 33-39.

[22] _____, 2001a, "Local Instrumental Variables," with J. Heckman, in C. Hsiao, K. Morimune, and J. Powell, eds., *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, (Cambridge: Cambridge University Press), 1-46.

[23] _____, 2001b, "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in M. Lechner and F. Pfeiffer, eds., *Econometric Evaluations of Active Labor Market Policies in Europe,* (Heidelberg; New York: Physica-Verlag), 1-23.

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

[25] Imbens, G. and W. Newey, 2001, "Identification and Inference in Triangular Simultaneous Equation Models without Additivity," unpublished working paper, UCLA and MIT.

[26] Manski, C., 1975, "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics,* **3**, 205-28.

[27] _____, (1988), "Identification of Binary Response Models Source," *Journal of the American Statistical Association* 83:403, 729-38

[28] Matzkin, R., 1991, "A Nonparametric Maximum Rank Correlation Estimator," in W. A. Barnett, J. L. Powell, and G. E. Tauchen, eds., *Nonparametric and Semiparametric Methods in Economics and Statistics,* (Cambridge University Press: Cambridge).

[29] Matzkin, R., 1992, "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica* 60:2, 239-70

[30] Matzkin, R., 1993, "Semiparametric Estimation of Monotone and Concave Utility Functions for Polychotomous Choice Models," *Econometrica* 59:5, 1315-27

[31] Matzkin, R., 2003, "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71:5, 1339-75.

[32] Newey, W., 1986, "Linear Instrumental Variable Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, **32**, 127-141.

[33] Rivers, D.c and Q. Vuong, 1988, "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics,* **39**, 347-366.

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

[35] Schafgans, M., 2000, "Finite Sample Properties for the Semiparametric Estimation of the Intercept of a Censored Regression Model," unpublished working paper, London School of Economics.

[36] Schafgans, M. and V. Zinde-Walsh, 2000, "On Intercept Estimation in the Sample Selection Model," unpublished working paper, London School of Economics.

[37] Silverman, B. W., 1978, "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives," *The Annals of Statistics*, **6**, 177-184.

[38] Vytlacil, E., 2002, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1): 331-41.

[39] Vytlacil, E., 2003, "Weak Separability, Additive Separability, and Linearity for Latent Indices of Threshold Crossing Models: Representation Results," unpublished working paper, Stanford Univeristy.

A Identification Proofs

Proof. (Lemma 3.1)

Consider the case where $p_1 > p_0$. (the case where $p_0 < p_1$ is symmetric). Consider the numerator of $h_1(p_0, p_1, x_1)$,

$$\begin{aligned} E(DY|X = x_1, P(Z) = p_1) - E(DY|X = x_1, P(Z) = p_0) & \\ &= \int_0^{p_1} E(Y_1|X = x_1, U = u)du - \int_0^{p_0} E(Y_1|X = x_1, U = u)du \\ &= \int_{p_0}^{p_1} E(Y_1|X = x_1, U = u)du \\ &= \int_{p_0}^{p_1} E(g(\nu(x_1, 1), \epsilon)|U = u)du, \end{aligned}$$

where the last equality is using assumption (A-2). Likewise, for the numerator of $h_0(p_0, p_1, x_0)$, we have

$$\begin{aligned} -\Big[E((1-D)Y|X = x_0, P(Z) = p_1) - E((1-D)Y|X = x_0, P(Z) = p_0)\Big] & \\ &= \int_{p_0}^{p_1} E(g(\nu(x_0, 0)\epsilon)|U = u)du. \end{aligned}$$

Thus,

$$h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) = \frac{1}{p_1 - p_0}\int_{p_0}^{p_1} E(g(\nu(x_1, 1), \epsilon) - g(\nu(x_0, 0), \epsilon)|U = u)du.$$

Using assumption (A-4), we have that the sign of this expression will be determined by the sign of $\nu(x_1, 0) - \nu(x_0, 1)$. Q.E.D..

Proof: (Lemma 3.1) Consider assertion (1). By Lemma 3.1, $\nu(\tilde{x}, 1) = \nu(x, 0)$ for any $\tilde{x} \in h_1^{-1}h_0(x)$. Thus,

$$\begin{aligned} E(DY|X \in h_1^{-1}h_0(x), P(Z) = p) & \\ &= E(\mathbf{1}[U \leq P(Z)]g(\nu(X, 1), \epsilon)|X \in h_1^{-1}h_0(x), P(Z) = p) \\ &= \int\left[\int \mathbf{1}[U \leq p]g(\nu(\tilde{x}, 1), \epsilon)dG(\tilde{x}|X \in h_1^{-1}h_0(x), P = p))\right]dF_{\epsilon,U} \\ &= \int\left[\int \mathbf{1}[U \leq p]g(\nu(x, 0), \epsilon)dG(\tilde{x}|X \in h_1^{-1}h_0(x), P(Z) = p)\right]dF_{\epsilon,U} \\ &= \int \mathbf{1}[U \leq p]g(\nu(x, 0), \epsilon)dF_{\epsilon,U} \\ &= E(DY_0|X = x, P = p) \end{aligned}$$

where $G(\tilde{x}|X \in h_1^{-1}h_0(x), P = p)$ is the distribution of X conditional on $X \in h_1^{-1}h_0(x), P = p$, and $F_{\epsilon,U}$ is the distribution of (ϵ, U). The first equality follows from plugging in the model for Y and D given by equations (1) and (2); the second equality follows from assumption (A-2), that $(X, Z) \perp\!\!\!\perp (\epsilon, U)$; and the third equality is using that $\nu(\tilde{x}, 1) = \nu(x, 0)$ for any $\tilde{x} \in h_1^{-1}h_0(x)$ by Lemma 3.1. Thus,

$$\begin{aligned} E(DY|X \in h_1^{-1}h_0(x), P = p) + E((1-D)Y|X = x, P = p) & \\ &= E(DY_0|X = x, P = p) + E((1-D)Y_0|X = x, P = p) \\ &= E(Y_0|X = x, P = p) \\ &= E(Y_0|X = x), \end{aligned}$$

so that

$$\begin{aligned} \int \Big(E(DY|X \in h_1^{-1}h_0(x), P = p) + E((1-D)Y|X = x, P = p) \Big) dG_{P|X}(p|x) & \\ &= \int E(Y_0|X = x) dG_{P|X}(p|x) \\ &= E(Y_0|X = x) \end{aligned}$$

and the result now follows immediately. Assertions (2) follow from an analogous argument, and assertion (3) follows from assertions (1) and (2). QED.

B Estimation Proofs: Main Results

$$\begin{aligned}\sqrt{N}(\tilde{\Delta} - E(Y_0|A_1 \cap A_2)) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\frac{\left[(1-D_i)Y_i + D_i \hat{q}(\hat{h}_{0i}, \hat{P}_i)\right] \hat{I}_{1i}\hat{I}_{2i}}{N^{-1}\sum_{i=1}^{n} \hat{I}_{1i}\hat{I}_{2i}} - E(Y_0|(X,Z) \in A_1 \cap A_2) \right] \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\left[(1-D_i)Y_i + D_i \hat{q}(\hat{h}_{0i}, \hat{P}_i) - E(Y_0|(X,Z) \in A_1 \cap A_2)\right] \hat{I}_{1i}\hat{I}_{2i}}{N^{-1}\sum_{i=1}^{n} \hat{I}_{1i}\hat{I}_{2i}}\end{aligned}$$

This is a multiple step estimator. In the first step. the joint density of (X, Z) is estimated using kernel density estimation. In the second step, the conditional expectation of D given Z is estimated using local polynomial regression of D on Z. The third step uses local polynomial regression to estimate the derivative with respect to $P(\cdot)$ of $E\big[(1-D)Y|X, P(Z)\big]$ and $E\big[DY|X, P(Z)\big]$. Since in reality, we never observe $P(Z_i)$, this step uses the estimated values, $\hat{P}(Z_i)$. On the other hand, since asymptotic variance of local polynomial regression based estimators is inversely related to the density of the conditioning variables, the estimation needs to be done in a region where this density is above a certain strictly positive level. For this reason in the later steps of our estimation process we use a trimming function which is based on the estimated density of (X, Z). In the fourth step we estimate $E\big[Y|D=1, h_1(P(Z), X), P(Z)\big]$ by local polynomial regression using the estimated values of $h_1(\cdot,\cdot)$ and $P(\cdot)$. This function, however, needs to be evaluated at $D = 1$ and the estimated values of the random functions $h_0(\cdot,\cdot)$ and $P(\cdot)$, which may or may not be in the support of $\big(h_1(P(Z), X), P(Z)\big)$. Since our estimator is well defined only when the supports of these random vectors overlap, we employ another trimming function, which is based on the estimated density of (h_1, P) and uses the estimated values of (h_1, P). The use of this second trimming function guarantees that our estimation is done on an estimated region where the supports of the two random vectors overlap.

The notation that is used to define our estimator is the following:

$$\begin{aligned} P(z) &= E(D|Z=z) & \hat{P}(z) &= \hat{E}(D|Z=z) \\ h_1(x, P(z)) &= \frac{\partial E\left[DY|X=x, P(Z)=P(z)\right]}{\partial P} & \hat{h}_1(x, \hat{P}(z)) &= \frac{\partial \hat{E}\left[DY|X=x, \hat{P}(Z)=\hat{P}(z)\right]}{\partial P} \\ h_0(x, P(z)) &= \frac{\partial E\left[-(1-D)Y|X=x, P(Z)=P(z)\right]}{\partial P} & \hat{h}_0(x, \hat{P}(z)) &= \frac{\partial \hat{E}\left[-(1-D)Y|X=x, \hat{P}(Z)=\hat{P}(z)\right]}{\partial P} \end{aligned}$$

$$\begin{aligned} q(h_0(x, P(z)), P(z)) &= E\left[Y|D=1, h_1(X, P(Z)) = h_0(x, P(z)), P(Z) = P(z)\right] \\ \hat{q}(\hat{h}_0(x, \hat{P}(z)), \hat{P}(z)) &= \hat{E}\left[Y|D=1, \hat{h}_1(X, \hat{P}(Z)) = \hat{h}_0(x, \hat{P}(z)), \hat{P}(Z) = \hat{P}(z)\right] \end{aligned}$$

$$I_{1i} := 1\left\{f_{X,Z}(X_i, Z_i) \geq q_{01}\right\} \qquad \hat{I}_{1i} := 1\left\{\hat{f}_{X,Z}(X_i, Z_i) \geq q_{01}\right\}$$

$$
\begin{aligned}
I_{2i} &:= 1\left\{f_{h_1(X,P(Z)),P(Z)}(h_0(X_i,P(Z_i)),P(Z_i)) \geq q_{02}\right\} \\
\hat{I}_{2i} &:= 1\left\{\hat{f}_{\hat{h}_1(X,\hat{P}(Z)),\hat{P}(Z)}(\hat{h}_0(X_i,\hat{P}(Z_i)),\hat{P}(Z_i)) \geq q_{02}\right\} \\
A_1 &:= \{(x,z) \in \text{supp}(X,Z) : f_{X,Z}(x,z) \geq q_{01}\} \\
A_2 &:= \left\{(x,z) \in \text{supp}(X,Z) : f_{h_1(X,P(Z)),P(Z)}(h_0(x,P(z)),P(z)) \geq q_{02}\right\}
\end{aligned}
$$

where we use the subscript i as a shorthand for the value one of these functions takes at (X_i, Z_i). For example, $\hat{P}_i$ is the shorthand notation for $\hat{P}(Z_i)$.

To study the asymptotic properties of our estimator, we break it into several pieces and study the behavior of each piece separately. In analyzing the behavior of each piece we rely on the analysis of Heckman, Ichimura and Todd (1998) extensively. In particular, the equicontinuity and Hoeffding, Powell, Stock and Stoker lemmas stated in Heckman, Ichimura and Todd (1998) are repeatedly used in our analysis. Therefore, before starting our analysis, it may be helpful two state these two lemmas. To state the two lemmas we need to define some notation: For $r = 1$ and 2, let $\mathcal{S}^r$ denote the r-fold product space of $\mathcal{S} \subset \mathbb{R}^d$ and define a class of functions Λ_N over $\mathcal{S}^r$. For any $\lambda_N \in \Lambda_N$, write λ_{N,i_r} as a short hand for either $\lambda_N(s_i)$ or $\lambda_N(s_{i_1}, s_{i_2})$, where $i_1 \neq i_2$. We define $U_N\lambda_N = \sum_{i_r} \lambda_{N,i_r}$, where $\sum_{i_r}$ denotes the summation over all permutations of r elements of $\{s_1, ..., s_N\}$ for $r = 1$ or 2. Then $U_N\lambda_N$ is called a U-process over $\lambda_N \in \Lambda_N$. For $r = 2$, we assume that $\lambda_N(S_i, S_j) = \lambda_N(S_j, S_i)$. Note that a normalizing constant might be included as a part of λ_N. We call a U-process degenerate if all conditional expectations given other elements are 0. When $r = 1$, this condition is defined to mean that $E\lambda_N = 0$.

In the following, we assume that Λ_N is a subset of $\mathcal{L}^2(\mathbb{P}^r)$, the $\mathcal{L}^2$ space defined over $\mathcal{S}^r$ using the product measure of $\mathbb{P}$, $\mathbb{P}^r$. $D_2(\tau, \Lambda_N)$ denotes the $\mathcal{L}^2$ packing number of Λ_N[19]. On the other hand, $||\lambda_N||_2 := \sqrt{\sum_{i_r} E(\lambda_{N,i_r})^2}$.

Equicontinuity Lemma (Heckman, Ichimura and Todd (1998)): Let $\{S_i\}_{i=1}^N$ be an iid sequence of random variables generated by $\mathbb{P}$. For a degenerate U-process $\{U_N\lambda_N\}$ over a separable class of functions $\Lambda_N \subset \mathcal{L}^2(\mathbb{P}^r)$ suppose the following assumptions hold:

(i) There exists an $F_N \in \mathcal{L}^2(\mathbb{P}^r)$ such that for any $\lambda_N \in \Lambda_N$, $|\lambda_N| < F_N$ such that $\limsup_{N\to\infty} \sum_{i_r} E(F_{N,i_r}^2) < \infty$;

(ii) For each $\delta > 0$, $\lim_{N\to\infty} \sum_{i_r} E(F_{N,i_r}^2 1\{F_{N,i_r} > \delta\}) = 0$;

(iii) There exists $\alpha(\tau)$ and $\overline{\tau} > 0$ such that for each $0 < \tau \leq \overline{\tau}$, $D_2(\tau, \Lambda_N) \leq \alpha(\tau)$ $\mathbb{P}$ almost surely and $\int_0^{\overline{\tau}} [\log \alpha(t)]^{r/2} \mathrm{d}t < \infty$.

Then for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$
\lim_{N\to\infty} P\left(\sup_{||\lambda_{1N} - \lambda_{2N}||_2 \leq \delta} |U_N(\lambda_{1N} - \lambda_{2N})| > \epsilon\right) = 0
$$

[19]The $\mathcal{L}^2$ packing number $D_2(\tau, T_0)$ for a subset T_0 of a metric space is defined as the largest J for which there exist points $t_1, ..., t_J$ in T_0 with $\sqrt{E|t_i - t_j|^2} > \tau$ for $i \neq j$.

Hoeffding, Powell, Stock and Stoker Lemma: Suppose $\{S_i\}_{i=1}^N$ is i.i.d., $U_N\lambda_N = (N(N-1))^{-1}\sum_{i_r}\lambda_N(S_i,S_j)$ where λ_N is symmetric in its arguments, $E[\lambda_N(S_i,S_j)]=0$, and $\hat{U}_N\lambda_N = N^{-1}\sum_{i=1}^N 2p_N(S_i)$, with $p_N(S_i)=E[\lambda_N(S_i,S_j)|S_i]$. If $E[\lambda_N(S_i,S_j)^2]=o(N)$, then $NE[(U_N\lambda_N - \hat{U}_N\lambda_N)^2]=o(1)$.

Now we are ready to state our assumptions. Suppose $\{\tilde{h}_{N1}\}$, $\tilde{K}_1$, $\{\tilde{h}_{N2}\}$ and $\tilde{K}_2$ denote the bandwidth parameter sequence and kernel function used to estimate $f_{X,Z}$ and $f_{h_1,P}$, respectively. Similarly, let $\{h_{NP}\}$, $\{h_{Nh}\}$ and $\{h_{Nq}\}$ and K^P, K^h and K^q denote the bandwidth sequences and kernel functions used in estimating, $P(Z)$, h_0 (h_1)[20] and q, respectively. We will call a function p-smooth if it is $p+1$ times continuously differentiable and its $p+1^{st}$ derivative is Holder continuous with parameter $0 < a \leq 1$[21].

Assumption B.1 *$\{D_i, Y_i, X_i, Z_i\}$ are i.i.d., (X_i, Z_i) takes values in $\mathbb{R}^{d_X}\times\mathbb{R}^{d_Z}=\mathbb{R}^d$, and $var(Y_i)<\infty$*

Assumptions related to the estimation of $f_{X,Z}$ and $f_{h_1,P}$:

Assumption B.2 *(a) $f_{X,Z}$ and $f_{h_1,P}$ are both uniformly continuous and have uniformly continuous first derivatives.*

(b) $f_{X,Z}$, P, h_0, h_1 and $f_{h_1,P}$ are all $\tilde{p}$-smooth with $\tilde{p} > d$[22]*.*

(c) Let $q_{01}>0$ and $q_{02}>0$ be such that

(i) q_{01} has a neighborhood U such that $f_{X,Z}(X,Z)$ has a continuous Lebesgue density that is strictly positive on U. Moreover for each $(x,z)\in f_{X,Z}^{-1}(U)$, $||Df_{X,Z}(x,z)||>0$.

(ii) q_{02} has a neighborhood V such that $f_{h_1,P}(h_1(X,P(Z)),P(Z))$ has a continuous Lebesgue density that is strictly positive on V. Moreover for each $(x,z)\in f_{X,Z}^{-1}(U)$, $||Df_{h_1,P}(h_0(x,P(z))||>0$.

(d) (i) For each $z\in\text{supp}(Z)$ such that there exists an $x\in\text{supp}(X)$ with $(x,z)\in f_{X,Z}^1(U)$, $||DP(z)||>0$.

(ii) For each $(x,z)\in f_{X,Z}^{-1}(U)$, $||D_x h_1(x,P(z))||>0$, and $||D_P h_1(x,P(z))||>0$.

(e) $\tilde{K}_1$ and $\tilde{K}_2$ satisfy Condition (C). $\tilde{K}_2$ is Lipschitz. Moreover, $\tilde{K}_1'$, and $\tilde{K}_2'$ satisfy parts (a), (b) and (d) of Condition (C), where

[20] We can use the same kernel function and bandwidth sequence in the estimation of h_0 and h_1.

[21] We use the same definiton as in Heckman, Ichimura and Todd. Namely, we say a function ϱ is Holder continuous at $X=x_0$ with constant $0<a\leq 1$ if $|\varrho(x,t)-\varrho(x_0,t)|\leq C||x-x_0||^a$ for some $C>0$ for all x and t in the domain of the function $\varrho(\cdot,\cdot)$. We assume that Holder continuity holds uniformly over t whenever there is an additional argument.

[22] Note that these conditions are used to guarantee that the composite function $f_{h_1,P}(h_1(x,P(z)),P(z))$ is $\tilde{p}_1$-smooth.

Definition B.1 Condition (C):

(a) $\tilde{K}$ *is uniformly continuous (with modulus of continuity* $w_{\tilde{K}}$*) and of bounded variation* $V(\tilde{K})$

(b) $\int |\tilde{K}(x)|\mathrm{d}x < \infty$ *and* $\tilde{K}(x) \to 0$ *as* $||x|| \to \infty$

(c) $\int \tilde{K}(x)\mathrm{d}x = 1$

(d) $\int \sqrt{||(x \log ||x||)||}|\mathrm{d}\tilde{K}(x)| < \infty$

(f) (i) $\tilde{h}_{N1} \to 0$, $\frac{\log N}{N\tilde{h}_{N1}^{d+1}} \to 0$.

(ii) $\tilde{h}_{N2} \to 0$, $\frac{\log N}{N\tilde{h}_{N2}^{3}} \to 0$, *and* $N\tilde{h}_{N2}^{12} \to c \in (0, \infty]$.

Assumptions related to the estimation of $E(D|Z)$**:**

Assumption B.3 *(a) Bandwidth sequence* $\{h_{NP}\}$ *satisfies* $h_{NP} \to 0$, $Nh_{NP}^{d_z}/\log N \to \infty$, *and* $Nh_{NP}^{2\overline{p}_P} \to c_P \in (0, \infty)$, *where* $d_z < \overline{p}_P \leq \tilde{p}$ [23]

(b) Kernel function K^P *is symmetric, supported on a compact set, and is Lipschitz continuous. Also it has moments of order* $p_P + 1$ *through* $\overline{p}_P - 1$ *that are equal to 0.*

Assumptions related to the estimation of $E(DY|P(Z), X)$ **and** $E(-(1-D)Y|P(Z), X)$**:**

Assumption B.4 *(a)* $\{h_{Nh}\}$ *satisfies* $Nh_{Nh}^{d_x+2}/\log N \to \infty$ *and* $Nh_{Nh}^{2(\overline{p}_h-1)} \to c_h < \infty$ *for some* $c_h \geq 0$, *and* $\overline{p}_h > d_x + 2$.

(b) Kernel function $K^h(\cdot)$ *is 1-smooth, symmetric and supported on a compact set. It has moments of order* $p + 1$ *through* $\overline{p}_h - 1$ *that are equal to zero.*

Assumptions related to the estimation of $E(Y|D = 1, h_1(X, P(Z)), P(Z)), P(Z))$**:**

Assumption B.5 *(a)* $\{h_{Nq}\}$ *satisfies* $Nh_{Nq}^2/\log N \to \infty$ *and* $Nh_{Nq}^{2\overline{p}_q} \to c_q < \infty$ *for some* $c_q \geq 0$ *and* $\overline{p}_q > 2$.

(b) Kernel function $K^q(\cdot)$ *is 1-smooth, symmetric and supported on a compact set. It has moments of order* $p + 1$ *through* $\overline{p}_q - 1$ *that are equal to zero.*

(c) There exists $\eta \in \mathbb{R}_+$ *such that* $P\big(P(Z) > \eta\big) = 1$.

[23] In principle, we could choose $\overline{p}_P = \tilde{p}$. But to control the bias of our local polynomial estimator, certain moments of the kernel function we use must be zero, and using $\tilde{p}$ requires more moments of this function to be 0.

We are now ready to study the asymptotic behavior of our estimator. Note that by the Mean Value Theorem we have

$$\begin{aligned}\sqrt{N}\big[\tilde{\Delta} - E(Y_0|A_1\cap A_2)\big] &= \Big\{\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N\big[(1-D_i)Y_i + D_i q(h_{0i},P_i) - E(Y_0|(X,Z)\in A_1\cap A_2)\big]\hat{I}_{1i}\hat{I}_{2i}\\ &+ \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N D_i\big[\hat{q}(h_{0i},P_i) - q(h_{0i},P_i)\big]\hat{I}_{1i}\hat{I}_{2i}\\ &+ \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N D_i\tfrac{\partial\hat{q}}{\partial h_1}(\tilde{h}_{0i},\tilde{P}_i)\big[\hat{h}_0(X_i,\hat{P}(Z_i)) - h_0(X_i,P(Z_i))\big]\hat{I}_{1i}\hat{I}_{2i}\\ &+ \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N D_i\tfrac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i)\big[\hat{P}(Z_i) - P(Z_i)\big]\hat{I}_{1i}\hat{I}_{2i}\Big\}\\ &\div \Big[N^{-1}\textstyle\sum_{i=1}^N\hat{I}_{1i}\hat{I}_{2i}\Big]\end{aligned} \tag{10}$$

where for each i, $(\tilde{h}_{0i},\tilde{P}_i)$ is between (h_{0i},P_i) and $(\hat{h}_{0i},\hat{P}_i)$.

We will study the asymptotic behavior of the denominator and each piece of the numerator separately. The asymptotic behavior of the last three terms of the numerator is largely determined by the asymptotic behavior of $\hat{P}(\cdot)$, $\hat{h}_0(\cdot,\hat{P}(\cdot))$ and $\hat{q}(h_0(\cdot,P(\cdot)),P(\cdot))$. The asymptotic properties of $\hat{P}$ can be obtained by applying Theorem 3 of HIT. Analyzing the asymptotic behavior of $\hat{h}_0(\cdot,\hat{P}(\cdot))$ requires simple modifications of Theorems 3 and 4 of HIT. The modifications are needed because h_0 itself is not a conditional expectation, but it is the derivative of one. Heckman, Ichimura and Todd are interested in the first element of the estimated coefficient vector, we are interested in the second element. Analyzing the asymptotic properties of $\hat{q}(h_0,P)$ is also slightly different because this is an estimator for the expectation of Y given $D=1$, h_1 and P evaluated at the value the random vector $\big(h_0,P\big)$ takes (and $D=1$). Evaluating this conditional expectation at the value $\big(h_0,P\big)$ takes is meaningful only when that value is an element of the support of $\big(h_1,P\big)$. To guarantee that this is indeed the case we have to use another trimming function. The details of our trimming function and how these three estimators behave asymptotically are given in Appendix C.

We now proceed as follows. Steps 1 and 2 (sections B.1 and B.2) examine the last two terms of the numerator of equation 10. Step 3 (section B.3) considers the second term of the numerator. In step 4 (section B.4), we consider the first term of the numerator of equation 10. In step 5 (section B.5), we consider the denominator of equation 10. The result stated in the text then immediately follows from Slutsky.

B.1 Step 1:

By adding and subtracting a term we observe that

$$\begin{aligned}\tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N D_i\tfrac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i)\big(\hat{P}(Z_i)-P(Z_i)\big)\hat{I}_{1i}\hat{I}_{2i} &= \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N D_i\big[\tfrac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i)-\tfrac{\partial q}{\partial P}(h_{0i},P_i)\big]\big(\hat{P}(Z_i)-P(Z_i)\big)\hat{I}_{1i}\hat{I}_{2i}\\ &+ \tfrac{1}{\sqrt{N}}\textstyle\sum_{i=1}^N D_i\tfrac{\partial q}{\partial P}(h_{0i},P_i)\big(\hat{P}(Z_i)-P(Z_i)\big)\hat{I}_{1i}\hat{I}_{2i}\end{aligned}$$

We first show that the first term of this expression is $o_p(1)$, so that the whole expression is asymptotically equivalent to the second term. By the results in Appendix C.1, we know that

$$\big[\hat{P}(z) - P(z)\big]\hat{I}_1(x,z) = N^{-1}\sum_{i=1}^{N}\psi_{NP}(D_i, Z_i; x, z) + \hat{b}_P(x,z) + \hat{R}_P(x,z), \tag{11}$$

where $E[\psi_{NP}(D_i, Z_i; X, Z)|X = x, Z = z] = 0$, $\text{plim}_{N\to\infty} N^{-1/2}\sum_{i=1}^{N}\hat{b}_P(X_i, Z_i) = b_P < \infty$, and $\text{plim}_{N\to\infty} N^{-1/2}\sum_{i=1}^{N}\hat{R}_P(X_i, Z_i) = 0$. Substituting in this expression, we obtain

$$\begin{array}{rcl} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\big[\frac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i) - \frac{\partial q}{\partial P}(h_{0i},P_i)\big]\big(\hat{P}(Z_i) - P(Z_i)\big)\hat{I}_{1i}\hat{I}_{2i} & = & \frac{1}{N\sqrt{N}}\sum_{i=1}^{N}\sum_{j=1}^{N} D_i\big[\frac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i) - \frac{\partial q}{\partial P}(h_{0i},P_i)\big]\psi_{NP}(D_j,Z_j;X_i,Z_i)\hat{I}_{2i} \\ + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\big[\frac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i) - \frac{\partial q}{\partial P}(h_{0i},P_i)\big]\hat{b}_P(X_i,Z_i)\hat{I}_{2i} & + & \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\big[\frac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i) - \frac{\partial q}{\partial P}(h_{0i},P_i)\big]\hat{R}_P(X_i,Z_i)\hat{I}_{2i}. \end{array}$$

By the results in Appendices C.1 and C.2, we know that $\hat{P}$ is uniformly consistent for P, and $\hat{h}_0(\hat{P}(\cdot),\cdot)$ is uniformly consistent for $h_0(P(\cdot),\cdot)$ on our region of estimation. Applying theorem 4 of Heckman, Ichimura and Todd to $\hat{q}$ for the set of observations for which $D_i = 1$, we also know that $\frac{\partial\hat{q}}{\partial P}(h,p)$ is uniformly consistent for $\frac{\partial q}{\partial P}(h,p)$ on $A_1 \cap A_2$[24]. Then using the equicontinuity lemma we can show that the probability limit of each of these terms is 0[25], so that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial\hat{q}}{\partial P}(\tilde{h}_{0i},\tilde{P}_i)\big(\hat{P}(Z_i) - P(Z_i)\big)\hat{I}_{1i}\hat{I}_{2i} \overset{AE}{=} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial q}{\partial P}(h_{0i},P_i)\big(\hat{P}(Z_i) - P(Z_i)\big)\hat{I}_{1i}\hat{I}_{2i}.$$

Using equation 11, we see that the latter term equals

$$\frac{1}{N\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial q}{\partial P}(h_{0i},P_i)\sum_{j=1}^{N}\psi_{NP}(D_j,Z_j;X_i,Z_i)\hat{I}_{2i} + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial q}{\partial P}(h_{0i},P_i)\hat{b}_P(X_i,Z_i)\hat{I}_{2i} + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial q}{\partial P}(h_{0i},P_i)\hat{R}_P(X_i,Z_i)\hat{I}_{2i}.$$

Using continuity of $\frac{\partial q}{\partial P}(h_{0i},P_i)$, compactness of $A_1 \cap A_2$, and the explicit form of $\hat{b}_P$, and $\hat{R}_P$, we can show that

$$b_{qP} := \text{plim}_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial q}{\partial P}(h_{0i},P_i)\hat{b}_P(X_i,Z_i)\hat{I}_{2i} < \infty$$

and

$$\text{plim}_{N\to\infty}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\frac{\partial q}{\partial P}(h_{0i},P_i)\hat{R}_P(X_i,Z_i)\hat{I}_{2i} = 0.$$

[24]Note that $A_1 \cap A_2 \subset \overline{A}_1$ where $\overline{A}_1 := \{(x,z) : f_{X,Z}(x,z) \geq q_{01} - \epsilon_{f1}\}$ with ϵ_{f1} denoting the same positive number as used in defining $\mathcal{H}_1$ in Appendix C.4

[25]The equicontinuity lemma applies to symmetric, degenerate U-processes. Even though these three sums are not necessarily symmetric and degenerate, by adding and subtracting some terms, we can break them into three pieces consisting of symmetric, degenerate U-processes, and two terms involving the expectations of the latter two of the initial sums. Verifying the conditions of the equicontinuity lemma for the symmetric, degenerate processes is straightforward. On the other hand, by the dominated convergence theorem the terms involving the expectations of the initial sums are all of the form $\sqrt{Nh_{NP}^{2\tilde{p}}}$ times some term that goes to 0.

On the other hand, using the equicontinuity lemma once more, we can show that

$$\frac{1}{N\sqrt{N}}\sum_{i=1}^{N} D_i \frac{\partial q}{\partial P}(h_{0i}, P_i) \sum_{j=1}^{N} \psi_{NP}(D_j, Z_j; X_i, Z_i) \left(\hat{I}_{2i} - I_{2i}\right) = o_P(1).$$

Combining these results, we conclude that

$$N^{-1/2}\sum_{i=1}^{N} D_i \tfrac{\partial \hat{q}}{\partial P}(\tilde{h}_{0i}, \tilde{P}_i)\left(\hat{P}(Z_i) - P(Z_i)\right)\hat{I}_{1i}\hat{I}_{2i} \overset{AE}{=} N^{-3/2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_i \tfrac{\partial q}{\partial P}(h_{0i}, P_i)\psi_{NP}(D_j, Z_j; X_i, Z_i) I_{2i} + b_{qP}.$$

Next, we focus on

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{D_i \frac{\partial q}{\partial P}(h_{0i}, P_i)\psi_{NP}(D_j, Z_j; X_i, Z_i) I_{2i}}{N\sqrt{N}} &= \sum_{i=1}^{N} \frac{D_i \frac{\partial q}{\partial P}(h_{0i}, P_i)\psi_{NP}(D_i, Z_i; X_i, Z_i) I_{2i}}{N\sqrt{N}} \\ &+ \sum_{i=1}^{N}\sum_{j\neq i} \frac{D_i \frac{\partial q}{\partial P}(h_{0i}, P_i)\psi_{NP}(D_j, Z_j; X_i, Z_i) I_{2i}}{N\sqrt{N}}.\end{aligned}$$

By applying a strong law of large numbers to the first of these terms we see that this expression is asymptotically equivalent to

$$\frac{E\left[D_i \varepsilon_i^P \frac{\partial q}{\partial P}(h_{0i}, P_i) e_1[M_{pN}^p(Z_i)]^{-1} e_1' K^p(0) I_{1i} I_{2i}\right]}{\sqrt{N} h_{NP}^{d_z}} + \sum_{i=1}^{N}\sum_{j\neq i} \frac{D_i \frac{\partial q}{\partial P}(h_{0i}, P_i)\psi_{NP}(D_j, Z_j; X_i, Z_i) I_{2i}}{N\sqrt{N}}.$$

Since $N h_{NP}^{2d_z} \to \infty$, this in turn is asymptotically equivalent to

$$N^{-3/2}\sum_{i=1}^{N}\sum_{j\neq i} D_i \tfrac{\partial q}{\partial P}(h_{0i}, P_i) I_{2i}\psi_{NP}(D_j, Z_j; X_i, Z_i) = \tag{12}$$
$$\tfrac{N-1}{N}\sum_{i=1}^{N}\sum_{j\neq i} \frac{\left[\frac{1}{2} D_i \frac{\partial q}{\partial P}(h_{0i}, P_i) I_{2i}\psi_{NP}(D_j, Z_j; X_i, Z_i) + \frac{1}{2} D_j \frac{\partial q}{\partial P}(h_{0j}, P_j) I_{2j}\psi_{NP}(D_i, Z_i; X_j, Z_j)\right]}{\sqrt{N}(N-1)}.$$

Since $\lim_{N\to\infty} \frac{N-1}{N} = 1$, the asympototic behavior of (12) is the same as the asymptotic behavior of $\sum_{i=1}^{N}\sum_{j\neq i} \frac{\zeta_N(D_i, X_i, Z_i, D_j, X_j, Z_j)}{\sqrt{N}(N-1)}$ where

$$\zeta_N(D_i, X_i, Z_i, D_j, X_j, Z_j) = \tfrac{1}{2}\left[D_i \tfrac{\partial q}{\partial P}(h_{0i}, P_i) I_{2i}\psi_{NP}(D_j, Z_j; X_i, Z_i) + D_j \tfrac{\partial q}{\partial P}(h_{0j}, P_j) I_{2j}\psi_{NP}(D_i, Z_i; X_j, Z_j)\right].$$

By the law of iterated expectations, $E\left[\zeta_N(D_i, X_i, Z_i, D_j, X_j, Z_j)\right] = 0$. Then by Hoeffding, Powell, Stock and Stoker lemma, if $E\left(\zeta_N(D_i, X_i, Z_i, D_j, X_j, Z_j)\right)^2 = o(N)$,

$$\begin{aligned}\frac{\sum_{i=1}^{N}\sum_{j\neq i}\zeta_N(D_i, X_i, Z_i, D_j, X_j, Z_j)}{\sqrt{N}(N-1)} &\xrightarrow{P} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} 2E[\zeta_N(D_i, X_i, Z_i, D_j, X_j, Z_j)|D_i, X_i, Z_i] \\ &= \sum_{i=1}^{N} \frac{E\left[D_j \frac{\partial q}{\partial P}(h_{0j}, P_j) I_{2j}\psi_{NP}(D_i, Z_i; X_j, Z_j)|D_i, X_i, Z_i\right]}{\sqrt{N}}.\end{aligned}$$

Next, we show that $E\zeta(D_i, X_i, Z_i, D_j, X_j, Z_j)^2 = o(N)$ under our basic assumptions. By the Cauchy-Schwarz inequality, it suffices to show that

$$E\left\{D_i^2\left(\tfrac{\partial q}{\partial P}(h_{0i}, P_i)\right)^2 I_{2i}^2 I_{1i}^2 (\varepsilon_j^P)^2 \left(e_1[M_{pN}^P(Z_i)]^{-1}\left[\left(\tfrac{Z_j - Z_i}{h_{NP}}\right)^{Q_p}\right]'\right)^2 h_{NP}^{-2d_z} K^P\left(\tfrac{Z_j - Z_i}{h_{NP}}\right)^2\right\} = o(N).$$

Since $Nh_{NP}^{2d_z} \to \infty$, and $\frac{1}{Nh_{NP}^{2d_z}} \to 0$, the required condition will hold if, for each N,

$$E\left\{ D_i \left(\frac{\partial q}{\partial P}(h_{0i}, P_i) \right)^2 I_{1i} I_{2i} (\varepsilon_j^P)^2 \left(e_1 [M_{pN}^P(Z_i)]^{-1} \left[\left(\frac{Z_j - Z_i}{h_{NP}} \right)^{Q_p} \right]' \right)^2 \left(K^P \left(\frac{Z_j - Z_i}{h_{NP}} \right) \right)^2 \right\} < \infty.$$

For sufficiently large N this is true, because M_{pN}^P is nonsingular, the kernel function is 0 outside a compact set and $\frac{\partial q}{\partial P}$ and K^P are continuous functions. Thus,

$$\begin{gathered} \tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^N D_i \frac{\partial \hat{q}}{\partial P}(\tilde{h}_{0i}, \tilde{P}_i) \big(\hat{P}(Z_i) - P(Z_i) \big) \hat{I}_{1i} \hat{I}_{2i} \overset{AE}{=} \\ N^{-1/2} \textstyle\sum_{i=1}^N E\left[D_j \frac{\partial q}{\partial P}(h_{0j}, P_j) I_{2j} \psi_{NP}(D_i, Z_i; X_j, Z_j) | D_i, Z_i, X_i \right] + b_{qP} \\ = N^{-1/2} \textstyle\sum_{i=1}^N E\left[D_j \frac{\partial q}{\partial P}(h_{0j}, P_j) I_{2j} \psi_{NP}(D_i, Z_i; X_j, Z_j) | Y_i, D_i, Z_i, X_i \right] + b_{qP}. \end{gathered}$$

B.2 Step 2:

By Appendix C.2, we know that

$$\big[\hat{h}_0(X_i, \hat{P}(Z_i)) - h_0(X_i, P(Z_i)) \big] \hat{I}_1(x, z) = N^{-1} \sum_{i=1}^N \psi_{Nh_0P}(D_i, Y_i, X_i, Z_i; x, z) + \hat{b}_{\hat{h}_0}(x, z) + \hat{R}_{\hat{h}_0}(x, z)$$

where $E[\psi_{Nh_0P}(D_i, Y_i, X_i, Z_i; X, Z) | X = x, Z = z] = 0$, $\text{plim}_{N\to\infty} N^{-1/2} \sum_{i=1}^N \hat{b}_{\hat{h}_0}(X_i, Z_i) = b_{h_0P} < \infty$, and $\text{plim}_{N\to\infty} N^{-1/2} \sum_{i=1}^N \hat{R}_{\hat{h}_0}(X_i, Z_i) = 0$. Then using arguments similar to those in step 1, we can show that

$$\begin{gathered} \tfrac{1}{\sqrt{N}} \textstyle\sum_{i=1}^N D_i \frac{\partial \hat{q}}{\partial h_1}(\tilde{h}_{0i}, \tilde{P}_i) \big[\hat{h}_0(X_i, \hat{P}(Z_i)) - h_0(X_i, P(Z_i)) \big] \hat{I}_{1i} \hat{I}_{2i} \overset{AE}{=} \\ \tfrac{1}{N\sqrt{N}} \textstyle\sum_{i=1}^N \sum_{j=1}^N D_i \frac{\partial q}{\partial h_1}(h_{0i}, P_i) \psi_{Nh_0P}(D_j, Y_j, X_j, Z_j; X_i, Z_i) I_{2i} + b_{qh_0P} \end{gathered}$$

where $b_{qh_0P} := \text{plim}_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{i=1}^N D_i \frac{\partial q}{\partial h_1}(h_{0i}, P_i) \hat{b}_{\hat{h}_0}(X_i, Z_i) \hat{I}_{2i} < \infty$. Then just as in **Step 1** we can break the sum involving ψ_{Nh_0P} into two pieces: one consisting of terms with the same index, and the other consisting of terms with different indices. For the first of these sums, we apply a strong law of large numbers and use $Nh_{Nh_0}^{2d_X+4} \to \infty$, and $Nh_{NP}^{2d_z} \to \infty$, to conclude that it is $o_p(1)$. As before, we start analyzing the sum consisting of terms with different indices by symmetrizing it first. Then using the definition of ε^{h_0}, ε^P, iterated law of expectations and the independence of observations from one another, one could show that the expectation of each term in this sum is 0. Moreover, since both $M_{pN}^h(P_i, X_i)$ and $M_{pN}^P(X_i, Z_i)$ are nonsingular for large N, $\partial q / \partial h_1$ and $\partial h_0 / \partial P$ are continuous, K^h and K^P are 0 outside a compact set, and $var(Y) < \infty$, and $Nh_{Nh_0}^{2(d_X+2)} \to \infty$ and $Nh_{NP}^{2d_z} \to \infty$, the second moment of each term of the symmetrized sum is

$o(N)$. Therefore by the Hoeffding, Powell, Stock and Stoker lemma,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i \frac{\partial \hat{q}}{\partial h_1}(\tilde{h}_{0i}, \tilde{P}_i)\left[\hat{h}_0(X_i, \hat{P}(Z_i)) - h_0(X_i, P(Z_i))\right]\hat{I}_{1i}\hat{I}_{2i} \stackrel{AE}{=}$$
$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} E\left[D_j \frac{\partial q}{\partial h_1}(h_{0j}, P_j) I_{2j}\psi_{Nh_0}\big(-(1-D_i)Y_i, P(Z_i), X_i; X_j, P(Z_j)\big)|Y_i, D_i, X_i, Z_i\right] +$$
$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} E\left[D_j \frac{\partial q}{\partial h_1}(h_{0j}, P_j) I_{2j}\frac{\partial h_0}{\partial P}(P(Z_j), X_j)\psi_{NP}\big(D_i, Z_i; X_j, Z_j\big)|Y_i, D_i, X_i, Z_i\right] + b_{qh_0P}.$$

B.3 Step 3:

By Appendix C.3, we know that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\big(\hat{q}(h_{0i}, P_i) - q(h_{0i}, P_i)\big)\hat{I}_{1i}\hat{I}_{2i} \stackrel{AE}{=}$$
$$\frac{1}{\sqrt{N}N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{D_i}{P(Z_i)}[M_{pN}^q(h_{0i}, P_i)]^{-1}\left[\left(\frac{(h_{1j}, P_j) - (h_{0i}, P_i)}{h_{Nq}}\right)^{Q_p}\right]' \frac{1}{h_{Nq}^2}K^q\left(\frac{(h_{1j}, P_j) - (h_{0i}, P_i)}{h_{Nq}}\right)\varepsilon_j^q I_{1j} I_{1i} I_{2i} + b_q,$$

where $\varepsilon_j^q = D_j Y_j - E\big[D_j Y_j | h_1\big(X_j, P(Z_j)\big), P(Z_j)\big]$.

As in the previous two steps, we break the summation in the above expression into two pieces: one containing terms with $i = j$, the other containing terms with $i \neq j$. We apply the strong law of large numbers to the first sum, and then use $Nh_{Nq}^4 \to \infty$ and the fact that the expectation of the remaining part of a typical term in this sum is finite to argue that the whole sum is $o_p(1)$. For the sum containing different indices, we use the the Hoeffding, Powell, Stock and Stoker lemma. By going through arguments that are almost identical to those in the previous two steps, we can show that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} D_i\big(\hat{q}(h_{0i}, P_i) - q(h_{0i}, P_i)\big) \stackrel{AE}{=} b_q + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} E\Bigg\{\frac{D_j}{P(Z_j)}e_1[M_{pN}^q(h_{0j}, P_j)]^{-1} I_{1j} I_{2j} I_{1i}$$
$$\times\left[\left(\frac{(h_{1i}, P_i) - (h_{0j}, P_j)}{h_{Nq}}\right)^{Q_p}\right]' \frac{\varepsilon_i^q}{h_{Nq}^2}K^q\left(\frac{(h_{1i}, P_i) - (h_{0j}, P_j)}{h_{Nq}}\right)|D_i, Y_i, X_i, Z_i\Bigg\}.$$

B.4 Step 4:

Here we study the numerator of the first term of equation (10),

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big[(1-D_i)Y_i + D_i q(h_{0i}, P_i) - E(Y_0|A_1 \cap A_2)\Big]\hat{I}_{1i}\hat{I}_{2i}.$$

Let $A := A_1 \cap A_2$ and

$$\delta_A(X_i, Z_i) := (1-D_i)Y_i + D_i q(h_{0i}, P_i) - E(Y_0|A_1 \cap A_2)$$

For $\tilde{I}_1 \in \mathcal{I}_1$, $\tilde{I}_2 \in \mathcal{I}_2$ such that $\tilde{I}_{1i} \neq I_{1i}$ or $\tilde{I}_{2i} \neq I_{2i}$ with positive probability, $E[\delta_A(X_i, Z_i)\tilde{I}_{1i}\tilde{I}_{2i}] \neq 0$. But with probability one $\hat{I}_{1i}\hat{I}_{2i}$ equals

$$I_{1i}\hat{I}_{2i} + [\hat{\sigma}_1(X_i, Z_i)]^{-1}\tilde{J}_-\left(\frac{f_{X,Z}(X_i,Z_i)-q_{01}}{\hat{\sigma}_1(X_i,Z_i)}\right)1\{\hat{f}(X_i, Z_i) > f_{X,Z}(X_i, Z_i)\}\left[\hat{f}(X_i, Z_i) - f_{X,Z}(X_i, Z_i)\right]\hat{I}_{2i}$$
$$+[\hat{\sigma}_1(X_i, Z_i)]^{-1}\tilde{J}_+\left(\frac{f_{X,Z}(X_i,Z_i)-q_{01}}{\hat{\sigma}_1(X_i,Z_i)}\right)1\{\hat{f}(X_i, Z_i)) < f_{X,Z}(X_i, Z_i)\}\left[\hat{f}(X_i, Z_i) - f_{X,Z}(X_i, Z_i)\right]\hat{I}_{2i}$$

where $\tilde{J}_-(u) = 1\{-1 \leq u < 0\}$, $\tilde{J}_+(u) = 1\{0 \leq u < 1\}$, and $\hat{\sigma}_1(X_i, Z_i) := \left|\hat{f}(X_i, Z_i) - f_{X,Z}(X_i, Z_i)\right|$. Similarly, for $f \in \mathcal{H}_1$, define, $\tilde{\sigma}_1(X_i, Z_i) := |f(X_i, Z_i) - f_{X,Z}(X_i, Z_i)|$, $\tilde{L}_{1i} = 1\{f(X_i, Z_i) > f_{X,Z}(X_i, Z_i)\}$. Then for $\tilde{I}_2 \in \mathcal{I}_2$,

$$N^{-3/2}\sum_{i=1}^N\sum_{j=1}^N \delta_A(X_i, Z_i)\tilde{I}_{2i}\tilde{L}_{1i}[\tilde{\sigma}_1(X_i, Z_i)]^{-1}\tilde{J}_-\left(\frac{f_{X,Z}(X_i,Z_i)-q_{01}}{\tilde{\sigma}_1(X_i,Z_i)}\right)$$
$$\times\left(\frac{1}{\tilde{h}_{N1}^d}\tilde{K}_1\left(\frac{(X_j,Z_j)-(X_i,Z_i)}{\tilde{h}_{N1}}\right) - E\left[\frac{1}{\tilde{h}_{N1}^d}\tilde{K}_1\left(\frac{(X_j,Z_j)-(X_i,Z_i)}{\tilde{h}_{N1}}\right)|X_i, Z_i\right]\right)$$
$$+N^{-3/2}\sum_{i=1}^N\sum_{j=1}^N \delta_A(X_i,Z_i)\tilde{I}_{2i}\tilde{L}_{1i}[\tilde{\sigma}_1(X_i,Z_i)]^{-1}\tilde{J}_- \quad \frac{f_{X,Z}(X_i,Z_i)-q_{01}}{\tilde{\sigma}_1(X_i,Z_i)} \quad E\ \frac{1}{\tilde{h}_{N1}^d}\tilde{K}_1\ \frac{(X_j,Z_j)-(X_i,Z_i)}{\tilde{h}_{N1}}\ |X_i,Z_i\ -f_{X,Z}(X_i,Z_i)$$

The first of these is an order one degenrate U-process which satisfies the conditions of the equicontinuity lemma. Therefore the first term is $o_p(1)$ for each element of the family of the functions we consider. As for the second term, using the rates of convergence in Silverman's article we can show that this term goes to 0 as well. On the other hand, the analysis of the term involving $\tilde{J}_+$ is symmetric. The last step in this section is to repeat these arguments for

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \delta_A(X_i, Z_i)I_{1i}[\hat{I}_{2i} - I_{2i}]$$

and conclude that

$$\frac{\sum_{i=1}^N\left[(1-D_i)Y_i+D_iq(h_{0i},P_i)-E(Y_0|A_1\cap A_2)\right]\hat{I}_{1i}\hat{I}_{2i}}{\sqrt{N}} \overset{AE}{=} \frac{\sum_{i=1}^N\left[(1-D_i)Y_i+D_iq(h_{0i},P_i)-E(Y_0|A_1\cap A_2)\right]I_{1i}I_{2i}}{\sqrt{N}}.$$

B.5 Step 5:

$$\frac{1}{N}\sum_{i=1}^N \hat{I}_{1i}\hat{I}_{2i} = \frac{1}{N}\sum_{i=1}^N I_{1i}I_{2i} + \frac{1}{N}\sum_{i=1}^N \hat{I}_{1i}[\hat{I}_{2i} - I_{2i}] + \frac{1}{N}\sum_{i=1}^N I_{2i}[\hat{I}_{1i} - I_{1i}]$$

By the law of large numbers, the first term on the right hand side converges to $P(A_1 \cap A_2)$. Now consider the second term, and note that $N^{-1}|\sum_{i=1}^N \hat{I}_{1i}[\hat{I}_{2i} - I_{2i}]| \leq N^{-1}\sum_{i=1}^N |\hat{I}_{2i} - I_{2i}|$. Our

trimming assumptions guarantee that $E|\hat{I}_{2i} - I_{2i}|$ approaches 0 as N tends to infinity. Therefore, for each fixed $\kappa > 0$,

$$P\left(N^{-1}\left|\sum_{i=1}^{N}\hat{I}_{1i}[\hat{I}_{2i} - I_{2i}]\right| > \kappa\right) \leq P\left(N^{-1}\sum_{i=1}^{N}|\hat{I}_{2i} - I_{2i}| > \kappa\right) \leq \frac{E|\hat{I}_{2i} - I_{2i}|}{\kappa} \to 0$$

and thus, the second term is $o_p(1)$.By an analogous argument, we can show that the last term is also $o_p(1)$.