Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects.*

J.P. Florens[†], J. J. Heckman[‡], C. Meghir[§] and E. Vytlacil[¶]

This draft, April 4, 2007

Abstract

We use the control function approach to identify the average treatment effect and treatment on the treated in models with a continuous endogenous regressor whose impact is heterogeneous. We assume a stochastic polynomial restriction on the form of the heterogeneity but, unlike alternative nonparametric control function approaches, ours does not require large support assumptions.

*We thank two anonymous referees and Whitney Newey for their comments. We also thank participants at the Berkeley-Stanford (March 2001) workshop on non-parametric models with endogenous regressors as well as participants at the UCL empirical microeconomics workshop for useful comments. J. Heckman would like to thank NIH R01-HD043411 and NSF SES-024158 for research support. C. Meghir would like to thank the Centre for Economics of Education and the ESRC for funding through the Centre for Fiscal policy at the IFS. E. Vytlacil would like to thank the NSF for financial support (NSF SES-05-51089). The views expressed in this paper are those of the authors and not necessarily those of the funders listed here. All errors are our own.

[†]IDEI, Toulouse

[‡]University of Chicago and University College Dublin

[§]IFS and UCL, c.meghir@ucl.ac.uk

[¶]Columbia University

1 Introduction

There is a growing theoretical and empirical literature on models where the impacts of discrete (usually binary) treatments are heterogeneous in the population.[1] The objective of this paper is to analyze non-parametric identification of treatment effect models with continuous treatments when the treatment intensity is not randomly assigned. This generally leads to models that are non-separable in the unobservables and produces heterogeneous treatment intensity effects. We impose a stochastic polynomial assumption on the heterogeneous effects, and exploit this restriction to use a control function approach to obtain identification without large support assumptions. Our approach has applications in a wide variety of problems, including demand analysis where price elasticities may differ across individuals, labor supply, where wage effects may be heterogeneous; or production functions, where the technology may vary across firms, at least in the short run.

Other recent papers on semiparametric and nonparametric models with nonseparable error terms and an endogenous, possibly continuous covariate include papers using quantile instrumental variable methods such as Chernozhukov and Hansen (2005) and Chernozhukov, Imbens, and Newey (2006), and papers using a control variate technique such as Altonji and Matzkin (2005), Blundell and Powell (2004), Chesher (2003), and Imbens and Newey (2002). See Chesher (2007) for a recent survey. The analysis of Imbens and Newey (2002) is perhaps the most relevant to our analysis, with the key distinction between our approach and their approach being a tradeoff between imposing a stochastic polynomial assumption on the outcome equation versus requiring a large support assumption. We further discuss the differences between our approach and their approach in Section 3.2.

2 The Model, Parameters of Interest and the Observables.

Let Y_d denote the potential outcome corresponding to level of treatment intensity d. Define $\varphi(d) = E(Y_d)$ and $U_d = Y_d - \varphi(d)$, so that, by construction,

$$Y_d = \varphi(d) + U_d. \tag{1}$$

[1] See, e.g., Roy, 1951; Heckman and Robb, 1985, 1986; Björklund and Moffitt, 1987; Imbens and Angrist, 1994; Heckman, 1997; Heckman, Smith, and Clements, 1997; Heckman and Honoré, 1990; Card, 1999, 2001; Heckman and Vytlacil, 2001, 2005, 2007a,b, who discuss heterogeneous response models.

Let D denote the realized treatment, so that the realized outcome Y is given by $Y = Y_D$. We will not explicitly consider observable regressors that directly affect Y_d. All of our analysis will be implicitly conditional on any such regressors. We assume

(A-1) *$\varphi(D)$ is K times differentiable in D (a.s.), and the support of D does not contain any isolated points (a.s.).*

We restrict attention to the case where the stochastic process U_d takes the form

$$U_d = \sum_{j=0}^{K} d^j \varepsilon_j, \qquad \text{with } \; E(\epsilon_j) = 0, \quad j = 0, ..., K, \tag{2}$$

allowing for heterogeneity of a finite set of derivatives of Y_d. This specification can be seen as a nonparametric, higher order generalization of the random coefficient model considered by Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2007). The normalization $E(\epsilon_j) = 0$, $j = 0, ..., K$, implies that $\frac{\partial^j}{\partial d^j} E(Y_j) = \frac{\partial^j}{\partial d^j}\varphi(d)$.[2]

Equations (1) and (2) can be restated as follows to emphasize that we are considering a nonseparable model:

$$Y = h(D, \epsilon) = \varphi(D) + \sum_{j=0}^{K} D^j h_j(\epsilon) \tag{3}$$

where ϵ need not be a scalar random variable. The notation of equation (3) can be mapped into the notation of equations (1) and (2) by setting $\varepsilon_j = h_j(\epsilon)$. Notice that we are not imposing that ϵ is a scalar random variable, and h need not be monotonic in ϵ.

One parameter of interest in this paper is the Average Treatment Effect,[3]

$$\Delta^{ATE}(d) = \lim_{\Delta d \to 0} \frac{E(Y_{d+\Delta d} - Y_d)}{\Delta d} \equiv \frac{\partial}{\partial d} E(Y_d) = \frac{\partial}{\partial d}\varphi(d) \tag{4}$$

which is the average effect of a marginal increase in the treatment if individuals were randomly assigned base treatment level d. Note that the average treatment effect depends on the base treatment level, and for any of the continuum of possible base treatment levels we have a different

[2]To see that $E(\epsilon_j) = 0$, $j = 0, ..., K$, is only a normalization, note that $\varphi(d) + \sum_{j=0}^{K} d^j \varepsilon_j = \left[\varphi(d) + \sum_{j=0}^{K} d^j E(\varepsilon_j)\right] + \sum_{j=0}^{K} d^j (\varepsilon_j - E(\varepsilon_j)) = \widetilde{\varphi}(d) + \sum_{j=0}^{K} d^j \widetilde{\varepsilon_j}$. Note that this is the appropriate normalization for $\frac{\partial}{\partial d}\varphi$ to denote the ATE.

[3]Blundell and Powell (2004) define the Average Structural Function (ASF) as $\mu(d) = E(Y_d)$. The average treatment effect is the derivative of the average structural function.

average treatment effect. We also consider the effect of treatment on the treated (TT), given by

$$\Delta^{TT}(d) = \lim_{\Delta d \to 0} \frac{E(Y_{d+\Delta d} - Y_d | D = d)}{\Delta d}$$
$$\equiv E(\frac{\partial}{\partial d_1} Y_{d_1} | D = d_2)\Big|_{d=d_1=d_2} = \frac{\partial}{\partial d}\varphi(d) + \sum_{j=1}^{K} j d^{j-1} E(\varepsilon_j | D = d)$$

which is the average effect among those currently choosing treatment level d of an incremental increase in the treatment while leaving their unobservables fixed.

We define the choice equation (the assignment mechanism to treatment intensity) as

$$D = g(Z, V) \tag{5}$$

where Z are observed covariates that enter the treatment choice equation but are excluded from the Y_d outcome equations. We make the following assumption

(A-2) *V is a scalar unobservable term whose distribution is absolutely continuous with respect to Lebesgue measure; g is strictly monotonically increasing in V and $Z \perp\!\!\!\perp (V, \varepsilon_0, ...\varepsilon_K)$.*

As long as D is a continuous random variable (conditional on Z), we can always represent D as a function of Z and a continuous scalar error term, with the function increasing in the scalar error term and the scalar error term independent of Z. Setting $V = F_{D|Z}(D|Z)$ and $g(Z, V) = F^{-1}_{D|Z}(V|Z)$, we have $D = g(Z, V)$ with g strictly increasing in the scalar V which is distributed unit uniform and independent of Z. However, the assumption that $g(Z, V)$ is monotonic in a scalar unobservable V with $Z \perp\!\!\!\perp (V, \varepsilon_0, ...\varepsilon_K)$ is restrictive. The construction $V = F(D|Z)$ and $D = F^{-1}_{D|Z}(V|Z) = g(Z, V)$ does not guarantee $Z \perp\!\!\!\perp (V, \varepsilon_0, ...\varepsilon_K)$.

Given assignment mechanism (5) and assumption (A-2), without loss of generality, we can impose the normalization that V is distributed unit uniform. Given these assumptions and the normalization that V is distributed unit uniform, we can recover V from $V = F(D|Z)$ and the function g from $g(Z, V) = F^{-1}_{D|Z}(V|Z)$.[4] Assignment mechanism (5) and assumption (A-2) will not be directly used to prove identification. However, we will use it to define the primitive conditions underlying the identification assumptions we make.

[4]Assignment mechanism (5) and assumption (A-2) are the same restrictions imposed by Imbens and Newey (2002) on the first stage equation for the endogenous regressor. We discuss the relationship of our approach and their approach in Section 3.2.

2.1 Education and Wages: A Simple Illustration

To illustrate the type of problem we analyze in this paper, consider a simple model of educational choice. Suppose that the agent receives wages Y_d at direct cost C_d if schooling choice d is made. We work with discounted annualized earnings flows. We write wages for schooling level d as Y_d, as well as the cost function for schooling as

$$\begin{aligned} Y_d &= \varphi_0 + (\varphi_1 + \varepsilon_1)d + \tfrac{1}{2}\varphi_2 d^2 + \varepsilon_0 \qquad I \\ C_d &= C_0(Z) + (C_1(Z) + v_1)d + \tfrac{1}{2}C_2(Z)d^2 + v_0 \quad II \end{aligned} \tag{6}$$

where ε_s and v_s $(s = 0, 1)$ are, respectively, unobserved heterogeneity in the wage level and in the return to schooling. These unobserved heterogeneity terms are the source of the identification problem. We impose the normalizations that $E(\varepsilon_s) = 0$, $E(v_s) = 0$, for $s = 0, 1$. We implicitly condition on variables such as human capital characteristics that affect both wages and the costs of schooling. Z are factors affecting the cost of schooling only, such as the price of education.

Assume that agents chose their level of education to maximize wages minus costs. Let D denote the resulting optimal choice of education. D solves the first order condition

$$(\varphi_1 - C_1(Z)) + (\varphi_2 - C_2(Z))\, D + \varepsilon_1 - v_1 = 0.$$

Assuming that $\varphi_2 - C_2(Z) < 0$ for all Z, the second order condition will be satisfied. This leads to an education choice equation (assignment to treatment intensity rule)

$$D = \frac{\varphi_1 - C_1(Z) + \varepsilon_1 - v_1}{C_2(Z) - \varphi_2}.$$

This example is a special case of the model given by equations (1), (2) and (5), with

$$\begin{aligned} \varphi(d) &= \varphi_0 + \varphi_1 d + \frac{1}{2}\varphi_2 d^2 \\ U_d &= \epsilon_0 + \epsilon_1 d \\ g(z, v) &= \frac{\varphi_1 - C_1(z) + F^{-1}_{\epsilon_1 - v_1}(v)}{C_2(z) - \varphi_2} \end{aligned}$$

where $V = F_{\varepsilon_1 - v_1}(\varepsilon_1 - v_1)$ with $F_{\varepsilon_1 - v_1}$ the cumulative distribution function of $\varepsilon_1 - v_1$. The object of the empirical analysis is to estimate the average return to education: $\Delta^{ATE}(d) = \varphi_1 + \varphi_2 d$, or TT, which here is give by $\Delta^{TT}(d) = (\varphi_1 + E(\epsilon_1 | D = d)) + \varphi_2 d$.

In this example, the treatment intensity is given by equation (5) with g strictly increasing in a scalar error term $V = F_{\varepsilon_1 - v_1}(\varepsilon_1 - v_1)$. The structure of the treatment intensity mechanism is sensitive to alternative specifications. Consider the same example as before, except now the second derivative of Y_d is also stochastic: $Y_d = \varphi_0 + (\varphi_1 + \varepsilon_1)d + \frac{1}{2}(\varphi_2 + \varepsilon_2)d^2 + \varepsilon_0$. The selection model becomes $D = \frac{\varphi_1 - C_1(Z) + \varepsilon_1 - v_1}{C_2(Z) - \varphi_2 - \varepsilon_2}$. In this case, the structural model makes D a function of $V = (\varepsilon_1 - v_1, \varepsilon_2)$, which satisfies $Z \perp\!\!\!\perp (V, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ but V is not a scalar error. We can still define $\tilde{V} = F(D|Z)$ and the function $\tilde{g}$ by $\tilde{g}(Z, \tilde{V}) = F^{-1}_{D|Z}(\tilde{V}|Z)$. With this construction, D is strictly increasing in a scalar error term $\tilde{V}$ that is independent of Z. With this construction, Z is not independent of $(\tilde{V}, \varepsilon_0, \varepsilon_1, \varepsilon_2)$.[5] This is an illustration of a case where assumption (A-2) does not hold. The fragility of the specification of equation (5) where g is strictly increasing in a scalar error term arises in part because, under rational behavior, heterogeneity in response to treatment (heterogeneity in the Y_d model) generates heterogeneity in selection into treatment intensity. This link can be broken if agents do not know their own treatment effect heterogeneity.

3 Identification Analysis.

In the case of binary treatment with heterogeneous impacts IV does not identify ATE (Heckman and Robb, 1986). The same is true in the models with heterogeneous treatments unless one imposes covariance restrictions between the errors in the assignment rule and the errors in the structural model. Following Newey and Powell (2003) and Darolles, Florens, and Renault (2002) consider a nonparametric IV strategy based on the identifying assumption that $E(Y - \varphi(D)|Z) = 0$. Suppose $K = 0$, which is the special case of no treatment effect heterogeneity. In this case, $Y_D = \varphi(D) + U_0$. We obtain the standard additive-in-unobservables model considered in the cited papers. The identification condition is $E(\epsilon_0|Z) = 0$. However, in the general case of treatment effect heterogeneity ($K > 0$), the IV identification restriction implies special covariance restrictions between the error terms. For example, suppose $K = 1$ and that $D = g(Z) + V$. Then $E(Y - \varphi(D)|Z) = 0$ requires $E(\epsilon_0|Z) = 0$ and $E(\epsilon_1 D|Z) = 0$, with the latter restriction generically equivalent to $E(\epsilon_1|Z) = 0$ and $E(\epsilon_1 V|Z) = 0$. In other words, in addition to the more standard type of condition that ϵ_0 be mean independent of the instrument, we now have a new restriction in the heterogeneous case that the covariance between the heterogeneous effect and the unobservables in the choice equation conditional on the instrument does not depend on the instrument.[6] Instead of following an instrumental variables approach, we explore

[5]Note that $\Pr(\tilde{V} \leq v|Z, \varepsilon_0, \varepsilon_1, \varepsilon_2) = \Pr\left[v_1 : \frac{\varphi_1 - C_1(Z) + \varepsilon_1 - v_1}{C_2(Z) - \varphi_2 - \varepsilon_2} \leq F^{-1}_{D|Z}(v)|Z, \varepsilon_0, \varepsilon_1, \varepsilon_2\right] \neq \Pr(\tilde{V} \leq v|\varepsilon_0, \varepsilon_1, \varepsilon_2)$.

[6]See Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2007).

identification through a control function.[7] We assume the existence of a (known or identifiable) control function $\tilde{V}$ that satisfies the following conditions:

(A-3) *Control Function Condition:* $E(\varepsilon_j \mid D, Z) = E(\varepsilon_j|\tilde{V}) = r_j(\tilde{V})$.[8]

and

(A-4) *Rank condition: D and $\tilde{V}$ are measurably separated, i.e., any function of D almost surely equal to a function of $\tilde{V}$ must be almost surely equal to a constant.*

A necessary condition for assumption (A-4) to hold is that the instruments Z have an impact on D.[9] We return later in this section to consider sufficient conditions on the underlying model that implies the existence of such a control variate $\tilde{V}$. Under these assumptions ATE and TT are identified.

Theorem 1. *Assume equations (1) and (2) hold with finite $K \geq 1$. Under assumptions (A-3) (control function condition), (A-4) (rank condition), and the smoothness and support condition (A-1),* ATE *and* TT *are identified.*

Proof. See Appendix. □

The control function assumption gives the basis for an empirical determination of the relevant degree of the polynomial in (2). If the true model is defined by a degree ℓ we have that for any $k > \ell$

$$\frac{\partial^k}{\partial d^k} E(Y|D = d, \tilde{V} = v) = \frac{\partial^k \varphi(d)}{\partial d^k}$$

which does not depend on v and thus is only a function of d. This property can be verified by testing whether the following equality holds almost surely:

$$\frac{\partial^k E(Y|D, \tilde{V})}{\partial D^k} \overset{a.s.}{=} E\left[\frac{\partial^k}{\partial D^k} E(Y|D, \tilde{V}) \middle| D\right], \qquad k > \ell.$$

[7] See Newey, Powell, and Vella (1999) for a control function approach for the case of separable models ($K = 0$).

[8] Note that our normalization $E(\epsilon_j) = 0, \quad j = 0, ..., K$, implies the normalization that $E(r_j(\tilde{V})) = 0, \quad j = 0, ..., K$,

[9] Measurable separability, which we maintain in this paper is just one way of achieving this. Alternatively, one could restrict the space of functions $\varphi(D)$ not to contain $r_j(\tilde{V})$ functions; this in turn can be achieved for example by assuming that $\varphi(D)$ is linear in D and r_j is non-linear as in the Heckman (1979) selection model. See also Heckman and Robb (1985, 1986) who discuss this condition.

3.1 Primitive Conditions for Control Function Assumption.

In the previous section the control function is assumed to be given, and it is assumed that the control function has the desired properties. The previous section did not use assignment rule (5) or condition (A-2). In this section, we use assignment rule (5) and condition (A-2) along with the normalization that V is distributed unit uniform. Under these conditions consider using $V = F_{D|Z}(D|Z)$ as the control function. This leads to the following corollary to Theorem 1:

Corollary 3.1. *Assume equations (1) and (2) hold with finite* $K \geq 1$ *and assume smoothness and support condition (A-1). If* D *satisfies assignment equation (5) and condition (A-2), and if* V *and* D *satisfy measurable separability (A-4), then* ATE *and* TT *are identified.*

In order for the conditions of Theorem 1 to be satisfied it is sufficient to verify that under the conditions in the corollary the control function assumption (A-3) is satisfied. This is deduced from the following argument. $Z \perp\!\!\!\perp (V, \varepsilon_0, ..., \varepsilon_K)$ implies $Z \perp\!\!\!\perp (\varepsilon_0, ..., \varepsilon_K)|V$, and hence $(Z, V) \perp\!\!\!\perp (\varepsilon_0, ..., \varepsilon_K)|V$ which implies that $D \perp\!\!\!\perp (\varepsilon_0, ..., \varepsilon_K)|V$, which implies the control function assumption.

Next consider the measurable separability condition (A-4). Measurable separability is a relatively weak condition, as illustrated by the following theorem.

Theorem 2. *Assume that* (D, V) *has a density with respect to Lebesgue measure in* R^2 *and denote by* S *its support and by* S^0 *the interior of the support. Further assume that i) any point in* S^0 *has a neighborhood such that the density is strictly positive within it and ii) any two points within* S^0 *can be connected by a continuous curve that lies strictly in* S^0*. Then measurable separability between* D *and* V *(A-4) holds.*

Proof. See Appendix. □

Measurable separability is a type of rank condition. To see this, consider the following heuristic argument. Consider a case where the condition is violated at some point in the interior of the support of (D, V), i.e. $h(D) = l(V)$. Hence $h(g(Z, V)) = l(V)$. Differentiating both sides of this expression with respect to Z, we obtain $\frac{\partial h}{\partial g}\frac{\partial g}{\partial Z} = 0$. If measurable separability fails, $\frac{\partial h}{\partial g} \neq 0$ and hence $\frac{\partial g}{\partial Z} = 0$ which means that g does not vary with Z. Note that the conditions in Theorem 2 are not very restrictive. For example, the conditional support of D can depend on V and vice versa.

Assignment rule (5) and condition (A-2) do not imply measurable separability (A-4). We now consider two examples where equation (5) and condition (A-2) hold but D and V are not measurably separable. In the first example, Z is a discrete random variable. In the second example, $g(z, v)$ is a discontinuous function of v.

First, suppose $Z = 0, 1$ and suppose that $D = g(z, v) = z + v$, with V uniform $(0, 1)$. Then (A-4) fails, i.e., D and V are not measurably separable. To see this, let $m_1(t) = t$ and let $m_2(t) = \mathbf{1}[t \leq 1]t + \mathbf{1}[t > 1](t - 1)$. Then $m_1(V) = m_2(D)$, but m_1 and m_2 are not a.s. equal to a constant. Now consider a second example. Suppose that $D = g_1(z) + g_2(v)$, where $g_2(t) = \mathbf{1}[t \leq .5]t + \mathbf{1}[t > .5](1 + t)$. Let g_1^{max} and g_1^{min} denote the maximum and minimum of the support of the distribution of $g_1(Z)$, and suppose that $g_1^{max} - g_1^{min} < 1$. Then (A-4) fails, i.e., D and V are not measurably separable. To see this, let $m_1(t) = \mathbf{1}[t \leq .5]$, let $m_2(t) = \mathbf{1}[t \leq .5 + g_1^{max}]$, and note that $m_1(V) = m_2(D)$ but that m_1 and m_2 do not (a.s.) equal a constant.

Assignment rule (5), condition (A-2), and regularity conditions that require Z to contain a continuous element and that g be continuous in v are sufficient to imply that measurable separability (A-4) holds. We prove the following theorem.

Theorem 3. *Suppose that D is determined by equation (5). Suppose that $g(z, v)$ is a continuous function of v. Suppose that, for any fixed v, the support of the distribution of $g(Z, v)$ contains an open interval. Then, under assumption (A-2), D and V are measurably separated ((A-4) holds).*

Proof. See Appendix. □

Note that, for any fixed v, for the support of the distribution of $g(Z, v)$ to contain an open interval requires that Z contains a continuous element. A sufficient condition for the support of the distribution of $g(Z, v)$ to contain an open interval is that (a) Z contains an element whose distribution conditional on the other elements of Z contains an open interval, and (b) g is a continuous monotonic function of that element. Thus, under the conditions of Theorem 3, V is identified by $V = F(D|Z)$ and both the control function condition (A-3) and the rank condition (A-4) hold with $\tilde{V} = V$.

3.2 Alternative Identification Analysis.

A general analysis of identification using the control function assumption without the polynomial structure is related to the work of Heckman and Vytlacil (2001) on the marginal treatment effect (MTE). That paper considers a binary treatment model, but their analysis may be extended to the continuous treatment case. A similar approach to identification of the "Average Structural Function" in semiparametric models with a continuous treatment D that is strictly monotonic in V is pursued by Blundell and Powell (2004) and Altonji and Matzkin (2005).

Most relevant to this note is the analysis of Imbens and Newey (2002), who follow this approach for a nonparametric model. They impose the same structure as we do on the first stage equation for the endogenous regressor as our assignment mechanism (5) and they invoke assumption (A-2). Thus, the control variate is again V, with V identified and with a distribution that can be normalized to be unit uniform. We again have $E(Y_d|D=d,V=v) \overset{as}{=} E(Y_d|V=v)$. Furthermore, assume that the support of (D,V) is the product of the support of the two marginal distributions, i.e., assume rectangular support. This assumption implies that the conditional support of D given V does not depend on V (and vice versa). It is stronger than the measurable separability assumption we previously used to establish identification. From these assumptions it follows that

$$E(Y|D=d,V=v) = E(Y_d|D=d,V=v)$$
$$= E(Y_d|V=v)$$

and

$$E(Y_d) = \int E(Y_d|V=v)dF(v).$$

Then $\varphi(d) = \int E(Y|D=d,\, V=v)dF(v)$ and is identified. Identification of $\varphi(d)$ in turn implies identification of $\Delta^{ATE} = \frac{\partial}{\partial d}\varphi(d)$. The rectangular support condition is needed to replace $E(Y_d|V=v)$ by $E(Y|D=d,\, V=v)$ for all v in the unconditional support of V in the previous integral. The rectangular support condition may not be satisfied and in general requires a large support assumption as illustrated by the following example. Suppose $D = g_1(Z)+V$. Let $\mathcal{G}_1$ denote the support (Supp) of the distribution of $g_1(Z)$. If Z and V are independent, then

$$\text{Supp}(V|D=d) = \text{Supp}(V|g_1(Z)+V=d) = \text{Supp}(V|V=d-g_1(Z)) = \{d-g : g \in \mathcal{G}_1\}$$

where the last equality uses the condition $Z \perp\!\!\!\perp V$. $\{d-g : g\in\mathcal{G}_1\}$ does not depend on d if and only if $\mathcal{G}_1 = \Re$.[10]

[10] For example, if $\mathcal{G}_1 = [a,b]$, then $\{d-g : g\in\mathcal{G}_1\} = \{d-g \,:\, g\in[a,b]\} = [d-b, d-a]$ which does not depend

Instead of imposing $E(Y_d|D=d, V=v) = E(Y_d|V=v)$, one could instead impose

$$\begin{aligned}\frac{\partial}{\partial d}E(Y|D=d, V=v) &= E(\tfrac{\partial}{\partial d}Y_d|D=d, V=v) \\ &= E(\tfrac{\partial}{\partial d}Y_d|V=v).\end{aligned} \tag{7}$$

$E(\frac{\partial}{\partial d}Y_d|V=v)$ is the marginal treatment effect of Heckman and Vytlacil (2001), adapted to the case of a continuous treatment. Instead of integrating $E(Y_d|V=v)$ to obtain $\varphi(d)$, one could instead integrate $E(\frac{\partial}{\partial d}Y_d|V=v)$ to obtain ATE or TT:

$$\int \tfrac{\partial}{\partial d}E(Y|D=d, V=v)dF(v) = \int \tfrac{\partial}{\partial d}E(Y_d|V=v)dF(v) = \Delta^{ATE}(d),$$

$$\begin{aligned}\int \tfrac{\partial}{\partial d_1}E(Y_{d_1}|D=d_2, V=v)dF(v|D=d_2)\Big|_{d=d_1=d_2} &= E\left(\tfrac{\partial}{\partial d_1}Y_{d_1}|D=d_2\right)\Big|_{d=d_1=d_2} \\ &= \tfrac{\partial}{\partial d_1}E\left(Y_{d_1}|D=d_2\right)\Big|_{d=d_1=d_2} = \Delta^{TT}(d).\end{aligned}$$

This is the identification strategy followed in Heckman and Vytlacil (2001), adapted to the case where D is a continuous treatment. As discussed in Heckman and Vytlacil (2001), we again have the rectangular support condition as a requirement to integrate up MTE to obtain ATE. Note that we do not require the rectangular support condition to integrate up $\frac{\partial}{\partial d}E(Y|D=d, V=v)$ to obtain TT. For TT, we only need to evaluate $\frac{\partial}{\partial d}E(Y|D=d, V=v)$ for v in the support of V conditional on $D=d$, not in the unconditional support of V.

While we do not require the rectangular support condition to integrate MTE to recover TT, we do require a support condition for equation (7) to hold. That equation requires that $E(Y|D=d, V=v)$ can be differentiated with respect to d while keeping v fixed. This property is closely related to measurable separability between D and V. Assume that there exists a (differentiable) function of D, $h(D)$ equal ($a.s.$) to a function of V, $m(V)$, which is not constant. Then we obtain

$$E(Y|D=d, V=v) \stackrel{as}{=} E(Y_d|V=v) + h(d) - m(V)$$

and

$$\frac{\partial}{\partial d}E(Y|D=d, V=v) \stackrel{as}{=} \frac{\partial}{\partial d}E(Y_d|V=v) + \frac{\partial}{\partial d}h(d)$$

which implies that equation (7) is violated. Thus, for TT, we still need measurable separability between D and V in order for equation (7) to hold.

There are trade-offs between the approach of this note versus an approach that identifies MTE/MTE-like objects and then integrates them to obtain the object of interest. The approach

on d if and only if $a=-\infty$ and $b=\infty$, i.e., if and only if $\mathcal{G}_1 = \Re$.

of this note requires the stochastic polynomial structure on U_D of equation (5) and higher order differentiability. These conditions are not required by Imbens and Newey (2002) or Heckman and Vytlacil (2001). The approach of this note does not require the large support assumption required by these alternative approaches. As shown by Theorem 2, measurable separability between D and V is a relatively mild restriction on the support of (D, V). As shown by Theorem 3, measurable separability between D and V follows from assignment mechanism (5) and Assumption (A-2) combined with a relatively mild regularity condition.

4 Conclusions

This paper considers the identification and estimation of models with a continuous endogenous regressor and non-separable errors when continuous instruments are available. We have presented an identification result using a control function technique. Our analysis imposes a stochastic, finite-order polynomial restriction on the outcome model but does not impose a large support assumption.

5 Appendix: Proofs of Theorems

Proof of Theorem 1

Suppose that there are two sets of parameters $(\varphi^1, r_K^1, ..., r_0^1)$ and $(\varphi^2, r_K^2, ..., r_0^2)$ such that

$$E(Y|D = d, \tilde{V} = v) = \varphi^i(d) + \sum_{k=0}^{K} d^k r_k^i(v), \; i = 1, 2,$$

where the conditional expectation on the left-hand side takes this form as a result of the control function assumption (A-3). Then

$$\left[\varphi^1(d) - \varphi^2(d)\right] + \sum_{k=0}^{K} d^k \left[r_k^1(v) - r_k^2(v)\right] = 0. \tag{8}$$

Given smoothness assumption (A-1), this implies

$$\frac{\partial^K}{\partial d^K}\varphi^1(d) - \frac{\partial^K}{\partial d^K}\varphi^2(d) + (K!)(r_K^1(v) - r_K^2(v)) = 0.$$

Measurable separability assumption (A-4) implies that if any function of d is equal to a function of v (a.s.) then this must be a constant (a.s.). Hence, $r_K^1(v) - r_K^2(v)$ is a constant a.s.. Hence,

$$r_K^1(v) - r_K^2(v) = E\left[r_K^1(\tilde{V}) - r_K^2(\tilde{V})\right].$$

This expression equals zero given our normalization that $E(\epsilon_K) = 0$. Hence,

$$r_K^1(v) - r_K^2(v) \overset{a.s.}{=} 0.$$

Considering the $(K-1)^{\text{st}}$ derivative of equation (8), we find that

$$\frac{\partial^{K-1}}{\partial d^{K-1}}\varphi^1(d) - \frac{\partial^{K-1}}{\partial d^{K-1}}\varphi^2(d) + (K!)d\Big[r_K^1(v) - r_K^2(v)\Big] + ((K-1)!)\Big[r_{K-1}^1(v) - r_{K-1}^2(v)\Big] = 0.$$

We have already shown that $r_K^1(v) = r_K^2(v)$, and thus

$$\frac{\partial^{(K-1)}}{\partial d^{(K-1)}}\varphi^1(d) - \frac{\partial^{(K-1)}}{\partial d^{(K-1)}}\frac{\partial}{\partial d}\varphi^2(d) + ((K-1)!)(r_{K-1}^1(v) - r_{K-1}^2(v)) = 0.$$

Using the logic of the previous analysis, we can show that $r_{K-1}^1(v) - r_{K-1}^2(v) \overset{a.s.}{=} 0$. Iterating this procedure for $k = K-2$,...,0, it follows that $r_k^1(v) - r_k^2(v) \overset{a.s.}{=} 0$ for all $k = 0, ..., K$. Again appealing to equation (8), it follows that $\varphi^1(d) - \varphi^2(d) \overset{a.s.}{=} 0$, and thus ATE is identified. Using the fact that $\varphi^1(d) - \varphi^2(d) \overset{a.s.}{=} 0$ and $r_k^1(v) - r_k^2(v) \overset{a.s.}{=} 0$ for all $k = 0, ..., K$, we also have that $\frac{\partial}{\partial d}\varphi^1 + \sum_{k=1}^{K} kd^{k-1}E[r_k^1(v)|d] = \frac{\partial}{\partial d}\varphi^2 + \sum_{k=1}^{K} kd^{k-1}E[r_k^2(v)|d] = 0$, and thus TT is identified.■

Proof of Theorem 2

Let (d, v) be a point of the interior of the support S^0. Let N^d denote a neighborhood of d and N^v a neighborhood of v such that $N^d \times N^v$ is included in S^0. The distribution of (D, V) restricted to $N^d \times N^v$ is equivalent to Lebesgue measure (i.e. has the same null sets). Then using Theorem 5.2.7 (Chapter 5) of Florens, Mouchart, and Rolin (1990) (D, V) restricted to $N^d \times N^v$ are measurably separated. This implies that if within that neighborhood $h(D) \overset{as}{=} l(V)$, then $h(D)$ and $l(V)$ are $a.s.$ constants. We need to show that this is true everywhere in the interior of the support. Consider any two points (d, v) and (d', v') in S^0. The theorem will be true if $h(d) = h(d')$. As S^0 satisfies the property (ii) in the theorem and is open by definition, there exists a finite number of overlapping open sets with non-empty overlaps, i.e. $\exists$ a finite sequence of neighborhoods $N_j^d \times N_j^v$, $j = 1, ..., J$ such that each $N_j^d \times N_j^v \subset S^0$ and $N_j^d \cap N_{j+1}^d \neq \emptyset$ and similarly for N_j^v. The first point (d, v) is in $N_1^d \times N_1^v$ and the second point (d', v') is in $N_J^d \times N_J^v$. Take $d_1 \in N_1^d$ and in the next overlapping neighborhood $d_2 \in N_2^d$. From the previous result (D, V) are measurably separated on $N_1^d \times N_1^v$ and on $N_2^d \times N_2^v$. Thus $h(d_i)$ $i = 1, 2$ is constant on each and thus constant on the union implying $h(d) = h(d_2)$. Iterating in this way along the sequence of neighborhoods until $N_J^d \times N_J^v$, it follows that $h(d) = h(d')$. Hence $h(D)$ is $a.s.$ constant and, because $h(D) \overset{as}{=} l(V)$, $l(v)$ is $a.s.$ constant.

■

Proof of Theorem 3

Let $\mathcal{Z}$ denote the support of the distribution of Z. Consider any two functions m_1 and m_2 such that $m_1(D) = m_2(V)$ a.s.. For (a.e. F_V) fixed v_0, using the assumption that Z and V are independent, it follows that $m_1(g(z, v_0)) = m_2(v_0)$ for a.e. z conditional on $V = v_0$ implies that m_1 is (a.s. F_Z) constant on $\{g(z, v_0) : z \in \mathcal{Z}\}$. Likewise, for a v_1 close to v_0, we have m_1 is constant on $\{g(z, v_1) : z \in \mathcal{Z}\}$. Using the fact that $g(z, v)$ is continuous in v and that $\{g(z, v) : z \in \mathcal{Z}\}$ contains an open interval for any v, we can pick v_1 sufficiently close to v_0 so that $\{g(z, v_0) : z \in \mathcal{Z}\}$ and $\{g(z, v_1) : z \in \mathcal{Z}\}$ have a nonnegligable intersection, and we thus conclude that m_1 is constant on $\{g(z, v) : z \in \mathcal{Z}, v = v_0, v_1\}$. Proceeding in this fashion, we have that m_1 is (a.s.) constant on $\{g(z, v) : z \in \mathcal{Z}, v \in [0, 1]\}$, and thus that m_1 is a.s. equal to a constant. ■

References

Altonji, J. G. and R. L. Matzkin (2005, July). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Björklund, A. and R. Moffitt (1987, February). The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics 69*(1), 42–49.

Blundell, R. and J. Powell (2004, July). Endogeneity in semiparametric binary response models. *Review of Economic Studies 71*(3), 655–679.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 5, pp. 1801–1863. New York: North-Holland.

Card, D. (2001, September). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Chernozhukov, V. and C. Hansen (2005, January). An iv model of quantile treatment effects. *Econometrica 73*(1), 245–261.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2006). Nonparametric identification and estimation of non-separable models. *Journal of Econometrics*. forthcoming.

Chesher, A. (2003, September). Identification in nonseparable models. *Econometrica 71*(5), 1405–1441.

Chesher, A. (2007). Identification of non-additive structural functions. In W. K. N. Richard Blundell and T. Persson (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Volume 3, Chapter 1. New York: Cambridge University Press. Presented at the Econometric Society Ninth World Congress. 2005. London, England.

Darolles, S., J.-P. Florens, and E. Renault (2002). Nonparametric instrumental regression. Working Paper 05-2002, Centre interuniversitaire de recherche en conomie quantitative, CIREQ.

Florens, J.-P., M. Mouchart, and J. Rolin (1990). *Elements of Bayesian Statistics.* New York: M. Dekker.

Heckman, J. J. (1979, January). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J. (1997, Summer). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources 32*(3), 441–462. Addendum published vol. 33 no. 1 (Winter 1998).

Heckman, J. J. and B. E. Honoré (1990, September). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J. 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*, Volume 10, pp. 156–245. New York: Cambridge University Press.

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

Heckman, J. J., J. A. Smith, and N. Clements (1997, October). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies 64*(221), 487–536.

Heckman, J. J. and E. J. Vytlacil (1998, Fall). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources 33*(4), 974–987.

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

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. Amsterdam: Elsevier. Forthcoming.

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

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and W. K. Newey (2002). Identification and estimation of triangular simultaneous equations models without additivity. Technical Working Paper 285, National Bureau of Economic Research.

Newey, W. K. and J. L. Powell (2003, September). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999, May). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–603.

Roy, A. (1951, June). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Vytlacil, E. J. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Wooldridge, J. M. (1997, October). On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters 56*(2), 129–133.

Wooldridge, J. M. (2003, May). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters 79*(2), 185–191.

Wooldridge, J. M. (2007). Instrumental variables estimation of the average treatment effect in correlated random coefficient models. In D. Millimet, J. Smith, and E. Vytlacil (Eds.), *Advances in Econometrics: Modeling and Evaluating Treatment Effects in Econometrics*, Volume 21. Amsterdam: Elsevier. Forthcoming.