INSTRUMENTAL VARIABLES, LOCAL INSTRUMENTAL VARIABLES AND CONTROL FUNCTIONS

Jean-Paul Florens
James Heckman
Costas Meghir
Edward Vytlacil

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP15/02

Instrumental Variables, Local Instrumental Variables and Control Functions[*]

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

July 5, 2004

Abstract

We consider the identification of the average treatment effect in models with continuous endogenous variables whose impact is heterogeneous. We derive a testable restriction that allows us to assess the degree of unobserved heterogeneity. Our analysis uses assumptions relating to the Local Instrumental Variables (LIV) approach and the control function approach.

[*] We 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. 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. All errors are our own.

[†] IDEI, Toulouse

[‡] University of Chicago

[§] IFS and UCL, c.meghir@ucl.ac.uk

[¶] Stanford University

1. Introduction

The common practice in empirical economic models is to assume that the unobservables are additively separable from the observables, particularly when the latter are endogenous. This is done because it is recognized that serious identification problems arise when such interactions are allowed for. However, more often than not such additivity is, to say the least, contrived and often inconsistent with the overall stochastic specification of the model. Good examples are demand functions, where either the price or the total expenditure impacts are likely to be heterogeneous; wage equations, where the returns to education are likely to vary with unobserved ability; labor supply, where wage effects may be heterogeneous; or production functions, where the technology may vary across firms, at least in the short run. In all these examples the model one may want to estimate includes a continuous endogenous variable whose impact varies over the population even conditional on observable characteristics. In this paper we address the non-parametric identification of the average effect of endogenous variables in such models.

There has been a growing theoretical and empirical literature on models where the impact of discrete (usually binary) treatments are heterogeneous in the population.[1] This leads to important identification questions and questions relating to the interpretability of standard methods such as instrumental variables.[2] Within this

[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, 2001a,a b; Heckman and Vytlacil 2000, 2001, who discuss heterogeneous response models.

[2] see, for example, Heckman and Robb (1986), Imbens and Angrist (1994), and Heckman (1997).

context, the issue of choosing an appropriate parameter of interest has arisen, since the heterogeneity in impacts implies a whole distribution of effects, rather than one fixed parameter as in the traditional literature. Parameters that have received a great deal of attention include the average treatment effect (ATE) which is the expected impact of the treatment on a randomly chosen individual, and the impact of treatment on the treated, which is the expected impact on a randomly chosen individual among those who chose to have treatment. In this paper we focus on continuous treatments, such as years of education, expenditure, income, prices etc. In this context we discuss parameters of interest and we focus on the identification and estimation of ATE, pointwise over the entire observed support of the treatment. For example, we consider the impact of a marginal increase in expenditure on budget shares at each value of expenditure.

It should be obvious that some structure has to be imposed on the nature of heterogeneity and the way it interacts with the endogenous variables. We express the model in counterfactual notation by specifying it as a stochastic process indexed by d, the endogenous treatment variable. The outcome we observe is then this stochastic process evaluated at an endogenous realization of d. We then put some structure on how the unobservables evolve with d, considering a linear and a quadratic random function in d as well as a more general structure.

We start investigating identification with instrumental variables (IV). We show that IV assumptions are not in general sufficient to identify ATE (or treatment on the treated). In fact, standard exclusion restrictions only identify parameters of interest in very simple stochastic setups. Even the LATE parameter of Imbens and Angrist, which has attracted so much attention, is only interpretable under conditions on the way individuals are assigned to treatment. We are thus motivated to go beyond the standard IV framework and consider the usefulness of the control function assumption (see Heckman, 1979 and Heckman and Robb, 1985) as well as a more recently developed *Local IV* approach (*LIV*, see Heckman and Vytlacil, 2000). We derive conditions under which the various assumptions are equivalent, which helps in their interpretation. We also derive a testable restriction with which to identify the "degree of unobserved heterogeneity."

1.0.1. Education and Wages: A Simple Illustration

To illustrate the type of problem we are concerned about in this paper we present a very simple model of education choice.

Suppose that the agent receives wages Y_d at the cost C_d if schooling choice d is made. The cost can include foregone earnings and other direct costs; everything is suitably scaled to reflect the fact that earnings flow over the entire working life-cycle. We write wages for schooling level d as Y_d, as well as the cost function for schooling as

$$\begin{aligned} Y_d &= \varphi_0(X) + (\varphi_1(X) + \varepsilon_1)d + \tfrac{1}{2}\varphi_2(X)d^2 + \varepsilon_0 \qquad I \\ C_d &= C_0(X,Z) + (C_1(X,Z) + v_1)d + \tfrac{1}{2}C_2(X,Z)d^2 + v_0 \quad II \end{aligned} \tag{1.1}$$

where ε_s and v_s $(s = 1, 2)$ reflect unobserved heterogeneity both in the wage level and in the return to schooling. These unobserved heterogeneity terms are the source of the identification problem. The variables X reflect human capital characteristics that affect both wages and the costs of schooling, while Z are factors affecting the cost of schooling only, such as the price of education.

Assume that agents chose education to maximize wages minus costs, and let

D denote the resulting optimal choice of education. Then D solves the first order condition

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

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

$$D = P(X, Z) + V \tag{1.2}$$

where

$$P(X, Z) = \frac{\varphi_1(X) - C_1(X, Z)}{\varphi_2(X) - C_2(X, Z)}$$

and

$$V = \frac{\varepsilon_1 - v_1}{\varphi_2(X) - C_2(X, Z)}. \tag{1.3}$$

The object of the empirical analysis is to estimate the average return to education, which here is given by $\varphi_1(X) + \varphi_2(X)d$ but which could in principle be a more complicated unknown nonlinear function of d. In this context the assignment rule is complicated by the fact that the errors are heteroscedastic. Moreover, this is not a single index model and the assignment rule does not satisfy monotonicity. For example, depending on the sign of $\varepsilon_1 - v_1$ it is possible that an increase in Z (which can be thought of as a policy variable such as tuition) may increase or decrease educational participation. This would not happen if the second derivative of the cost function did not depend on Z, in which case we would be back to a single index model conditional on X.

Our discussion of identification will implicitly look at a number of cases, including simpler situations. We will start with a simple case of homogeneous returns to d (education here) and then proceed to more complicated settings.

2. The Model, Some Parameters of Interest and the Observables.

We consider models in the class

$$Y_d \quad = \quad \varphi(d, X) + U_d$$

where d defines the level of treatment intensity and where we define $\varphi(d, X) = E(Y_d|X)$ which implies by construction that $E(U_d|X) = 0$. The assumption that $E(U_d|X) = 0$ is just a normalization; in other words we do not assign any causal interpretation related to changes in the value of X. Thus the derivatives of $\varphi(d, X)$ with respect to X need not have any causal interpretation. This is very much in the spirit of the treatment effects literature, where no causal interpretation is attached to the impact of X. In this sense we are not identifying a complete structural model here. It can be binary, which is the case that has been studied extensively in the literature; it can be discrete ordered. However, in this paper we focus on the case where d is continuous. An example would be a demand function, where d is the price of a good and where unobservables are not additively separable.

For the purposes of economic evaluation we are interested in certain aspects of φ as well as potentially in the joint distribution of the U_d. A parameter of interest that follows naturally from the definition of the model is the Average Treatment

Effect,

$$\Delta^{ATE}(d,x) = \frac{\partial}{\partial d}E(Y_d|X=x) = \frac{\partial}{\partial d}\varphi(d,x) \tag{2.1}$$

or higher order derivatives of the average.[3]

Some of the proofs will have the structure of first identifying $E(Y_d|X=x) \equiv \varphi(d,x)$, and then using identification of $\varphi(d,x)$ to identify $\frac{\partial}{\partial d}E(Y_d|X=x) = \frac{\partial}{\partial d}\varphi(d,x)$. Some of the proofs will have the same structure but with $\varphi(d,x)$ only identified up to an unknown, additive function of x. We will thus need conditions under which $\frac{\partial}{\partial d}\varphi(d,x)$ is well defined to have identification of $\frac{\partial}{\partial d}\varphi(d,x)$ a.s. follow from identification of $\varphi(d,x)$ a.s. up to an additive, unknown function of x.

In general identification results require some structure to be imposed on the stochastic process U_d. Typically we will require some continuity and possibly additional smoothness conditions. In general we can think of approximating U_d by a sum of known functions of d weighted by random coefficients, i.e.

$$U_d = \sum_{j=0}^{K} \alpha_j(d)\varepsilon_j, \tag{2.2}$$

where $\alpha_j(d)$ are the first elements of suitable basis of the space of functions and the ε_j are the random components of the stochastic process. In this paper we will consider the case of a power series in d. We will start by considering the usual *zero* order case, including a summary of existing results. We will then

[3] An equivalent expression is

$$\lim_{\Delta d \to 0} \frac{E(Y_{d+\Delta d} - Y_d|D=d, X=x)}{\Delta d},$$

consider identification in the more general higher order cases. We subsequently discuss diagnostic tests for higher order heterogeneity (i.e. higher order random coefficients).

We now complete the model by introducing a description of the mechanism assigning a particular treatment level to each individual, denoted by D. We define

$$D \;=\; P(X,Z)+V$$

where we define $E(D|X,Z) = P(X,Z)$. In the sequel we will use the variables Z as instruments, which only determine the level of treatment (D), in ways that will be defined precisely.

At this point it is useful to define the notion of an expected outcome at treatment intensity d_1 given that the individual chose/was assigned to treatment d_2. This is denoted by $E\left(Y_{d_1}|D=d_2,X=x\right)$. Given this and the model of treatment assignment we can also define a commonly used parameter, which is the treatment on the treated (TT)[4]

$$\begin{aligned}\Delta^{TT}(d,x) &= [\frac{\partial}{\partial d_1}E\left(Y_{d_1}|D=d_2,X=x\right)]_{d=d_1=d_2}\\ &= [\frac{\partial}{\partial d_1}\varphi(d_1,x)+\frac{\partial}{\partial d_1}E(U_{d_1}|D=d_2,X=x)]_{d=d_1=d_2}.\end{aligned}$$

Clearly if we can observe all outcomes (actual and counterfactual) independently of the choice of treatment d, there is obviously no identification issue.

[4] An equivalent expression is

$$\lim_{\Delta d\to 0}\ \frac{E(Y_{d+\Delta d}-Y_d|D=d,X=x)}{\Delta d},$$

Thus, to set the scene for the discussion of identification we assume we observe realizations of the random variable $Y = Y_D$ and of D as well as the relevant X and Z. Thus we can never observe the counterfactual outcome, i.e. the outcome $Y_{d'}$ for some value d' different from the actual chosen treatment level.

In order to better understand the issues of identification we need to define

$$g(d_1, d_2, x, z) = E(U_{d_1} | D = d_2, X = x, Z = z) \tag{2.3}$$

In other words the function $g(d_1, d_2, x, z)$ is the conditional expectation of the random error term corresponding to treatment level d_1 when the choice that is made by the individual is to take the treatment level d_2. Thinking of an education choice example, $g(d_1 = 9, d_2 = 10, z, x)$ would be the expected value of the unobservable part of the outcome equation at nine years of education for someone choosing ten years of education. In the case where $d_1 = d_2$, we get the conditional expectation of the outcome at d when the choice is in fact d. Thus

$$E(Y | D = d, X = x, Z = z) = \varphi(d, x) + \tilde{g}(d, x, z) \tag{2.4}$$

where we have defined $\tilde{g}(d, x, z) = g(d, d, z, x)$. Since the data itself only identifies the conditional expectation in equation (2.4), $\varphi^1(d, x)$ and $\varphi^2(d, x)$ are observationally equivalent if we can find two functions $g^1(d_1, d_2, z, x)$ and $g^2(d_1, d_2, z, x)$ such that $\varphi^1 + \tilde{g}^1 \overset{as}{=} \varphi^2 + \tilde{g}^2$. The average treatment effect, $\frac{\partial}{\partial d}\varphi(d, x)$ is identified, if any two observationally equivalent functions $\varphi(d, x)$ and $\varphi^1(d, x)$ have the same first derivative, i.e. $\frac{\partial}{\partial d}\varphi(d, x) = \frac{\partial}{\partial d}\varphi^1(d, x)$. Moreover, the effect of treatment on the treated $[\frac{\partial}{\partial d_1}\varphi(d_1, x) + \frac{\partial}{\partial d_1}g(d_1, d_2, z, x)]_{d=d_1=d_2}$ is identified, if, for any functions $\varphi^1(d_1, x)$, $g^1(d_1, d_2, z, x)$, $\varphi + \tilde{g} \overset{as}{=} \varphi^1 + \tilde{g}^1$ implies $[\frac{\partial}{\partial d_1}\varphi + \frac{\partial}{\partial d_1}g]_{d=d_1=d_2} \overset{as}{=} [\frac{\partial}{\partial d_1}\varphi^1 + \frac{\partial}{\partial d_1}g^1]_{d=d_1=d_2}$.

We now discuss the identification of certain parameters of interest under different assumptions. The way we approach the problem is first to start by looking at identification in the simpler homogeneous impact model. In that context we first consider identification under the standard orthogonality conditions, then under a control function assumption and lastly by assuming that the function we wish to identify satisfies the Local Instrumental Variables equation. In general, these conditions are not equivalent, though we proceed to derive conditions under which these assumptions are equivalent. This leads to a set of assumptions under which ATE is identified by any of these conditions.

We then proceed to look at a model with heterogeneous impacts. We show that the usual orthogonality conditions no longer identify ATE (as is well understood in the treatment effects literature); we then proceed to show that the model is identified using an extension of the control function assumption we made earlier in the homogeneous impact model. We allow for heterogeneity where the first derivatives of the function of interest are additive in the unobservables; we then generalize to more complex forms. Up to that point many of the identification results rely on the assignment rule to treatment belonging to the single index family. We then explore relaxing this assumption, which has important implications for the type of economic models we can handle. In the next section we approach the problem in a more general fashion and we derive conditions on the control function which imply that LIV can be used to identify ATE.

3. The Common Treatment Effects Model

We start discussing identification within the more conventional common treatment effect model where $U_d \equiv U$. Of course in this context the Average Treatment Effect and the Treatment on the Treated are identical. The main issue that arises in this context is that of the non-parametric identification of an otherwise standard simultaneous equations model. Below we consider instrumental variables, control function, and Local instrumental Variables approaches. All these approaches are based on alternative assumptions and we will show that they all identify the average treatment effect given the validity of their assumptions. Nevertheless, they are different and the assumptions invoked by no one approach implies the assumptions invoked in another approach. However, we show below that suitable independence assumptions unify the conditions imposed by all of these approaches. The distinction between the approaches will become particularly interesting when we deal with the heterogeneous treatment effects model.

Traditionally, researchers have focused on instrumental variables, which will be our point of departure. Hence we assume that

A1. Regularity condition $\varphi(D, X)$ is differentiable in D (a.s.), and the support of D conditional on X does not contain any isolated points (a.s.).

This regularity condition (**A1**) will be assumed throughout.

A2. $E(U|X, Z) = E(U|X)$ (Exclusion restriction)

Given the definition of $\varphi(d, X) \equiv E(Y_d|X)$, we have by construction that $E(U|X) = 0$ and thus by **A2** that $E(U|X, Z) = 0$.

We also need a rank condition which ensures that our instrument has explanatory power. In linear models this assumption takes a relatively simple form, requiring that the instruments are correlated with D (conditional on X). However, in the context of nonparametric identification, we need to take into account that we do not generally know the form of the function $\varphi(d,x)$; here we require a more general dependence condition between D and the instruments Z that, loosely speaking, ensures that any function of D is correlated with some function of Z.

Definition 3.1. *We say that D is strongly identified by Z given X if for any function $\lambda(D,X)$ we have that $E(\lambda(D,X)|X,Z) \overset{a.s.}{=} 0 \Longrightarrow \lambda(D,X) \overset{a.s.}{=} 0$.*[5]

This assumption can be viewed as a non-parametric extension of the rank condition. An interpretation is that any conceivable function of D is correlated with some function of Z. Thus we state the following assumption.

A3. D is strongly identified by Z given X.

We now state the first result in terms of a theorem:[6]

Theorem 3.2. *Assume that the exclusion restriction (**A2**) holds, and that D can be strongly identified by the instrument Z given explanatory variables X (**A3**). Then $E(Y_d|X=x) \equiv \varphi(d,x)$ is identified.*

Proof. See Appendix. ■

[5] Equivalently we can write that $E(\Psi(D,X)|X,Z) \overset{a.s.}{=} \Psi_0(X) \Rightarrow \Psi(D,X) \overset{a.s.}{=} \Psi_0(X)$. This assumption is related to the concept of completeness as used in the statistical literature on minumum variance unbiased estimation. Newey and Powell (1989) and Darolles, Florens, and Renault (2002) use a similar condition.

[6] An analogous result is proved by Newey and Powell (1989) and Darolles, Florens and Renault (2002).

In order to contrast with the identification results in the heterogeneous treatment effects section, we should emphasize that here identification does not require independence of the unobservables from the instrument, but just mean independence of U from Z. It imposes no structure on the model driving the treatment choice D, other than the strong identifiability condition **A3.** The first stage equation, $D = P(X, Z) + V$, has played no role in this analysis. The only restriction on the relationship between Z and D needed by the theorem is the strong identification assumption (**A3**).

Theorem 3.2 provides identification of the function $\varphi(d, x)$ (a.s.), while our parameter of interest is the derivatives of this function with respect to d. Combining Lemma 3.4 below and Theorem 3.2, we immediately have the following result.

Corollary 3.3. *Assume (**A1**), (**A2**), and (**A3**). Then* $\frac{\partial}{\partial d}E(Y_d|X = x) = \frac{\partial}{\partial d}\varphi(d, x)$ *is identified.*

This corollary follows from the following lemma:

Lemma 3.4. *Assume (**A1**). Then identification of* $\varphi(d, x)$ *(a.s.) up to an additive function of* x *implies identification of* $\frac{\partial}{\partial d}\varphi(d, x)$.

Proof. See Appendix. ■

Heckman and Vytlacil (1999) propose a new approach to the identification of causal effects, the Local Instrumental Variables (LIV) approach. The key LIV assumption is

A4 (LIV)**.**

$$E(\frac{\partial}{\partial D}\varphi(D, X)|Z = z, X = x) = \frac{\frac{\partial E(Y|X,Z)}{\partial z_j}}{\frac{\partial E(D|X,Z)}{\partial z_j}} \ \forall j$$

In the standard linear IV model this condition holds immediately. However, this does not follow from the usual orthogonality conditions in non-linear models. It states that the causal effect averaged over all values of the treatment and *at a given value of the instrument* is a scaled version of the marginal effect of the instrument on the expected outcome.

We wish to identify ATE pointwise, i.e., we wish to identify $\frac{\partial}{\partial d}\varphi(d,x)$ for each d and x in the appropriate support. In contrast, **A4** immediately provides identification of an average of ATE, averaging over the distribution of D conditional on X and Z. However, if D is strongly identified by Z given X than we can use **A4** to obtain identification of ATE pointwise as shown by the following theorem.

Theorem 3.5. *If D is strongly identified by Z given X (**A3**), and if the expectation of ATE conditional on Z and X satisfies the LIV condition (**A4**), then ATE ($\Delta^{ATE}(d,x) = \frac{\partial}{\partial d}\varphi(d,x)$) is identified.*

Proof. See Appendix. ■

Note that LIV does not impose much structure on the assignment rule. It should be stressed that assumption **A4** is generally not implied and does not imply assumptions (**A2**) and (**A3**), which characterize instrumental variables. Of course the problem is that assumption (**A4**) is unusual and possibly difficult to relate directly to economic theory. However, an additional assumption unifies the IV and LIV approaches and makes them equivalent. This assumption now imposes restrictions on the assignment rule. Thus, let $p(D|X,Z)$ be the conditional density of D given X and Z. Then we assume that

A5 *Single Index:* V is independent of Z given X.

(**A5**) implies that $p(D|Z,X) = p(D - P(X,Z)|X)$. We will also consider the following, stronger condition.

A5′ (V,U) are jointly independent of Z given X.

To proceed we need to define measurable separability:

Definition 3.6. *We say that Q_1 and Q_2 are measurably separated if whenever a function of Q_1 is almost surely equal to a function of Q_2 this is the constant function.*

Theorem 3.7. *Assume that:*

1. *Z and X are measurably separated,*
2. *the support of the conditional distribution of D given $X = x$, $Z = z$ is an interval $(D^L_{x,z}, D^U_{x,z})$ (possibly infinite) and φ satisfies: $\varphi(D^L_{x,z}, x)p(D^L_{x,z}|x,z) = \varphi(D^U_{x,z}, x)p(D^U_{x,z}|x,z) = 0$ where $p(d|x,z)$ is the conditional density w.r.t. the Lebesgue measure of D given $X = x, Z = z$,*
3. *all functions involved are smooth and square integrable,*
4. *V is independent of Z given X (**A5**).*

*Then the exclusion restriction (**A2**) and the LIV assumption (**A4**) are equivalent. Conversely, if (**A2**) and (**A4**) are equivalent for any function φ then the single index assumption (**A5**) holds*

Proof. See Appendix. ■

In other words if all the dependence of D on Z comes through a single function (conditional on X) then the LIV and Instrumental Variables can be used to identify ATE under the same conditions. Note that the single index assumption **A5,** is imposing both that the treatment can be written as an additively separable function of observables and unobservables and that the unobservables are independent of Z. This assumption is not innocuous.

The literature on selection models and non-linear simultaneous equations models have also used the control function approach to identify ATE.[7] The control function can be defined as follows: Let $\tilde{V}$ be a real-valued (square integrable) function of (D, X, Z) such that the σ-field generated by (D, X, Z) is identical to the σ-field generated by $(\tilde{V}, X, Z)$.[8] The function $\tilde{V}$ is called a control function (see Heckman and Robb, 1985). Generally the assumptions that allow identification using a control function are not equivalent to those that justify the IV and LIV approaches. Formally, we assume

A6. *Control Function:* There exists a real valued function $h(\tilde{V}, X)$ such that $E(Y|D, Z, X) = \varphi(D, X) + h(\tilde{V}, X)$, where $\tilde{V}$ is a real-valued (square integrable) function of (D, X, Z) such that the σ-field generated by (D, X, Z) is identical to the σ-field generated by $(\tilde{V}, X, Z)$.

Essentially this imposes that the dependence of the distribution of the unobservables in the outcome equation (U) on the unobservable in the assignment equation (V) and on the instrument Z operate through the same channel, i.e. through this function $\tilde{V}$. This usually will turn out to be the residual of the assignment equation, i.e., $\tilde{V} = V$ where $V \equiv D - P(X, Z)$. For identification purposes we need to be able to distinguish the two functions φ and h. Thus we

[7] see Heckman (1979) and in the context of non-parametric simultaneous equations models see Newey, Powell and Vella (1999).

[8] This property is obtained if $\tilde{V}$ is a one to one measurable function of D given X and Z.

will also need to impose that the control function has some independent variation from D conditional on X. This notion is formalized in the following assumption.

A7. *Rank condition:* D and $\tilde{V}$ are measurably separated given X, i.e., any function of D and X almost surely equal to a function of $\tilde{V}$ and X must be a function of X only.

A necessary condition for assumption **A.7** to hold is that the instruments Z have an impact on D.[9]

Theorem 3.8. *Assume that we can write $E(Y|D,X,Z) = \varphi(D,X) + h(\tilde{V},X)$ (**A6**), and that D and $\tilde{V}$ are measurably separated given X (**A7**). Then $E(Y_d|X = x) = \varphi(d,x)$ is identified up to an additive function of x.*

Proof. See Appendix. ■

Applying Lemma 3.4, we state the following result,

Corollary 3.9. *Assume (**A1**), (**A6**), and (**A7**). Then $\frac{\partial}{\partial d}E(Y_d|X = x) = \frac{\partial}{\partial d}\varphi(d,x)$ is identified.*

Finally, independence unifies the assumptions invoked by these approaches and makes them equivalent. Thus we present two equivalence results

Theorem 3.10. *The single index assumption (**A5**) and the control function assumption (**A6**) with $\tilde{V} = V$ imply the exclusion restriction (**A2**) and the LIV*

[9] Measurable separability, which we maintain in this paper is just one way of achieving this. Alternatively, one could restrict the space of functions $\varphi(D,X)$ not to contain $h(\tilde{V},X)$ functions; this in turn can be achieved for example by assuming that $\varphi(D,X)$ is linear in D and h is nonlinear as in the Heckman (1979) selection model.

*assumption (***A4***). Hence under independence (***A5***) and the control function assumption (***A6***) with $\tilde{V} = V$, as well as under the rank condition (***A7***) the control function approach provides a solution which satisfies the IV, LIV assumptions.*

Proof. See Appendix. ■

Theorem 3.11. *Assume that (V, U) are jointly independent of Z given X (***A5′***). Then the exclusion restriction (***A2***), the control function restriction (***A6***) and the single index assumption (***A5***) hold. Hence under independence (***A5′***) (and the rank conditions) the three approaches (*control function, IV *and* LIV*) are equivalent and all identify* ATE.

Proof. See Appendix. ■

In conclusion, as has been known at least since the seminal paper of Newey and Powell (2003) with additive separable errors instrumental variables can identify the function $\varphi(d|X)$ so long as strong identifiability holds, which is just a generalization of the rank condition to a non-parametric framework. It also turns out that the control function assumption and the LIV assumptions if true identify this function. However all these assumptions are not equivalent and none imply any of the others. The IV assumption has the appealing feature that it is often justifiable from economic theory and one can design randomized experiments where an instrument such as tuition or a price is randomly assigned, which allows identification of the structural effects. However under conditional independence of the errors from the instrument Z given X, all approaches are equivalent. Thus equivalence requires us to say something about the assignment rule. As we shall now see, with heterogeneous treatment effects identification will also in general rely on extra assumptions on the assignment rule.

4. Models with Heterogeneous Treatment Effects

We now discuss the class of models that were the original motivation of this paper, namely models where the impact of the treatment D is heterogeneous. We focus on the case where the realization of the treatment is correlated with the impact of the treatment. This can happen, for instance, when the allocation to treatment depends on the individual's potential benefit from the treatment intensity, such as in our introductory example.

Identification results require some structure to be imposed on the stochastic process U_d. Typically we will require some continuity and possibly additional smoothness. We will study the case where U_d is given by a finite order polynomial in d,

$$U_d = \sum_{j=0}^{K} d^j \varepsilon_j, \tag{4.1}$$

where we have adopted the convention that $0^0 = 1$.

Usually, models allow just the level of the outcome variable to be random. However, here we also allow the higher order derivatives to be random. For the binary treatment case a linear form ($K = 1$) is completely general. However, with more than one outcome for D or in particular for D continuous this specification is constraining.[10] Therefore, we develop our analysis for K of any finite order.

[10] If d takes $K + 1$ values, than a $K + 1$ order polynomial will be completely general. In particular, if d takes $K + 1$ values, $d \in \{d_0, d_1, ..., d_K\}$, then it is always possible to define

The analysis may be seen as an approximation to a more general set of possible stochastic functional forms.[11]

We now discuss the assumptions we will be using. All our specifications require the exclusion of a continuous instrument from the outcome equation. Thus we impose

A2′. $E(U_d|X,Z) = E(U_d|X)\ \forall\ d$ (Exclusion restriction)

Imposing equation (4.1), restriction **A2′** is equivalent to $E(\varepsilon_k|X,Z) = E(\varepsilon_k|X)$ for all $k = 0,...,K$. We normalize the errors such that $E(\varepsilon_k|X) = 0\ \forall\ k$.

In equation (2.3), we defined the conditional expectation of the unobservable for outcome d_1 when the choice made is d_2. Under linear heterogeneity on the unobservables (equation (4.1) with $K = 1$), this function takes the form

$$g(d_1, d_2, x, z) = d_1 r_1(d_2, x, z) + r_0(d_2, x, z),$$

where each term is defined by

$$r_1(d_2, x, z) = E(\varepsilon_1|D = d_2, X = x, Z = z),$$

$$r_0(d_2, x, z) = E(\varepsilon_0|D = d_2, X = x, Z = z),$$

random variables $\varepsilon_0, ...\varepsilon_K$ such that $U_d = \sum_{j=0}^{K} d^j \varepsilon_j$ for all $d \in \{d_0, d_1, ..., d_K\}$. To see this, define $U = [U_{d_0}, U_{d_1}, ..., U_{d_K}]'$. Define the $K \times K$ matrix A to have (i,j) element given by $A_{i,j} = (i-1)^j$. Note that A is invertible. Define $\epsilon = A^{-1}U$, and let ϵ_k denote the $k-1$ element of ϵ. The constructed $\{\epsilon_k : k = 0, ..., K\}$ then satisfy the desired property, $U_d = \sum_{j=0}^{K} d^j \varepsilon_j$ for all $d \in \{d_0, d_1, ..., d_K\}$.

[11] If d takes values only on a compact interval, the Weierstrass theorem implies that U_d can be approximated uniformly by a polynomial function for any realization such that U_d is a continuous function of d. Assuming that U_d is a continuous function of d w.p.1, we have that w.p.1 we can approximate U_d uniformly by a polynomial function where the coefficients of the polynomial function depend on the realization and are thus random variables.

and hence the conditional expectation of the outcome at level of intensity d when d was actually chosen (see 2.4) becomes

$$E(Y|D = d, X = x, Z = z) = \varphi(d, x) + dr_1(d, x, z) + r_0(d, x, z). \tag{4.2}$$

Hence identification relates to our ability to characterize (some aspects) of $\varphi(d, x)$, $r_1(d, x, z)$ and $r_0(d, x, z)$. Note that the parameter, "Treatment on the Treated" can now be expressed as

$$\Delta^{TT}(d, x) = \frac{\partial}{\partial d}\varphi(d, x) + E[r_1(d, z, x)|D = d, X = x].$$

This framework is fundamentally different from the one earlier on and generally standard exclusion restrictions are not sufficient to identify all the parameters of interest. We show by an example that ATE is not identified generally just with exclusion restrictions.

Note that identification of ATE is equivalent to the implication that for any functions $\varphi^*(d, x)$, $r^*(d, z, x)$ and $h^*(d, z, x)$ that satisfy

$$\varphi^*(d, x) + dr^*(d, z, x) + h^*(d, z, x) = 0 \tag{4.3}$$

it must be that[12]

$$\frac{\partial \varphi^*}{\partial d} = 0.$$

[12] To see this, take two values of $\varphi(d, x)$, $r_1(d, z, x)$ and $r_0(d, z, x)$, e.g. φ^s, r^s, and h^s for $s = 1, 2$. These are observationally equivalent if they generate the same $E(Y|D = d, X = x, Z = z)$. Let us take the difference which gives $\varphi^1 - \varphi^2 + d(\ r^1 - r^2) + h^1 - h^2 = 0$, or $\varphi^* + d\ r^* + h^* = 0$. Identification condition of ATE requires that under the orthogonality conditions, this equation implies that $\frac{\partial \varphi^1}{\partial d} - \frac{\partial \varphi^2}{\partial d} \equiv \frac{\partial \varphi^*}{\partial d} = 0$. For the TT parameter, the corresponding condition is that $\varphi^* + d\ r^* + h^* = 0$ implies $\frac{\partial \varphi^*}{\partial d} + E(r^*|D, X) = 0$. Note that neither condition is stronger or weaker than the other.

We have extended the definition of g to the case of linear heterogeneity specified by equation (4.1) with $K = 1$. The definition further extends in the obvious way to the case of higher order polynomial heterogeneity.

4.1. Instrumental Variables and Heterogeneous Treatment Effects

First consider instrumental variables in the case of linear heterogeneity, taking equation (4.1) with $K = 1$. Imposing the exclusion restriction **A2′** restricts the set of admissible functions r_1 and r_0 defined above. Thus we have that $E(r_1(D,X,Z)|X,Z) = E(r_1(D,X,Z)|X)$ and $E(r_0(D,X,Z)|X,Z) = E(r_0(D,X,Z)|X)$. The question is whether the functions that satisfy these conditions and solve equation (4.3) are such that $\frac{\partial\varphi^*}{\partial d} = 0$. In this case IV would identify the model, subject to the strong identification condition. In general this is not the case as the following counter example shows.

Let us consider for simplicity a case without X variables, Z is a positive random variable and the distribution of D conditional on Z satisfies: $E(D|Z) = Var(D|Z) = E((D-Z)^3|Z) = Z$. The above implies that $E(D^2|Z) = Z^2 + Z$ and $E(D^3|Z) = Z + 3Z^2 + Z^3$. Now we suppose that $r^*(d,z) = d^2 - (z + z^2)$ (hence $E(r^*(D,Z)|Z) = 0$) and that $\varphi^*(d) = -2d^2 + d$. Now suppose h^* satisfies

$$\begin{aligned} h^*(d,z) &= -[\varphi^*(d) + dr^*(d,z)] \\ &= 2d^2 - d - d(d^2 - (z + z^2)) \end{aligned}$$

One can easily check that $E(h^*(D,Z)|Z) = 0$. With these chosen functions the orthogonality conditions are satisfied and equation (4.3) is satisfied, but clearly $\frac{\partial\varphi^*}{\partial d} \neq 0$. Note that this example is not in contradiction with the assumption that D is strongly identified by Z.

With additional conditions, Instrumental Variables will identify ATE. Heckman and Vytlacil (1998) analyzed instrumental variables applied to a linear model with a random coefficient. Their model is a special case of that considered here, with a linear structure imposed on $\varphi(D, X)$ and with $K = 1$ in equation (4.1). They considered the following assumption restricting the relationship of the unobservables in the outcome and assignment equation.

A2′′. $E(\epsilon_1 V|X, Z) = E(\epsilon_1 V|X)$ (Covariance restriction)

Note that the example of nonidentification considered above violates **A2′′**. Under **A2′′**, it is possible to obtain positive results for IV for the case of linear heterogeneity as shown by the following theorem.

Theorem 4.1. *Assume that equation (4.1) holds with $K = 1$. Assume that the exclusion restriction (**A2′**) holds, that the covariance restriction **A2′′** holds, and that D can be strongly identified by the instrument Z given explanatory variables X (**A3**). Then $E(Y_d|X = x) \equiv \varphi(d, x)$ is identified up to an additive function of x.*

Proof. See Appendix. ■

Combining Lemma 3.4 and Theorem 4.1, we immediately have the following result.

Corollary 4.2. *Assume (**A1**), (**A2′**), (**A2′′**) and (**A3**). Then $\frac{\partial}{\partial d}E(Y_d|X = x) = \frac{\partial}{\partial d}\varphi(d, x)$ is identified.*

The assumption that $E(\epsilon_1 V|X, Z) = E(\epsilon_1 V|X)$ is not innocuous. Consider, for example, the model $D = \tilde{P}(X, Z, \tilde{V})$ with $Z \perp\!\!\!\perp (\epsilon_1, \tilde{V})|X$. The independence property stated in terms of $\tilde{V}$ in the "structural" model does not imply that $E(\epsilon_1 V|X,Z) = E(\epsilon_1 V|X)$ where V is defined as a deviation from a conditional expectation. For example, consider $\tilde{P}(X,Z,\tilde{V}) = P(X,Z) + \sigma(X,Z)\tilde{V}$, so that $V = \sigma(X,Z)\tilde{V}$. In this case, $E(\epsilon_1 V|X,Z) = \sigma(X,Z)E(\epsilon_1 \tilde{V}|X)$, so that **A2**$''$ does not hold.

In particular, if the unobservables in the equation determining the level of the treatment are additively separable from the observables *and* the unobservables in the outcome equation and the treatment equation are jointly independent from the instruments Z then IV identifies ATE. In the additive separability case this means that the impact of the instrument Z on treatment intensity is the same across people with different unobservables. Interestingly, a purely randomly assigned value of the instrument Z would not be sufficient to identify ATE using IV, unless the separability condition held in the model. A key point is that with heterogeneous treatment effects identification will require stronger than usual assumptions on the model assigning individuals to different levels of treatment. In what follows, identification will rely precisely on assumptions not only on the equation of interest but also on the assignment rule and the way its stochastic structure relates to the outcome equation.

4.2. Identification Based on the Control Function Approach

The covariance restriction required for IV is probably far too strong for most problems. We now explore identification through the control function which may be more appealing. We start with the case of linear heterogeneity, given by equation (4.1) with $K = 1$. The definition of the control function is as above.

However, we extend the analysis of the earlier section on homogeneous treatment effects by replacing assumption (**A4**) with

A8. *Control function II.* There exist two real valued functions $r_0(\tilde{V}, X)$ and $r_1(\tilde{V}, X)$ such that

$$E(Y|D, X, Z) = \varphi(D, X) + Dr_1(\tilde{V}, X) + r_0(\tilde{V}, X). \tag{4.4}$$

where $\tilde{V}$ is a real valued (square integrable) function of (D, X, Z) such that the σ-field generated by (D, X, Z) is identical to the σ-field generated by $(\tilde{V}, X, Z)$. Alternatively, this expression is obtained by assuming that $E(\varepsilon_0|D, X, Z) = E(\varepsilon_0|\tilde{V}, X) = r_0(\tilde{V}, X)$ and $E(\varepsilon_1|D, X, Z) = E(\varepsilon_1|\tilde{V}, X) = r_1(\tilde{V}, X)$. The assumption is distinct from the standard orthogonality condition unless we assume that (ε_0, V) and (ε_1, V) are both conditionally independent of Z given X in which case (**A8**) holds with $\tilde{V} = V$.

A9. *Normalization*: $E(r_1(\tilde{V}, X)|X) = 0.$[13]

In addition, we will need a smoothness/support condition similar to **A1**, but now assumed to hold conditional on $(\tilde{V}, X)$.

A1′. $\varphi(D, X)$ is differentiable in D (a.s.), and the support of D conditional on $(X, \tilde{V})$ does not contain any isolated points (a.s.).

Theorem 4.3. *Assume equation (4.1) holds with $K = 1$. Under assumptions (***A5***) (rank condition),* control function II *(***A8***) the* normalization restriction

[13] To see that **A9** is only a normalisation, note that

$$\begin{gathered}\varphi + Dr + h = \\ (\varphi + DE(r|X)) + D(r - E(r|X)) + h = \\ \widetilde{\varphi} + D\widetilde{r} + \widetilde{h}.\end{gathered}$$

Note that **A9** is the appropriate normalisation for $\frac{\partial}{\partial d}\varphi$ to denote the ATE.

*(***A9***), and the smoothness and support condition (***A1$'$***),* ATE *and* TT *are identified in the heterogeneous treatment effects model presented above.*

Proof. See Appendix. ■

The analysis can be extended to higher order heterogeneity. Thus, consider the more general where $K \geq 1$. Consider

A8$'$. *Control function III.* There exist real valued functions $r_k(\tilde{V}, X)$ for $k = 0, ..., K$, such that

$$E(Y|D, Z, X) = \varphi(D, X) + \sum_{k=0}^{K} D^k r_k(\tilde{V}, X). \tag{4.5}$$

where again $\tilde{V}$ be a real valued (square integrable) function of (D, X, Z) such that the σ-field generated by (D, X, Z) is identical to the σ-field generated by $(\tilde{V}, X, Z)$. We also impose

A9$'$. *Normalization*: $E(r_k(\tilde{V}, X)|X) = 0$ for $k = 0, ..., K$.

A1$''$. $\varphi(D, X)$ is K-times differentiable in D (a.s.), and the support of D conditional on $(X, \tilde{V})$ does not contain any isolated points (a.s.).

Theorem 4.4. *Assume equation (4.1) holds with finite $K \geq 1$. Under assumptions (***A5***) (rank condition),* control function III (**A8$'$**) *the* normalization restriction *(***A9$'$***), and the smoothness and support condition (***A1$''$***),* ATE *and* TT *are identified in the heterogeneous treatment effects model presented above.*

Proof. See Appendix. ■

The case of the control function with $\tilde{V} = V$ can be directly related to the Marginal Treatment Effect of Heckman and Vytlacil (2001). Consider the case where d is a continuous scalar variable. We have that

$$\begin{aligned}\frac{\partial}{\partial d}E(Y|D=d,V=v,X=x) &= \frac{\partial}{\partial d}\varphi(d,x)+\sum_{k=1}^{K}kd^{k-1}r_k(v,x)\\ &= E(\frac{\partial}{\partial d}\varphi(d,x)|D=d,V=v,X=x)\end{aligned}$$

Thus, given the control function assumptions, a change in d holding V and X constant identifies the average effect of a change in the treatment level among those with the given values of (V,X). In this case, the derivative of $E(Y|D=d,V=v,X=x)$ with respect to d identifies the average effect of treatment for a particular subgroup, in a manner similar to the marginal treatment effect of Heckman and Vytlacil (2001).

5. Testing for the Degree of Heterogeneity

The approach we described above allows one to identify models with random higher order derivatives. In this section we derive the basis for a diagnostic informing us about the degree of heterogeneity. So, for example, if the null hypothesis is that $U_d=\varepsilon_0$ (common treatment effects model) then we can test this hypothesis by testing that $r_1(v,x)=0$ in equation (4.4). This can be repeated for higher order heterogeneity. In fact, within the control function approach this suggests a way of finding the degree of heterogeneity required.

More generally, within the control function framework we can test for the degree of heterogeneity without explicitly estimating the model. Consider the null hypothesis that the degree of heterogeneity is ℓ versus the alternative that it is $k>\ell$. Then under the null hypothesis and within the framework of the control function assumptions we must have that, for all $k > \ell$,

$$E\left[\frac{\partial^k E(Y|D=d,V=v,X=x)}{\partial d^k}\bigg|V=v\right]$$
$$= E\left\{E\left[\frac{\partial^k}{\partial d^k}E(Y|D=d,V=v,X=x)\bigg|d\right]\bigg|V=v\right\}. \quad (5.1)$$

Letting $k = \ell + 1$ for example, provides a test of the hypothesis that the degree of heterogeneity is ℓ.

To see where this expression comes from suppose the degree of heterogeneity is $k-1$, i.e., following assumption (A-8) assume that $E(Y|D=d,V=v,X=x) = \varphi(d,x) + \sum_{j=1}^{k-1} d^j r_j(v,x) + r_0(v,x)$. Then the k^{th} order derivative of $E(Y|d)$ must satisfy

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

Then taking expectations of the above with respect to d and then v we get that

$$E\left[\frac{\partial^k}{\partial d^k}E(Y|D=d,V=v,X=x)\bigg|D=d\right] = \frac{\partial^k \varphi(d)}{\partial d^k} \quad (5.2)$$

$$E\left[\frac{\partial^k}{\partial d^k}E(Y|D=d,V=v,X=x)\bigg|V=v\right] = E\left[\frac{\partial^k \varphi(d)}{\partial d^k}\bigg|V=v\right]. \quad (5.3)$$

By substituting for $\frac{\partial^k \varphi(d)}{\partial d^k}$ from equation (5.2) into equation (5.3) we obtain the expression which is the basis of our test.

6. Identification with Two Index Assignment Rules.

In many ways the assumptions that we have made up to now can be very restrictive, particularly in relation to the assignment rule, which has been assumed to have a single index structure.[14] However, many interesting economic models will not satisfy this condition, so we now consider identification in cases with more complex assignment rules as in the equations below:

$$\begin{aligned} Y_d &= \varphi(d, X) + d\varepsilon_1 + \varepsilon_0 \quad (I) \\ D &= P(X, Z) + f(X, Z)v \quad (II) \end{aligned} \tag{6.1}$$

To prove identification we will make the following assumptions:

A10 Independence and Normalizations: (v, ε_1) are jointly independent of Z given X. We also assume as a normalization rule that $E(\varepsilon_1|X, Z) = 0$ and $E(v|X, Z) = 0$. We further impose the normalization that $\mathrm{Var}(v \mid X, Z) = 1$.

A11 Exclusion: $E(\varepsilon_0|X, Z) = 0$.

Theorem 6.1. *When outcomes Y_d are given by equation (6.1.I) and the assignment rule by equation (6.1.II), the average treatment effect is identified if the following conditions hold: strong identification* **A3**, *independence* **A10** *and exclusion* **A11.**

Proof: $P(X, Z)$ and $f(X, Z)$ are immediately identified from $P(X, Z) = E(D|X, Z)$ and $f(X, Z) = \sqrt{\mathrm{Var}(D|X, Z)}$. Denote the density of D conditional on X, Z as $p(D|X, Z)$ and the density of V as $\tilde{p}(V)$. Motivated by classical instrumental variables consider the relationship

$$\frac{\partial}{\partial z}\left\{\frac{1}{f}E(Y|X = x, Z = z)\right\} =$$

$$\frac{\partial}{\partial z}\left\{\frac{1}{f}\int \varphi(t_d, x)p(t_d|X = x, Z = z)dt_d\right\} + \qquad I$$

$$\frac{\partial}{\partial z}\left\{\frac{1}{f}\int t_d r_1(t_d, x, z)p(t_d|X = x, Z = z)dt_d\right\} \quad II$$

[14] Vytlacil (2002) this is equivalent to the monotonicity assumption imposed in the LATE model of Imbens and Angrist (1995) in the context of a binary treatment.

For the purposes of identification it would be sufficient to show that under our conditions part II above is zero and then to show that the remaining expression provides a unique solution to ATE. We define $\delta = \frac{d}{f}$ with density

$$q(\delta|x,z) = \tilde{p}(\delta - \frac{P(x,z)}{f(x,z)}).$$

We now define the function ρ to satisfy the following differential equation:

$$\frac{\partial q}{\partial z} + \rho \frac{\partial q}{\partial \delta} = 0. \tag{6.2}$$

Given the structure of the assignment rule we get that $\rho = \frac{\partial}{\partial z}\left(\frac{P}{f}\right)$ and this is a function of Z and X only and not of d. We can now write the structure of II as

$$\begin{aligned} II &= \tfrac{\partial}{\partial z}\left[\textstyle\int \delta r_1(f\delta,x,z) q d\delta\right] \\ &= \textstyle\int \delta \frac{\partial r_1(f\delta,x,z)}{\partial z} q d\delta + \int \delta r_1(f\delta,x,z) \frac{\partial q}{\partial z} d\delta \\ &= \textstyle\int \delta \frac{\partial r_1(f\delta,x,z)}{\partial z} q d\delta - \rho \int \delta r_1(f\delta,x,z) \frac{\partial q}{\partial \delta} d\delta \\ &= \textstyle\int \delta \frac{\partial r_1(f\delta,x,z)}{\partial z} q d\delta + \rho \int \left[r_1(f\delta,x,z) + \delta \frac{\partial r_1(f\delta,x,z)}{\partial \delta}\right] q d\delta \\ &= \textstyle\int \delta \left\{\frac{\partial r_1(f\delta,x,z)}{\partial z} + \rho \frac{\partial r_1(f\delta,x,z)}{\partial \delta}\right\} q d\delta \end{aligned} \tag{6.3}$$

where $\rho \int r_1(f\delta,x,z) q d\delta = \frac{\rho}{f}\int r_1(t_d,x,z)p(t_d|x,z)dt_d = 0$ by assumption **A10.** A sufficient condition for II to be zero is then that the term in $\{\}$ in equation (6.3) is zero for every value of δ. This means that the conditional expectation of the error term ε_1, given D, X and Z satisfies the differential equation (6.2). This will be the case if we can write this conditional expectation as

$$r_1(f\delta,x,z) = \tilde{r}_1(\frac{d-P}{f}, x). \tag{6.4}$$

This in turn is true by assumption **A10**.

The nest step in the identification proof is to show that, given this condition, the equation

$$\begin{gathered}\tfrac{\partial}{\partial z}\left\{\tfrac{1}{f(x,z)}E(Y|X=x,Z=z)\right\} = \\ \tfrac{\partial}{\partial z}\left\{\tfrac{1}{f(x,z)}\int \varphi(t_d,x)p(t_d|X=x,Z=z)dt_d\right\}\end{gathered} \tag{6.5}$$

has a unique solution for ATE. Lack of identification would imply that there are two functions φ_1 and φ_2 which satisfy equation (6.5). For any such φ_1 and φ_2, we have that

$$\frac{\partial}{\partial z}\left\{\frac{1}{f(x,z)}\int\left(\varphi_1(t_d,x)-\varphi_2(t_d,x)\right)p(t_d|X=x,Z=z)dt_d\right\}=0$$

which implies

$$\left\{\int\left(\varphi_1(t_d,x)-\varphi_2(t_d,x)\right)p(t_d|X=x,Z=z)dt_d\right\}=c(x)f(x,z) \tag{6.6}$$

If $c(x)=0$, strong identification implies that $\varphi_1(d,x)=\varphi_2(d,x)$. Otherwise this equation implies that $f(x,z)$ is in the range of the conditional expectations operator. By the assumption of strong identification (**A3)**, we can define uniquely a function $\varphi_0(d,x)$ such that

$$\int \varphi_0(t_d,x)p(t_d|X=x,Z=z)dt_d=f(x,z) \tag{6.7}$$

which can be calculated given knowledge of $f(x,z)$. Substituting equation (6.7) into equation (6.6) we get that

$$\left\{\int\left(\varphi_1(t_d,x)-\varphi_2(t_d,x)-c\varphi_0(t_d,x)\right)p(t_d|X=x,Z=z)dt_d\right\}=0$$

which by strong identification is true if and only if

$$\varphi_1(t_d, x) - \varphi_2(t_d, x) - c(x)\varphi_0(t_d, x) = 0.$$

If $\varphi_0(t_d, x) = 0$, then ATE is identified. If $\varphi_0(t_d, x) \neq 0$, then divide through by $\varphi_0(t_d, x) \neq 0$ to obtain

$$\frac{\varphi_1(t_d, x)}{\varphi_0(t_d, x)} - \frac{\varphi_2(t_d, x)}{\varphi_0(t_d, x)} - c(x) = 0,$$

which implies that $\frac{\partial}{\partial \mathrm{d}}\left[\frac{\varphi_1}{\varphi_0}\right]$ is identified. However, since φ_0 is a known function, ATE is identified as well.■

As the reasoning of the proof demonstrates, identification is implied by weaker conditions since from equation (6.3) we can see that the term in {} brackets, weighted by the normalized treatment intensity needs to be "on average" zero. In addition, independence is just a sufficient condition for equation (6.4) to hold but not necessary. However, the more general conditions are unfamiliar and hard to interpret from an economic point of view. The simpler condition is quite familiar in the context of the control function approach, such as Heckman's selection model. Thus, we have shown that under conventional assumptions it is possible to identify ATE non parametrically even when the assignment rule does not have the usual single index structure.

7. Estimation and Implementation

In our companion paper, Florens, Heckman, Meghir and Vytlacil (2003), we develop estimation strategies that correspond to the identification strategies considered in this paper. We now provide an overview of their analysis.

7.1. Local Instrumental Variables

We start by considering *LIV*. We simplify the problem by ignoring all Xs. Estimation can be thought of as conditional on X. We suppose the existence of p instruments Z. The problem is to solve for $\frac{\partial \varphi(d)}{\partial d}$ from the set of integral equations.

$$E\left(\frac{\partial}{\partial d}\varphi(d)\middle| Z=z\right) = \frac{\frac{\partial E(Y|Z=z)}{\partial z_j}}{\frac{\partial E(D|Z=z)}{\partial z_j}} \equiv \lambda_j(z) \qquad \forall j = 1, ..., p. \tag{7.1}$$

In the presence of more than one instrument z, the problem is overidentified. This is manifested in two ways. One is the number of equations in (7.1). The other is due to the fact that $E(\frac{\partial}{\partial d}\varphi(d)|Z = z)$ is a function of "too many" variables. We solve the first problem by replacing $\lambda_j(z)$ for a weighted sum, i.e. $\lambda(z) = \Sigma_{j=1}^{p}\gamma_j(z)\lambda_j(z)$. We discuss below the optimal choice of the weights γ_j. Now we proceed to discuss the second problem for which one solution was developed in Darolles, Florens and Renault (2002).

The idea is to replace equation (7.1) by its conditional expectation, given d. Hence we get

$$E\left[E(\frac{\partial}{\partial d}\varphi(d)|Z=z)\middle| D=d\right] = E\left[\lambda(z)\middle| D=d\right] \tag{7.2}$$

This is a Fredholm type I integral equation and it is an ill posed problem. It can be regularized using the Tikhonov regularization and then a solution for $\frac{\partial}{\partial d}\varphi(d)$ can be found. In particular regularization takes place by adding $\alpha\frac{\partial}{\partial d}\varphi(d)$ on the left hand side. In the next step the expectations are replaced by their estimates. In particular on the left hand side we use kernel functions to represent the expectations, while the right hand side is estimated by kernel in a first step.

One problem with the approach in Darolles, Florens and Renault (2002) is that it involves the inversion on a matrix whose dimension is the sample size. For large data sets, such as those found in administrative sources, this may be impractical. We now suggest an alternative form of regularization.

Write the equation to be solved as

$$E\left(\frac{\partial}{\partial d}\varphi(d)\middle| Z = z\right) = \frac{1}{p}\sum_{j=1}^{p}\gamma_j\frac{\frac{\partial E(Y|Z)}{\partial z_j}}{\frac{\partial E(D|Z)}{\partial z_j}}$$

where p are the number of instruments and γ_j are known weights. We use the shorthand notation for this equation

$$A\psi = \lambda$$

where $A : \psi \to E(\psi|Z = z)$ is the linear operator mapping from the set of real square integrable functions of d $(L^2(D))$ to the set of square integrable functions on z $(L^2(Z))$, in both cases with respect to the true distribution of D and Z respectively. We define the function ψ by $\psi = \frac{\partial}{\partial d}\varphi(d)$. Finally we have defined $\lambda = \frac{1}{p}\sum_{j=1}^{p}\gamma_j\frac{\frac{\partial E(Y|Z)}{\partial z_j}}{\frac{\partial E(D|Z)}{\partial z_j}}$ which is a function we estimate directly from the data.

We define the dual operator of A to be A^* which is the operator that equates the scalar products[15]

$$< A\psi, \mu >_Z = < \psi, A^*\mu >_D$$

where μ is any square integrable function of z (with respect to the density of d). Hence A^* is an operator mapping from $L^2(Z)$ to $L^2(D)$.

[15] We need to recall the following definitions. $< a(z), b(z) >= \int a(z)b(z)f(z)dz$. Say a function is square integrable if the variance of the function is finite. Define the norm of a square integrable function, $\psi \in L^2_D$, to be $||\psi|| = \left[\int \psi^2(D)f(d)dd\right]^{1/2}$. The norm of an operator A is defined as $||A|| = \sup ||A\psi||$ where ψ is any function such that $||\psi|| \leq 1$

¿From the definition of the dual operator A^* it follows that

$$A^*\mu = E\left[\mu(Z)\middle| D\right]$$

We suppose that all expectations are replaced by kernel estimates. Clearly the problem $\widehat{A}\psi = \widehat{\lambda}$ is ill-posed. Consequently, we consider a regularized solution based on the Landweber-Fridman regularization (see Kress, 1999). According to this the regularized solution has the form

$$\widehat{\psi}^{(m_N)} = a\sum_{k=0}^{m_N}\left(I - \widehat{A}^*\widehat{A}\right)^k \widehat{A}^*\widehat{\lambda}$$

where m_N is the number of terms in the sum and depends on the sample size. The speed of convergence of the estimator depends on the way that m_N increases with the sample size.

This can be computed by using the following recursion

$$\widehat{\psi}^{(m_N)} = \left(I - a\widehat{A}^*\widehat{A}\right)\widehat{\psi}^{j-1} + a\widehat{A}^*\widehat{\lambda}$$

The parameter a is chosen so that the recursion converges and this requires that

$$0 < a < \frac{1}{||A||^2} \equiv 1.$$

One possible choice for a is 1/2. Our companion paper considers the optimal choice of the weights $\gamma_j(z)$ and of a.

7.1.1. Control Function Estimation

There are a number of ways of approaching the estimation problem in this case. One way would be to extend the Newey, Powell and Vella (1999) approach and use series estimation. We approach the problem in a different way, much in the spirit of the backfitting method we suggested for LIV in the previous section.

Under the control function assumptions the functions φ, r and h solve the following problem

$$S = \min_{\varphi,r,h} \int \left[E\left(y|d,v\right) - \left(\varphi + dr + h\right)\right]^2 \mathrm{d}P(d|z) \tag{7.3}$$

which has the following first order conditions

$$\begin{aligned} &\textstyle\int \tilde{\varphi}[E\left(y|D=d,V=v\right) - (\varphi + dr + h)]\mathrm{d}P(d|z) = 0 \quad I \\ &\textstyle\int \widetilde{r}d[E\left(y|D=d,V=v\right) - (\varphi + dr + h)]\mathrm{d}P(d|z) = 0 \quad II \\ &\textstyle\int \tilde{h}[E\left(y|D=d,V=v\right) - (\varphi + dr + h)]\mathrm{d}P(d|z) = 0 \quad III \end{aligned} \tag{7.4}$$

where $\tilde{\varphi}$, $\widetilde{r}$ and $\tilde{h}$ are any functions of d and of v respectively. In a next step we integrate over v in expression 7.4 I and over d in expressions II and III, which directly imply that

$$\begin{aligned} E(y|d) &= \varphi + dE(r|d) + E(h|d) \qquad I \\ E(dy|v) &= E(d\varphi|v) + rE(d^2|v) + E(d|v)h \quad II \\ E(y|v) &= E(\varphi|v) + rE(d|v) + h \qquad III \end{aligned} \tag{7.5}$$

The equations in (7.5) can be solved for the unknown functions φ, r and h. One way of doing this is to follow a recursive iterative solution. First we can estimate $E(y|d)$, $E(dy|v)$ and $E(y|v)$ using kernel from the data. Then, starting from an initial value of φ, we can use II and III in equations (7.5) to obtain solutions to the control functions r and h. We can then use I to update φ and we can keep iterating. However it is also possible to solve this in one shot and we demonstrate this below.

First we can use expressions II and III to eliminate h and r from I in equations (7.5). Following this we obtain

$$\begin{aligned}
&\varphi - dE\left\{\tfrac{1}{\sigma^2(v)}\left(E(d\varphi|v) - E(d|v)E(\varphi|v)\right)|d\right\} - \\
&E\left\{\tfrac{1}{\sigma^2(v)}\left(E(d^2|v)\,E(\varphi|v) - E(d|v)E(d\varphi|v)\right)|d\right\} = \\
&E(y|d) - dE\left\{\tfrac{1}{\sigma^2(v)}\left(E(dy|v) - E(d|v)E(y|v)\right)|d\right\} - \\
&E\left\{\tfrac{1}{\sigma^2(v)}\left(E(d^2|v)\,E(y|v) - E(d|v)E(dy|v)\right)|d\right\}
\end{aligned} \tag{7.6}$$

where $\sigma^2(v) = E(d^2|v) - (E(d|v))^2$. This expression can be written compactly as $(I-T)\varphi = E(y|d) - Ty$, where T is compact. This is a Fredholm type II integral equation, which can be solved directly by inverting $I-T$ on the set of functions that satisfy a normalization rule.

The procedure described above is capable of estimating the function $\varphi(d)$. However, if we are interested in estimating just the ATE parameter $\frac{\partial\varphi(d)}{\partial d}$, we can obtain a computationally simpler problem by noting that

$$\frac{\partial}{\partial d}E(Y|D=d, X=x, Z=z) = \frac{\partial}{\partial d}\varphi(d,x) + r_1(v,x) \tag{7.7}$$

The method we presented above can now be simplified to identify just the two components on the right hand side of equation (7.7). In particular, the first order conditions will have just two equations. These can either be used to derive an iterative algorithm as before or to write down a one-shot solution, which would be based on a simplified version of I and II of equation (7.5). This is computationally simpler since we do not need to estimate the function h. However we have not established whether the two approaches differ in efficiency terms.

8. Conclusions

In this paper we have considered the identification and estimation of models with a continuous endogenous variable (or in any case discrete where the levels have a cardinal interpretation, like years of education) and non-separable errors when continuous instruments are available. We have presented three methods: Instrumental Variables, Local Instrumental Variables and Control Function. These methods rely on different underlying assumptions, which we derive. We also derive conditions under which all methods are equivalent. These conditions always involve independence assumptions of the unobservables from the instruments. Our estimation strategy for all our methods are based on Kernel smoothing and the estimators are solutions of integral equations. Finally, we provide tests for the degree of heterogeneity which allows us to assess the overall specification of the model.

9. Appendix I: Proofs of theorems

Proof of Theorem 3.2

Let φ_2 and φ_1 be two functions satisfying the assumptions. Then from **A2** we get that

$$E(\varphi_2(D,X) - \varphi_1(D,X)|X,Z) \overset{a.s.}{=} 0.$$

Assumption **A3** then implies

$$\varphi_2(D,X) - \varphi_1(D,X) \overset{a.s.}{=} 0.$$

■

Proof of Lemma 3.4

The proof is stated for the case where $\varphi(d, x)$ is identified a.s.. The proof extends trivially to the more general case where $\varphi(d, x)$ is identified a.s. up to an additive function of x.

$\varphi(d, x)$ is identified a.s. by assertion. We thus need to show that if $\varphi_1(D, X) = \varphi_2(D, X)$ a.s., and both φ_1 and φ_2 satisfy condition (**A3**), then $\frac{\partial}{\partial d}\varphi_1(D, X) = \frac{\partial}{\partial d}\varphi_2(D, X)$ a.s..

Let Ω denote the set of (d, x) points such that $\varphi_1(d, x) - \varphi_2(d, x) = 0$, such that $\frac{\partial}{\partial d}\varphi_1(d, x)$ and $\frac{\partial}{\partial d}\varphi_2(d, x)$ exist, and such that d is not an isolated point of the support of D conditional on $X = x$. Ω is an intersection of sets that occur with probability one, and thus $\Pr[(D, X) \in \Omega] = 1$.

Will use proof by contradiction. Let $\Lambda = \{(d, x) : \frac{\partial}{\partial d}\varphi_1(d, x) \neq \frac{\partial}{\partial d}\varphi_2(d, x)\}$. Assume that $\Pr[(D, X) \in \Lambda] > 0$, which implies that $\Pr[(D, X) \in \Lambda \bigcap \Omega] > 0$. For any $(d, x) \in \Lambda \bigcap \Omega$, $\varphi_1(d, x) = \varphi_2(d, x)$, and the partial derivatives of each exists, so that $\frac{\partial}{\partial d}\varphi_1(d, x) \neq \frac{\partial}{\partial d}\varphi_2(d, x)$ implies that there exists a radius $r > 0$ such that $\varphi_1(d', x) \neq \varphi_2(d', x)\ \forall d' \in B(d, r) \setminus d$. d is not an isolated point of the support of D conditional on $X = x$, and thus $\Pr[D \in B(d, r) \setminus d | X = x] > 0$ so that $\Pr[\varphi_1(D, X) \neq \varphi_2(D, X) | X = x] > 0$. This holds for a set of x values with positive probability, and thus $\Pr[\varphi_1(D, X) \neq \varphi_2(D, X)] > 0$, contradicting the assumption that the two functions are equal a.s.. ■

Proof of theorem 3.5

Let φ_1 and φ_2 be two functions satisfying assumption **A4**. Then

$$E\left(\frac{\partial \varphi_1}{\partial d} - \frac{\partial \varphi_2}{\partial d} | Z = z, X = x\right) \overset{a.s.}{=} 0$$

which implies $\frac{\partial \varphi_1}{\partial d} - \frac{\partial \varphi_2}{\partial d} \overset{a.s.}{=} 0$ under the strong identification assumption **A3**.

Proof of theorem 3.7

Note first that assumption **A2** (IV) is equivalent to

$$\frac{\partial}{\partial z_j} E(U|X = x, Z = z) \overset{a.s.}{=} 0 \ \forall j \tag{9.1}$$

under smoothness conditions. Condition (9.1) implies that:

$$\begin{aligned} \frac{\partial}{\partial z_j} E(Y|X = x, Z = z) &= \frac{\partial}{\partial z_j} \int_{D^L_{x,z}}^{D^U_{xz}} \varphi(t_d, x) p(t_d|x, z) dt_d \\ &= \int_{D^L_{x,z}}^{D^U_{xz}} \varphi(t_d, x) \frac{\partial}{\partial z_j} p(t_d|x, z) dt_d \end{aligned}$$

where we used

$$\varphi(D^L_{x,z}, x) p(D^L_{x,z}|x, z) = \varphi(D^U_{x,z}, x) p(D^L_{x,z}|x, z) = 0. \tag{9.2}$$

The LIV assumption (**A4**) says that:

$$\frac{\partial}{\partial z_j} E(Y|X = x, Z = z) = \frac{\partial}{\partial z_j} P(x, z) \int_{D^L_{x,z}}^{D^U_{x,z}} \frac{\partial \varphi}{\partial t_d}(t_d, x) p(t_d|x, z) dt_d. \tag{9.3}$$

Integrating by parts and using (9.2), we can write (9.3) as

$$\frac{\partial}{\partial z_j} E(Y|X = x, Z = z) = -\frac{\partial P}{\partial z_j}(x, z) \int_{D^L_{x,z}}^{D^U_{x,z}} \varphi(t_d, x) \frac{\partial}{\partial t_d} p(t_d|x, z) dt_d \tag{9.4}$$

Then IV and LIV are equivalent if and only if (9.2) and (9.4) are equivalent, i.e.:

$$\int_{D^L_{x,z}}^{D^U_{x,z}} \varphi(t_d, x) \left\{ \frac{\partial}{\partial z_j} p(t_d|x, z) + \frac{\partial P(x, z)}{\partial z_j} \frac{\partial}{\partial t_d} p(t_d|x, z) \right\} dt_d = 0 \tag{9.5}$$

The assumption **A5** ($V \perp\!\!\!\perp Z | X$) implies that:

$$p(t_d|x,z) = \tilde{p}(t_d - P(x,z)|x,z) = \tilde{p}(t_d - P(x,z)|x) \tag{9.6}$$

where $\tilde{p}$ is the density of V given ($X = x$ and $Z = z$). Under (9.6), equation (9.5) is satisfied and the first part of the theorem is proved.

However if IV and LIV are equivalent for any φ, (9.5) is satisfied for any φ and then the term between brackets vanishes. The partial differential equations

$$\frac{\partial}{\partial z_j} p(t_d|x,z) = \frac{\partial P}{\partial z_j}(x,z) \frac{\partial}{\partial d} p(t_d|x,z) \quad \forall j \tag{9.7}$$

implies there exists $\tilde{p}$ verifying (9.6) or equivalently $V \perp\!\!\!\perp Z | X$.■

Proof of theorem 3.8

Let (φ_1, h_1) and (φ_2, h_2) be two sets of functions satisfying assumption **A6**. Then

$$\varphi_1(D,X) - \varphi_2(D,X) \overset{a.s.}{=} h_2(\tilde{V},X) - h_1(\tilde{V},X).$$

By (**A7**), this implies that $\varphi_1(D,X) - \varphi_2(D,X)$ is a.s. a function of X alone.■

Proof of theorem 3.10

Assumption **A6** with $\tilde{V} = V$ means that $E(U|D,Z,X) = E(U|V,X)$ a.s..

Then

$$\begin{aligned} E(U|Z,X) &\overset{a.s.}{=} E(E(U|D,Z,X)|Z,X) \\ &\overset{a.s.}{=} E(E(U|V,X)|Z,X) \qquad \text{Control Function} \\ &\overset{a.s.}{=} E(E(U|V,X)|X) \qquad \text{Conditional Independence} \end{aligned}$$

because $V \perp\!\!\!\perp Z|X$ (which implies $(V, X) \perp\!\!\!\perp Z|X$). Since $E(U|Z, X)$ is a.s. a function of X only we have that $E(U|Z, X) \overset{a.s.}{=} E(U|X)$

Proof of theorem 3.11 $(V, U) \perp\!\!\!\perp Z|X$ implies a. $V \perp\!\!\!\perp Z|X$ (single index assumption), b. $U \perp\!\!\!\perp Z|X$ (IV assumption) and c. $U \perp\!\!\!\perp Z|X, V$. Moreover $E(U|Z, X, D) = E(U|X, Z, V) = E(U|X, V)$ (control function assumption).

Proof of Theorem 4.1

Let φ_2 and φ_1 be two functions satisfying the assumptions. Then

$$\begin{aligned} E(Y_D - \varphi_j(D, X)|X, Z) &= E(D\epsilon_1 + \varepsilon_0|X, Z) \\ &= E((P(X, Z) + V)\epsilon_1 + \varepsilon_0|X, Z) \\ &= P(X, Z)E(\epsilon_1|X, Z) + E(V\epsilon_1|X, Z) + E(\varepsilon_0|X, Z) \\ &= E(V\epsilon_1|X), \end{aligned}$$

with the last equality following from **A2′** and **A2″**. Thus,

$$E(\varphi_2(D, X) - \varphi_1(D, X)|X, Z) \overset{a.s.}{=} M(X).$$

with $M(X) = E(\epsilon_1 V|X)$. Assumption **A3** then implies

$$\varphi_2(D, X) - \varphi_1(D, X) \overset{a.s.}{=} M(X).$$

■

Proof of Theorem 4.3

Suppose there are two sets of functions $(\varphi^1, r_1^1, r_0^1)$ and $(\varphi^2, r_1^2, r_0^2)$ such that

$$E(Y|D = d, \tilde{V} = v, X = x) =$$

$$\varphi^i(d, x) + dr_1^i(v, x) + r_0^i(v, x), \;\; i = 1, 2$$

Then

$$\left[\varphi^1(d,x)-\varphi^2(d,x)\right]+d\left[r_1^1(v,x)-r_1^2(v,x)\right]+\left[r_0^1(v,x)-r_0^2(v,x)\right]=0.$$

Given assumption **A1′**, this implies

$$\frac{\partial}{\partial d}\varphi^1(d,x)-\frac{\partial}{\partial d}\varphi^2(d,x)+\left[r_1^1(v,x)-r_1^2(v,x)\right]=0.$$

Measurable separability implies that if any function of d and x is equal to a function of v and x (a.s.) then this must be a function of x only. Hence $r_1^1(v,x)-r_1^2(v,x)$ is a function of x only. Hence,

$$r_1^1(v,x)-r_1^2(v,x)=E\left[r_1^1(\tilde{V},X)-r_1^2(\tilde{V},X)|X=x\right].$$

The above is equal to zero under **A9**. Hence,

$$\frac{\partial}{\partial d}\varphi^1(d,x)=\frac{\partial}{\partial d}\varphi^2(d,x).$$

and thus ATE is identified. Since $r_1^1(v,X)=r_1^2(v,X)$, we have $\frac{\partial}{\partial d}\varphi_1+E[r_1^1(v,X)|X,d]=\frac{\partial}{\partial d}\varphi_2+E[r_1^2(v,X)|X,d]$ and thus TT is identified as well. ■

Proof of Theorem 4.4

Suppose 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

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

Then

$$\left[\varphi^1(d,x)-\varphi^2(d,x)\right]+\sum_{k=0}^{K}d^k\left[r_k^1(v,x)-r_k^2(v,x)\right]=0. \qquad (9.8)$$

Given assumption $A1''$, this implies

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

A5 implies that if any function of d and x is equal to a function of v and x (a.s.) then this must be a function of x only. Hence, $r^1_K(v,x) - r^2_K(v,x)$ is a function of x only. Hence,

$$r^1_K(v,x) - r^2_K(v,x) = E\left[r^1_K(\tilde{V},X) - r^2_K(\tilde{V},X)|X = x\right].$$

The above is equal to zero under $\mathbf{A9}'$. Hence,

$$r^1_K(v,x) - r^2_K(v,x) \overset{a.s.}{=} 0.$$

Considering the $K-1$ derivative of equation (9.8), we have

$$\frac{\partial^{K-1}}{\partial d^{K-1}}\varphi^1(d,x) - \frac{\partial^{K-1}}{\partial d^{K-1}}\varphi^2(d,x) + \\ (K!)d\left[r^1_K(v,x) - r^2_K(v,x)\right] + ((K-1)!)\left[r^1_{K-1}(v,x) - r^2_{K-1}(v,x)\right] = 0.$$

We have already shown $r^1_K(v,x) = r^2_K(v,x)$, and thus

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

Following a parallel analysis as that used above, we can now show that $r^1_{K-1}(v,x) - r^2_{K-1}(v,x) \overset{a.s.}{=} 0$. Iterating this procedure for $k = K-2$,...,0, we have that $r^1_k(v,x) - r^2_k(v,x) \overset{a.s.}{=} 0$ for all $k = 0,...,K$. Thus, again appealing to equation (9.8), we have $\varphi^1(d,x) - \varphi^2(d,x) \overset{a.s.}{=} 0$, and thus ATE is identified. Using that $\varphi^1(d,x) - \varphi^2(d,x) \overset{a.s.}{=} 0$ and $r^1_k(v,x) - r^2_k(v,x) \overset{a.s.}{=} 0$ for all $k = 0,...,K$, we also have that $\frac{\partial}{\partial d}\varphi^1 + \sum_{k=1}^K kd^{k-1}E[r^1_k(v,X)|X,d] = \frac{\partial}{\partial d}\varphi^2 + \sum_{k=1}^K kd^{k-1}E[r^2_k(v,X)|X,d] = 0$, and thus TT is identified.■

10. References

1. Blundell, R. and J. Powell (2002), "Endogeneity in non-parametric and semiparametric regression Models," *Econometric Society World Meeting*, Seattle.

2. Darolles, S., J. P. Florens and E. Renault (2002), "Nonparametric Instrumental Regression," mimeo IDEI, Toulouse.

3. Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2003), "Estimators for the Average Treatment Effect in Nonparametric Models with Heterogeneous Returns," unpublished working paper.

4. Heckman, James J. (1979), "Sample Selection Bias as a Specification Error" *Econometrica*, Vol. Jan., 47, No. 1. , pp. 153-162.

5. Heckman, J. J., H. Ichimura, and P. Todd (1997), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies,* 65, 261-294.

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

7. Heckman, J. J. , R. LaLonde, and J. Smith (1998), "The Economics and Econometrics of Active Labor Market Programs," forthcoming, *Handbook of Labor Economics III,* O. Ashenfelter and D. Card, editors.

8. Heckman, J. J. and R. Robb (1985), "Alternative Methods for Evaluating The impact of Interventions," in *Longitudinal Analysis of Labor Market Data,* New York, NY: Wiley.

9. 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 *Drawing Inference from Self-Selected Samples*, ed. by H. Wainer. NY: Springer-Verlag, 63-107.

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

11. Heckman, J. J. and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model," *Journal of Human Resources,* 33, 974-987.

12. Heckman, J. J and E. Vytlacil (2000), "Local Instrumental Variables," NBER Working Paper No. T0252.

13. Imbens, G. and J. D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica,* 62, 467-475.

14. Imbens, G. and W. Newey (2001), "Identification and Inference in Triangular Simultaneous Equations Models without additivity," mimeo MIT and UCLA.

15. Kress, R. (1999), *Linear Integral Equations,* Springer, New York.

16. Newey, W. and J. Powell (2003), "Instrumental Variable Estimation of Nonparametric Models", *Econometrica* Volume 71: Issue 5, September 2003

17. Newey, W., J. Powell and F. Vella (1999), “Non-Parametric Estimation of Triangular Simultaneous Equations Models,” *Econometrica* 67, 565-603

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

19. Vytlacil, E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70(1): 331-41.

Instrumental Variables, Local Instrumental Variables and Control Functions[*]

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

June 4, 2002

Abstract

We consider the identification and estimation of certain parameters of interest in models with continuous endogenous variables whose impact is heterogeneous. We provide a test that allows us to assess the degree of unobserved heterogeneity. Our identification and estimation approaches use assumptions relating to the Local Instrumental Variables (LIV) approach and the control function approach.

[*] We thank participants at the Berkeley-Stanford (March 2000) workshop on non-parametric models with endogenous regressors as well as participants at the UCL empirical microeconomics workshop for useful comments. C. Meghir thank the ESRC for funding through the Centre for Fiscal policy at the IFS. All errors are our own.

[†] IDEI, Toulouse

[‡] Univesrity of Chicago

[§] IFS and UCL, c.meghir@ucl.ac.uk

[¶] Stanford University