

Nonadditive Models with Endogenous Regressors *

Guido W. Imbens[†]

First Draft: July 2005

This Draft: February 2006

Abstract

In the last fifteen years there has been much work on nonparametric identification of causal effects in settings with endogeneity. Earlier, researchers focused on linear systems with additive residuals. However, such systems are difficult to motivate by economic theory. In many cases it is precisely the nonlinearity of the system and the presence of unobserved heterogeneity in returns (and thus non-additivity in the residuals) that leads to the type of endogeneity problems that economists are concerned with. In the more recent literature researchers have attempted to characterize conditions for identification that do not rely on such functional form or homogeneity assumptions, instead relying on assumptions that are more tightly linked to economic theory. Such assumptions often include exclusion and monotonicity restrictions and (conditional) independence assumptions. In this paper I will discuss part of this literature. I will focus on a two-equation triangular (recursive) system of simultaneous equations with a single endogenous regressor and a single instrument, with the main interest in the outcome equation relating the outcome to the (endogenous) regressor of interest. The discussion will include settings with binary, continuous, and discrete regressors.

JEL Classification: C14, C21, C52

Keywords: *Identification, Nonparametric Models, Nonadditive Models, Monotonicity, Endogeneity, Instrumental Variables*

*This paper was presented as an invited lecture at the Econometric Society World Congress held in London, August 2005. Financial support for this research was generously provided through NSF grant SES 0136789 and SES 0452590. I am grateful for comments by Richard Crump, Whitney Newey, and Edward Vytlačil.

[†]Department of Economics, and Department of Agricultural and Resource Economics, University of California at Berkeley, 330 Giannini Hall, Berkeley, CA 94720-3880, and NBER. Electronic correspondence: imbens@econ.berkeley.edu, <http://elsa.berkeley.edu/users/imbens/>.

1 Introduction

In the last fifteen years there has been much work on identification of causal effects under weak conditions in settings with endogeneity. Earlier, researchers focused on linear systems with additive residuals. However, such systems are difficult to motivate by economic theory. In many cases it is the nonlinearity of the system and the presence of unobserved heterogeneity in returns (and thus non-additivity in the residuals) that leads to the type of endogeneity problems that economists are concerned with. In the more recent literature researchers have attempted to characterize conditions for identification that do not rely on such functional form or homogeneity assumptions, instead relying solely on assumptions that are more tightly linked to economic theory. Such assumptions typically include exclusion and monotonicity restrictions and (conditional) independence assumptions.

In this paper I will discuss part of this literature. I will focus on a two-equation triangular (recursive) system of simultaneous equations with a single endogenous regressor and a single instrument. Although much of the earlier literature on simultaneous equations focused on larger systems (in fact, much of the theoretical literature studied systems with an arbitrary number of endogenous regressors and an arbitrary number of instruments despite the rare occurrence of systems with more than two endogenous variables in empirical work and the practical difficulty of even finding a single credible instrument), many applications have this two-equation form and the framework is sufficiently rich for discussing the nature of the identification problems that are studied here. I focus on identification of the outcome equation relating the outcome to the regressor of interest. The latter is potentially endogenous. It is itself determined in the second equation, the choice equation, partly by a set of instruments and partly by unobserved residuals. The endogeneity of the regressor arises from the correlation between the residuals in the outcome and choice equations. A natural setting for such models is one where an economic agent chooses an action to optimize some objective function with incomplete information regarding the objective function. The discussion will include settings with binary endogenous variables (see, among others, Heckman and Robb, 1984; Manski, 1990; Imbens and Angrist, 1994; Blundell and Powell, 2003; Vytlacil, 2000) continuous endogenous regressors (Newey and Powell and Vella, 1999; Chesher, 2003; Imbens and Newey, 2002; Chernozhukov and Hansen, 2005; Darolles, Florens, and Renault, 2001; Altonji and Matzkin, 2005) and discrete regressors (Angrist and Imbens, 1995; Chesher, 2005; Das, 2005).

Such a triangular system corresponds to a special and potentially restrictive form of endogeneity, especially with a single unobserved component combined with monotonicity in the choice equation. It contrasts with a part of the literature on endogeneity where researchers have refrained from imposing any restrictions on the form of the relationship between the endogenous regressor and the instruments beyond assuming there is one, instead merely assuming independence of the instrument and the unobserved component in the outcome equation (e.g., Newey Powell, 2003; Newey, Powell, and Vella, 1999; Darolles, Florens, and Renault, 2001; Hall and Horowitz, 2003; Chernozhukov, and Hansen, 2005; Chernozhukov, Imbens and Newey, 2005). There has been little discussion regarding tradeoff between the benefit in terms of identification of making such assumptions and the cost in terms of potential misspecification.

It is also important to note that the motivation for the endogeneity considered here differs from that arising from equilibrium conditions. I will refer to the latter as *intrinsically simultaneous* models, in contrast to the *recursive* or triangular models considered here. In the leading case of the intrinsically simultaneous equations model, the supply and demand model, there are two (sets of) agents, each characterized by a relation describing quantities as a function of prices. Endogeneity of prices arises by the equilibrium condition that quantities supplied and demanded are equal. Although in linear cases the models corresponding to this type of endogeneity are essentially the same as those for the recursive models considered here, some of the assumptions considered in the current paper are more plausible in the recursive system than in the intrinsically simultaneous set up, as will be discussed in more detail below.

The form of the endogeneity I consider in this paper implies that conditioning on some unobserved variable would suffice to remove the endogeneity. One possibility is to condition directly on the unobserved component from the choice equation, but more generally there can be functions of the unobserved component that suffice to eliminate endogeneity. This approach is not directly feasible because these variables are not observed. However, in some cases it is possible to indirectly adjust for differences in these variables. Methods developed for doing so depend critically on the nature of the system, i.e., whether the endogenous regressor is binary, discrete or continuous. A benefit of this approach is that the identification results I will discuss here are constructive, and so they are closely tied to the actual methods for adjusting for endogeneity.

A major theme of the current paper, and of my work in this area more generally, is the choice of estimands. In settings with heterogeneous effects and endogenous regressors it can often be difficult to infer the effects of policies that affect all agents, or that move some agents far from their current choices. Instead it can be much easier to evaluate policies that move agents locally by eliminating endogeneity problems for some subpopulations even when the instruments are not useful for eliminating endogeneity problems for other subpopulations. For the subpopulations these instruments induce exogenous variation in the regressor that allows the researcher to infer *local* causal effects over some range of the regressor. However, the amount of variation in the instruments and the strength of their correlation with the endogenous regressor defines these subpopulations. This has led to some concern that such local effects are in general of limited interest as they do not directly correspond to specific policy parameters. However, it is important to keep in mind that if the endogeneity problem were entirely eliminated by observing the residuals that lead to the endogeneity this would not lead to additional variation in the exogenous regressors and would therefore still limit the ability to make precise inferences about the causal effects for the entire population. These limits on the identification can be addressed in two ways. One is to acknowledge the lack of identification and report ranges of values of the estimand of primary interest in a bounds approach (e.g., Manski, 1990, 2003). Alternatively, one can impose additional restrictions on the outcome equation that would allow for extrapolation. Before doing so, or in addition to doing so, it may be useful, however, to study the subpopulations for which one can (point) identify causal effects. These subpopulations will be defined in terms of the instruments and the individual's responses to them. They need not be the most interesting subpopulations from the researcher's perspective, but it is important

to understand the share of these subpopulations in the population and how they differ from other subpopulations in terms of outcome distributions that can be compared. This strategy is similar in spirit to the focus on internal validity in biomedical trials where typically the stress is on conducting careful clinical trials without requiring that these equally satisfy external validity concerns.

In the next section I will set up the basic framework. I will discuss both the potential outcome framework popularized by Rubin (1974) that is now standard in the program evaluation literature for the binary regressor case as well as the equation-based framework traditionally used in the simultaneous equations literature in econometrics following the Cowles foundation work (see Hendry and Morgan (1997) for a historical perspective on this). In Section 3 I will discuss in some detail two economic examples that lead to the type of structure that is studied in this paper. This will illustrate how the models discussed in the current paper can arise in economic settings with agents in an environment with incomplete information.

I will then discuss the role of multi-valued endogenous regressors and multi-valued instruments. I will discuss separately three cases. First the setting with a binary endogenous regressor. In that case Imbens and Angrist (1994) show that the average effect of the regressor is identified only for a subpopulation they call compliers. I then discuss the case with a continuous endogenous regressor. This case is studied in Imbens and Newey (2002) who present conditions for identification of what they call the average conditional response function in the non-additive case. Finally I discuss the case with a discrete endogenous regressor. Here I build on the work by Angrist and Imbens (1995). (A different approach to the discrete case is taken by Chesher (2005) who combines assumptions about the outcome equation with assumptions on the choice equation and focuses on local aspects of the regression function.) In the current literature these three cases have received separate treatment, often with distinct assumptions (e.g., weak monotonicity in the binary regressor case versus strict monotonicity and continuity in the continuous regressor case). I will provide some comments linking the three cases more closely, and discuss formulations of the key assumptions that are underlying all three. These assumptions include monotonicity type assumptions in both observables and unobservables, as well as smoothness assumptions. I will also discuss the role of binary versus multi-valued and continuous instruments in all three cases.

2 The Model

The basic set up I consider in this paper is the following two-equation structural model:

$$Y_i = g(X_i, \varepsilon_i), \tag{2.1}$$

$$X_i = h(Z_i, \eta_i). \tag{2.2}$$

Both equations are structural equations, describing causal relations between the right-hand side and left-hand side variables. The system is triangular or recursive rather than simultaneous, with X entering the equation determining Y , but not the other way around. This differs from the recursive form of the general simultaneous equations model (e.g., Hausman, 1983),

where the recursive nature is by construction. In contrast to the recursive form in such linear simultaneous equations models the unobserved components in (2.1) and (2.2) are potentially correlated.

The first equation, (2.1), is the equation of primary interest. I will refer to it as the *outcome* equation. It is a primitive of the model and describes the causal or behavioral relation between a scalar outcome Y_i and the scalar regressor of primary interest X_i . In the examples I consider this is a mechanical relation such as a production function, not under the control of the agent. The two arguments of this production function are the regressor of interest X_i and an unobserved component denoted by ε_i . This unobserved component can be a vector or a scalar. We will largely refrain from making assumptions concerning the dependence of this function on its arguments.

The second equation, (2.2), describes the behavioral relation between the potentially endogenous regressor X_i and a single, or set of, instruments Z_i . In the case with a binary endogenous regressor this relation is often referred to as the *participation* or *selection* equation. With a potentially multivalued regressor that is less appropriate and here I will generally refer to it as the *choice* equation. In part this terminology makes the point that this equation often describes a choice made by an economic agent, in contrast to the outcome equation which typically describes a mechanical relation such as a production function. The endogenous regressor also depends on an unobserved component η_i . Again this unobserved component can be a vector, although I will often make assumptions that essentially reduce the dimension of η_i to one. If the unobserved component is a vector of arbitrary size there is little loss of generality compared to not specifying the equation.

This type of triangular structural model, in particular in combination with some of the assumptions that will be discussed later, is less appropriate for settings where the endogeneity arises from equilibrium conditions. Such intrinsically simultaneous settings often have more than one unobserved component in the second equation that would result from the equilibrium. For example, consider a demand and supply system with the demand function

$$Q = q^d(P, X, \varepsilon), \tag{2.3}$$

and the supply function

$$Q = q^s(P, Z, \nu). \tag{2.4}$$

The reduced form for the equilibrium price depends on both unobserved components in addition to the potential instruments,

$$P^e = g(Z, X, \varepsilon, \nu). \tag{2.5}$$

If both supply and demand are additive in the unobserved component then the equilibrium price can be written in terms of a scalar unobserved component, and the model is similar to the setting in (2.1)-(2.2) that is the primary focus of this paper. Outside of such additive models it is generally impossible to write the equilibrium price in terms of a single unobserved component.

See for a more detailed discussion of such models that are intrinsically simultaneous in a non-additive setting the recent work by Chernozhukov and Hansen (2005), Chernozhukov, Imbens and Newey (2005), Benkard and Berry (2005) and Matzkin (2005).

The formulation of the model in equations (2.1) and (2.2) is common in the econometric literature on simultaneous equations with continuous endogenous regressors. In the modern literature on the binary endogenous regressor case a slightly different set up is typically used based on the potential outcomes framework developed by Rubin (1973). This amounts to writing the outcome for unit i as a function of the regressor x as $Y_i(x)$. In the equation-based model this equals $Y_i(x) = g(x, \varepsilon_i)$, with the observed outcome $Y_i = Y_i(X_i)$. (It is interesting to note that Haavelmo uses the same potential outcomes notation with the explicit distinction between x as the argument in the function and the observed value X_i in his early work on simultaneous equations, e.g., Haavelmo (1943).) Similar to the outcome function $Y_i(x)$ the value of the regressor X would be written in this framework as a function of the instrument z as $X_i(z) = h(z, \eta_i)$, with the observed outcome equal to $X_i = X_i(Z_i)$. If we do not restrict ε_i and η_i to be scalars, there is no essential loss of generality in writing the model as (2.1) and (2.2). However, restricting either ε_i or η_i to be scalar, in particular in combination with the assumption of monotonicity would be restrictive. I will return to this issue later.

I maintain the following assumption throughout the discussion.

Assumption 2.1 (INDEPENDENCE)

The instrument Z_i is independent of the pair of unobserved components (ε_i, η_i) .

In the potential outcome formulation this amounts to assumption that $\{(Y_i(x), X_i(z))\}_{z,x}$ are jointly independent of Z_i .

This assumption embodies two notions. First, and this is somewhat implicit in the formulation of the model with z not an argument of $g(\cdot)$, there is no direct causal effect of Z on Y . This is typically referred to as an exclusion restriction. Second, Z is exogenous with respect to X and Y so that the causal effect of Z on X and Y can be estimated by comparing units with different values of Z . This is like a random assignment assumption. It can be weakened by allowing for additional exogenous covariates. In the binary case Angrist, Imbens and Rubin (1996) discuss the distinctions between these assumptions in more detail.

Assumption 2.1 is a strong assumption. It is common in much of the recent identification literature that considers models with non-additive unobserved components, including both settings with binary regressors (Imbens and Angrist, 1994), and with continuous regressors (Matzkin, 2003; Chernozhukov and Hansen, 2005; Chernozhukov, Imbens and Newey, 2005; Imbens and Newey, 2002). In the next section I will discuss some examples that show how it can arise in economic models. Various weaker versions have been considered in the literature. Three of those deserve special mention. First, researchers have often assumed mean independence $\mathbb{E}[\varepsilon_i|Z_i] = 0$ and $\mathbb{E}[\eta_i|Z_i] = 0$, instead of full independence (Darolles, Florens and Renault, 2001; Newey, Powell and Vella, 1999). Unless the model is additive in η and ε , this is generally not sufficient for identification. A second potential disadvantage of the mean-independence assumption is that it ties the identifying assumptions to the functional form of the model. Second, in a very different approach Manski and Pepper (2000) discuss one-sided

versions of the exclusion restriction where the instrument may increase but not decrease the outcome. Such assumptions may be more justified in specific economic models than the full independence assumption. Third, more recently Chesher (2003, 2005) has suggested local versions of the independence assumption where at a particular value x specific quantiles of the distribution of $Y(x)$ do not vary with Z .

Endogeneity of X arises in this system of equations through the correlation of the two unobserved components ε and η . This correlation implies that X and ε are potentially correlated and therefore methods that treat X as exogenous are in general not appropriate. The exogeneity of the instrument embodied in Assumption 2.1 implies that although X and ε are potentially correlated, the correlation arises from their joint dependence on η . Hence conditional on η , the regressor X and ε are independent, and X is exogenous. An alternative way of formulating the endogeneity problem in this model is therefore that X is exogenous only conditional on an unobserved covariate. A similar set up occurs in Chamberlain (1983) where conditional on an unobserved permanent component regressors are exogenous in a panel data setting. This argument illustrates some limits to the type of identification results we can hope for. Even if we were to infer the value of η for each individual, either through direct observation or through estimation, we could not hope to learn about the relation between Y and X other than conditional on η . In other words, given knowledge of η we could identify the conditional mean $\mathbb{E}[Y|X, \eta]$ on the joint support of (η, X) . This gives a causal relation between Y and X on this joint support, but in general it will not be informative outside this support. In particular if the conditional support of $X|\eta$ varies by η we will not be able to integrate over the marginal distribution of η , and in that case we will not be able to infer variation in the conditional distribution of $Y(x)$ by x alone. There are a number of directions one can take in that case. First, one can focus on questions that do not require integration over the marginal distribution of η . This may take the form of focusing on subpopulations with overlap in the distributions of η and X , or focusing on quantiles. Second, one can obtain bounds on the effects of interest. Third, one can make additional assumptions on the conditional distribution of the outcomes given η and X that allow for the extrapolation necessary to obtain point identification.

As this discussion illustrates, this unobserved component η plays an important role in the analysis. Conditional on this variable the regressor of interest is exogenous, and this motivates two approaches to identification and estimation. First, in some cases it can be estimated consistently. The leading case of this arises in settings with X continuous and $h(z, \eta)$ strictly monotone in η (Blundell and Powell, 2003; Imbens and Newey, 2002). Given a consistent estimator for η one can then in the second stage regress Y on X controlling for $\hat{\eta}$. This is a generalization of the control function approach (e.g., Heckman and Robb, 1984; Blundell and Powell, 2004).

In other cases, however, one cannot estimate η consistently. Even in that case the conditional distribution of Y given X and Z can be interpreted as a mixture of conditional distributions of Y given X and η . The second approach to identification and estimation exploits the fact that in some cases these mixtures can be disentangled, often requiring additional assumptions. The leading case here is the binary case where a weak monotonicity condition is sufficient to identify the local average treatment effect (Imbens and Angrist, 1994). In both cases it is important

that η is a scalar with $h(z, \eta)$ (weakly) monotone in η . The first step in analyzing these cases is to note that one need not condition on η itself. It may be sufficient to condition on a function of η in order to eliminate endogeneity of the covariate. Especially when η is continuous and X takes on only few values this may simplify the problem considerably. I will call this function the *type* of a unit. It will be denoted by $T_i = T(\eta_i)$ for unit i . By conditioning on the type of a unit the endogeneity of the regressor can be eliminated. Formally,

Definition 2.1 *The type of a unit is a function $T(\eta)$ such that*

$$\varepsilon \perp X \mid T(\eta).$$

Let \mathbb{T} be the set of values that T takes on. If X and Z are discrete there are choices for the type function T that take on only a finite number of values even if η is continuous. With either Z or X continuous there may be an uncountable infinite number of types, with in the worst case $T(\eta) = \eta$. The type of a unit has some similarities to the notion of the balancing score in settings with unconfoundedness or selection on observables (e.g., Rosenbaum and Rubin, 1983). Like the propensity score the definition of a type is not unique. Under the independence assumption the unobserved component η satisfies this definition and so does any strictly monotone transformation of η . It is useful to look for the choice of type that has the least variation, the same way in the treatment evaluation literature we look for the balancing score that is the coarsest function among all possible balancing scores. In the program evaluation setting with selection on observables the solution is the propensity score. Here the optimal (coarsest) choice of the type function is any function that is constant on sets of values of η that for all z lead to the same value of X :

$$T(\eta) = T(\eta') \quad \text{if } h(z, \eta) = h(z, \eta') \quad \forall z \in \mathbb{Z}.$$

This implies that we can write the choice equation in terms of the type as $X_i = \tilde{h}(Z_i, T_i)$.

Much of the identification discussion in this paper will therefore focus on identification of

$$\beta(x, t) = \mathbb{E}[Y | X = x, T = t], \tag{2.6}$$

the conditional expectation of the outcome given the regressor of interest and the type, on the joint support of (X, T) . Because conditional on the type the regressor is exogenous, this conditional expectation has a causal interpretation as a function of x . The main issue in identification will be whether one can either infer (and estimate) the type directly and estimate $\beta(x, t)$ by regressing Y on X and the estimated type \hat{T} , or indirectly infer $\beta(x, t)$ for some values of x and some types from the joint distribution of (Z, X, Y) . There will also be some discussion relating the function $\beta(x, t)$ to policy parameters. Here the limits on the identification stemming from the restriction of the identification to the joint support of (X, T) will be important. Many policy questions will involve values of $\beta(x, t)$ outside the support of (X, T) . Such questions are only partially identified under the assumptions considered in the current discussion. To obtain point identification the researcher has to extrapolate $\beta(x, t)$ from the joint support of (X, T) to other areas. This may be more credible in some cases (e.g., if $\beta(x, t)$ is flat in t and the additional values of $\beta(x, t)$ required involve extrapolation only over t) than in others.

I will consider a couple of assumptions beyond the independence assumption that involve monotonicity of some form. The role of monotonicity assumptions in identification has recently been stressed by Imbens and Angrist (1994), Matzkin (2003), Altonji and Matzkin (2005), Athey and Imbens (2006), Chernozhukov and Hansen (2005) and others. First, I consider two monotonicity assumptions on the choice equation. These assumptions are closely related to separability conditions (e.g., Goldman and Uzawa, 1964; Pinkse, 2001).

Assumption 2.2 (WEAK MONOTONICITY IN THE INSTRUMENT)

If $h(z, \eta) > h(z', \eta)$ for some (z, z', η) , then $h(z, \eta') \geq h(z', \eta')$ for all η' .

The second monotonicity condition concerns monotonicity in the unobserved component of the choice function.

Assumption 2.3 (WEAK MONOTONICITY IN THE UNOBSERVED COMPONENT)

If $h(z, \eta) > h(z, \eta')$ for some (z, η, η') , then $h(z', \eta) \geq h(z', \eta')$ for all z' .

Sufficient, but not necessary, for the second monotonicity assumption is that η is a scalar and that $h(z, \eta)$ is nondecreasing in η . The two monotonicity assumptions are substantively very different, although closely related in some cases. Both can aid in identifying the average causal effect of changes in the covariate. Neither is directly testable.

I will also consider strict monotonicity versions of all three assumptions.

Assumption 2.4 (STRICT MONOTONICITY IN THE INSTRUMENT)

If $h(z, \eta) > h(z', \eta)$ for some (z, z', η) , then $h(z, \eta') > h(z', \eta')$ for all η' .

Assumption 2.5 (STRICT MONOTONICITY IN THE UNOBSERVED COMPONENT)

If $h(z, \eta) > h(z, \eta')$ for some (z, η, η') , then $h(z', \eta) > h(z', \eta')$ for all z' .

The strict monotonicity assumptions are particularly restrictive in settings with discrete choices where they restrict the number of support points for the instrument or the unobserved component to be equal to the number of values that the choice can take on. Additivity of the choice equation in the instruments and the unobserved component directly implies 2.4 and 2.5, but the combination of these two assumptions is still much weaker than additive separability.

These monotonicity assumptions 2.3 and 2.5 are much less plausible in settings where the endogeneity arises from equilibrium conditions. As equation (2.5) shows, with non-additive demand and supply functions the reduced form for the price is generally a nonseparable function of the two unobserved components.

3 Two Economic Examples and Some Policies of Interest

In this section I discuss two economic examples where the type of triangular systems studied in the current paper can arise. In both examples an economic agent faces a decision with the payoff depending partly on the production function that is the primary object of interest in our analysis. In the first example the agent faces a binary decision. In the second example the decision is a continuous one. There are two critical features of the examples. First, the payoff

function of the agents differs from the production function that the econometrician is interested in. This difference generates exogenous variation in the regressor. Second, the production function is non-additive in the unobserved component. It is the unobserved heterogeneity in returns that follows from this non-additivity that leads agents with identical values of observed characteristics to respond differently to the same incentives. This important role of non-additivity is highlighted by Athey and Stern (1998) in their discussion of complementarity. Note that the importance of non-additivity in these models is different from that in supply and demand models. In supply and demand models endogeneity of prices arises from equilibrium conditions. It requires neither nonlinearity nor nonadditivity. In the current setting endogeneity arises in a way similar to the endogeneity in the binary selection models developed by Heckman (1978).

I will also discuss some estimands that may be of interest in these settings and how they relate to the model specified in equations (2.1) and (2.2). Traditionally researchers have focused on estimation of the function $g(x, \varepsilon)$ itself. This can be unwieldy when the regressor takes on a large number of values. The estimands I consider here fall into two categories. The first category consists of summary measures of the effect of the endogenous regressor on the outcome of interest. These include average differences or average derivatives, for the population as a whole or for subpopulations. The second set consists of the effects of specific policies that change the incentives for individuals in choosing the value of the potentially endogenous regressor.

3.1 Two Economic Examples

Example 1: (JOB TRAINING PROGRAM)

Suppose an individual faces a decision whether or not to enroll in a job training program. Life-time discounted earnings y is a function of participation in the program $x \in \{0, 1\}$ and ability ε : $y = g(x, \varepsilon)$. Ability ε is not under the control of the individual, and not observed directly by either the individual or the econometrician. The individual chooses whether to participate by maximizing expected life-time discounted earnings minus costs associated with entering the program conditional on her information set. This information set includes a noisy signal of ability, denoted by η , and a cost shifter z . The signal for ability could be a predictor such as prior labor market history. The cost of entering the program depends on an observed cost shifter z such as the availability of training facilities nearby. Although I do not explicitly allow for this, the costs could also depend on the signal η , if merit-based financial aid is available. Hence utility is

$$U(x, z, \varepsilon) = g(x, \varepsilon) - c(x, z),$$

and the optimal choice satisfies

$$\begin{aligned} X &= \operatorname{argmax}_{x \in \{0, 1\}} \mathbb{E} \left[U(x, Z, \varepsilon) | \eta, Z \right] = \operatorname{argmax}_{x \in \{0, 1\}} \left[\mathbb{E} \left[g(x, \varepsilon) | \eta, Z \right] - c(x, Z) \right]. \\ &= \begin{cases} 1 & \text{if } \mathbb{E}[g(1, \varepsilon) | \eta, Z] - c(1, z) \geq \mathbb{E}[g(0, \varepsilon) | \eta, Z] - c(0, z) \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Thus, $X = h(Z, \eta)$, so that equations (2.1) and (2.2) are satisfied.

Let us return to the two crucial features of the set up. The first point concerns the importance of the distinction between the payoff function of the individual ($U(x, z, \varepsilon) = g(x, \varepsilon) - c(x, z)$) and the production function that is the focus of the researcher ($g(x, z)$). Suppose the individual were interested in maximizing $g(x, \varepsilon)$ without subtracting the cost $c(x, z)$. In that case Z would not be a valid instrument since it would not affect the choice. On the other hand, if the researcher is interested in the causal effect of X on the objective function of the individual, $U(x, z, \varepsilon)$, then Z is not a valid instrument because it cannot be excluded from the objective function. Validity of an instrument requires it to shift the objective function without entering the production function.

The second point is the non-additivity of the production function in its unobserved component. If the production function $g(x, \varepsilon)$ were additive in ε , the participation decision would be the solution to $\max_x g(x) - c(x, Z)$. In that case the optimal solution $X = h(Z, \eta)$ would not depend on the signal for ability η . As a result all individuals with the same level of the instrument Z would make the same choice and the instrument and regressor would be perfectly correlated. In order to generate individual variation in the endogenous regressor conditional on the instrument it is essential to have unobserved heterogeneity in the returns, that is, nonadditivity of the production function in the unobserved component.

Consider now the assumptions introduced in Section 2. The independence assumption could be plausible in this case if there is individual-level variation in the costs associated with attending the training program that are unrelated to individual characteristics that affect earnings. This may be plausible if the costs are determined by decisions taken by different agents. For example, the costs could be determined by location decisions taken by administrative units. The monotonicity conditions can both be plausible in this example. Suppose the costs are monotone in the instrument. This implies that the choice function is monotone in Z . Das (2001) discusses a number of examples where monotonicity of the choice function in the unobserved component is implied by conditions on the economic primitives using monotone comparative statics results (e.g., Milgrom and Shannon, 1994; Athey, 2002).

One could generalize this model to allow for multi-valued X , e.g., education. In that case this model is closely related to the models for educational choices such as those used by Card (2001). See also Das (2001) for a discussion of a similar model. \square

Example 2: (PRODUCTION FUNCTION)

This is a non-additive extension of the classical problem in the estimation of production functions, e.g., Mundlak (1963). Consider a production function that depends on three inputs. The first input is observable to both the firm and the econometrician, and is variable in the short run (e.g., labor). It will be denoted by x . The second input is observed only by the firm and is fixed in the short run (e.g., capital or management). It will be denoted by η . Finally, the third input is unobserved by the econometrician and unknown to the firm at the time the labor input is chosen, e.g., weather. It will be denoted by ν : thus $y = g(x, \eta, \nu)$. Note that now the unobserved component in the outcome equation, ε , consists of two elements, η and ν .

The level of the input x is chosen optimally by the firm to maximize expected profits. At the time the level of the input is chosen the firm knows the form of the production function, the

level of the capital input η and the value of a cost shifter for the labor input, e.g., an indicator of the cost of labor inputs. This cost shifter is denoted by z . Profits are the difference between production times price (normalized to equal one), and costs, which depend on the level of the input and the observed cost shifter z :

$$\pi(x, z, \eta, \nu) = g(x, \eta, \nu) - c(x, z),$$

so that the firm solves the problem

$$X = \operatorname{argmax}_x \mathbb{E}[\pi(x, Z, \eta, \nu) | \eta, Z] = \operatorname{argmax}_x \left[\mathbb{E}[g(x, \eta, \nu) | \eta] - c(x, Z) \right].$$

Thus, $X = h(Z, \eta)$, so that equations (2.1) and (2.2) are satisfied.

Again, it is crucial that there is a difference between the payoff function for the agent and the production function that is the object of interest for the researcher, and that the production function is nonadditive in the unobserved component. If, for example, $g(x, \eta, \nu)$ were additive in η , the optimal level of the input would be the solution to $\max_x g(x, \nu) - c(x, Z)$. In that case the optimal solution $X = h(Z, \eta)$ would not depend on η and all firms with the same level of the instrument Z would choose the same level of the labor input irrespective of the amount of capital. \square

3.2 Policies of Interest

Next I want to discuss some specific policies that may be of interest and how they relate to the model specified in equations (2.1) and (2.2). Traditionally researchers have focused on identification and estimation of the production function $g(x, \varepsilon)$ itself. There are two concerns with this focus. First, it may not be enough. Evaluation of specific policies often requires knowledge of the joint distribution of X and ε in addition to knowledge of $g(x, \varepsilon)$. Second, it may be too much. Identification of the entire function $g(x, \varepsilon)$ can require very strong support conditions. To avoid the second problem researchers have often reported summary statistics of the production function. A leading example of such a summary statistic in the binary regressor case is the difference in the average value of the function at the two values of the regressor, in that setting referred to as the average treatment effect, $\mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon)]$. Another approach is to report average derivatives (e.g., Powell, Stock, Stoker, 1989). Such statistics are very useful ways of summarizing the typical (e.g., some average) effect of the regressor even though they rarely correspond to the effect of a policy that may actually be considered for implementation. Such policies typically involve changing the incentives for individuals to make particular choices. Only as a limit does this involve mandating a specific level of the choice variable for all individuals. In general policies changing the incentives require researchers to estimate both the outcome equation and the choice equation. Since the choice behavior of an individual or unit is wholly determined by the value of the instrument and the type of the unit, these policy effects can often be expressed in terms of two objects, first the expected production function given the agent type,

$$\beta(x, t) = \mathbb{E}[g(x, \varepsilon) | T],$$

and second the joint distribution of the type and regressor, $f_{XT}(x, t)$.

Here I want to mention briefly three examples of parameters that may be of interest to report in such an analysis. The first two are of the summary statistic type, and the last one corresponds to a more specific policy. Blundell and Powell (2003) focus on the identification and estimation of what they label the *average structural function* (ASF, see also Chamberlain, (1983)), the average of the structural function $g(x, \varepsilon)$ over the marginal distribution of ε ,

$$\mu(x) = \int g(x, \varepsilon) F_\varepsilon(d\varepsilon). \quad (3.7)$$

This is an attractive way of summarizing the effect of the regressor, although it does not correspond to a particular policy. By iterated expectations the average structural function can also be characterized in terms of the conditional average response function:

$$\mu(x) = \int \beta(x, t) F_T(dt).$$

A second summary measure corresponds to increasing for all units the value of the input by a small amount. In the continuous regressor case the per-unit effect of such a change on average output is

$$\mathbb{E} \left[\frac{\partial g}{\partial x}(X, \varepsilon) \right] = \int \int \frac{\partial g}{\partial x}(x, \varepsilon) F_{\varepsilon|X}(d\varepsilon|x) F_X(dx) = \mathbb{E} \left[\frac{\partial \beta}{\partial x}(X, T) \right], \quad (3.8)$$

where the last equality holds by changing the order of differentiation and integration. This average derivative parameter is analogous to the average derivatives studied in Powell, Stock and Stoker (1989) in the context of exogenous regressors. Note that this average derivative is generally *not* equal to the average derivative of the average structural function,

$$\mathbb{E} \left[\frac{\partial \mu}{\partial x}(X) \right] = \int \int \frac{\partial g}{\partial x}(x, \varepsilon) F_\varepsilon(d\varepsilon) F_X(dx),$$

where the derivative of the production function is averaged over the product of the marginal distributions of ε and X rather than over their joint distribution. Equality holds if X and ε are independent ($F_{\varepsilon|X}(\varepsilon|x) = F_\varepsilon(\varepsilon)$, and thus X is exogenous), or if the derivative is constant (e.g., in the linear model).

An example of a more specific policy is the implementation of a ceiling on the value of the input at \bar{x} . This changes the optimization problem of the firm in the production function example (Example 2) to

$$X = \operatorname{argmax}_{x \leq \bar{x}} \mathbb{E} [\pi(x, Z, \eta, \nu) | \eta, Z] = \operatorname{argmax}_{x \leq \bar{x}} [\mathbb{E} [g(x, \eta, \nu) | \eta] - c(x, Z)].$$

Those firms who in the absence of this restriction would choose a value for the input that is outside the limit now choose the limit \bar{x} (under some conditions on the production and cost functions), and those firms whose optimal choice is within the limit are not affected by the policy, so that under these conditions the new input is $\ell(X) = \min(X, \bar{x})$, and the resulting average production is

$$\mathbb{E} [g(\ell(X), \eta, \nu)] = \mathbb{E} [\beta(\ell(X), T)]. \quad (3.9)$$

One example of such a policy would arise if the input is causing pollution, and the government is interested in restricting its use. Another example of such a policy is the compulsory schooling age, with the government interested in the effect such a policy would have on average earnings.

More generally one may be interested in policies that change the incentives in a way that leads all agents who currently make the choice X to make the same new choice $\ell(X)$. The average outcome that is the result from such a policy has the form $\mathbb{E}[g(\ell(X), \eta, \nu)]$. Note that even in the context of standard additive linear simultaneous equations models knowledge of the regression coefficients and knowledge of the function $\ell(X)$ would not be sufficient for the evaluation of such policies—unless X is exogenous this would also require knowledge of the joint distribution of (X, η) , not just the effect of a unit increase in X on Y .

The identification of the average effects of policies of this type can be difficult compared to (3.8) partly because the policy does not correspond to a marginal change: for some individuals the value under the new policy can be substantively far away from the value in the current environment. It can therefore be useful to define a local version of such a policy. Consider a parametrization of the policy by a parameter γ , so that under the new policy the value of the regressor for an individual whose current value is x is $\ell(x, \gamma)$, with $\ell(x, 0) = x$ corresponding to the current environment. Assuming $\ell(x, \gamma)$ is sufficiently smooth, we can focus on

$$\mathbb{E} \left[\left. \frac{\partial g}{\partial \gamma}(\ell(X, \gamma), \eta, \nu) \right|_{\gamma=0} \right] = \mathbb{E} \left[\left. \frac{\partial \beta}{\partial \gamma}(\ell(X, \gamma), T) \right|_{\gamma=0} \right], \quad (3.10)$$

the average effect of a marginal change in the incentives. For example, one may be interested in the effect of a new tax on the quantity traded in a particular market. Rather than attempt to estimate the effect of the new tax at its proposed level, it may be more credible to estimate the derivative of the quantity traded with respect to the new tax, evaluated at the current level of the tax.

4 A Binary Endogenous Regressor: Local Average Treatment Effects

In this section I discuss the case with a binary endogenous regressor. First I focus on the case where the instrument is also binary. Next I will consider the case with Z multi-valued or even continuous. This section relies heavily on the discussion in Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996).

4.1 A Binary Instrument

With both the regressor and instrument binary the four types can fully describe the set of responses to all levels of the instrument, irrespective of the cardinality of η . It is useful to list them explicitly:

$$T_i = \begin{cases} (0, 0) \text{ (never – taker)} & \text{if } h(0, \eta_i) = h(1, \eta_i) = 0, \\ (0, 1) \text{ (complier)} & \text{if } h(0, \eta_i) = 0, h(1, \eta_i) = 1, \\ (1, 0) \text{ (defier)} & \text{if } h(0, \eta_i) = 1, h(1, \eta_i) = 0, \\ (1, 1) \text{ (always – taker)} & \text{if } h(0, \eta_i) = h(1, \eta_i) = 1. \end{cases}$$

The labels nevertaker, complier, defier and alwaystaker (Angrist, Imbens and Rubin, 1996) refer to the setting of a randomized experiment with noncompliance, where the instrument is the (random) assignment to the treatment and the endogenous regressor is an indicator for the actual receipt of the treatment. Compliers are in that case individuals who (always) comply with their assignment, that is, take the treatment if assigned to it and not take it if assigned to the control group.

In this case we cannot infer the type of a unit from the observed variables of Y_i , X_i , and Z_i . To see this, consider Table 1. Each pair of observed values (Z, X) is consistent with two of

Table 1: TYPE BY OBSERVED VARIABLES

		Z_i	
		0	1
$X_i = h(Z_i, \eta_i)$	0	Nevertaker/Complier	Nevertaker/Defier
	1	Alwaystaker/Defier	Alwaystaker/Complier

the four types.

This changes if one is willing to assume monotonicity. This can be done in two ways, monotonicity in the observed (Assumption 2.2) or monotonicity in the unobserved component (Assumption 2.3) of the choice function. In the case with Z and X binary these two assumptions are both equivalent to ruling out the presence of both compliers and defiers, and it is therefore sometimes referred to as the “no-defiance” assumption (Balke and Pearl, 1994; Pearl, 2000). Suppose Assumption 2.2 holds and $h(z, \eta)$ is non-decreasing in z . Then if $h(0, \eta) = 1$, it must be that $h(1, \eta) = 1$ because $h(1, \eta) \geq h(0, \eta)$. Hence there can be no defiers. (Similarly, if $h(z, \eta)$ is non-increasing in z the presence of compliers would be ruled out.) Now suppose Assumption 2.3 holds. Suppose that there is a complier with η_0 such that $h(0, \eta_0) = 0$ and $h(1, \eta_0) = 1$. Now consider an individual with $\eta_1 > \eta_0$. Then $h(1, \eta_1) = 1$ because $h(1, \eta_1) \geq h(1, \eta_0) = 1$. For an individual with $\eta_1 < \eta_0$ it must be that $h(0, \eta_1) = 0$ because $h(0, \eta_1) \leq h(0, \eta_0) = 0$. Hence no individual can be a defier. So, again the assumption is equivalent to ruling out the presence of either compliers or defiers. Vytlacil (2000), discussing the relation between this model and the selection type models developed by Heckman (e.g., Heckman, 1978; Heckman and Robb, 1984), shows that this assumption is also equivalent to the existence of an additive latent index representation of the choice function, $h(z, \eta) = 1\{m(z) + \eta \geq 0\}$. Note that monotonicity is not a testable assumption based on the joint distribution of (Z_i, X_i) . Obviously, if we specified the assumption as requiring that $h(z, \eta)$ is nondecreasing in z , there would a testable implication, but simply requiring monotonicity does not impose any restrictions on the distribution of (Z_i, X_i) . It does impose some fairly weak conditions on the conditional distribution of Y_i given (Z_i, X_i) that stem from the mixture implications of the model. See

Imbens and Rubin (1997) and Balke and Pearl (1994).

Monotonicity is not sufficient to identify the type of an individual given the value of Z and X . Although it rules out the presence of defiers, it does not eliminate all mixtures. Table 3 shows the additional information from the monotonicity assumption. Consider individuals with

Table 2: TYPE BY OBSERVED VARIABLES

		Z_i	
		0	1
$X_i = h(Z_i, \eta_i)$	0	Nevertaker/Complier	Nevertaker
	1	Alwaysstaker	Alwaysstaker/Complier

$(Z_i, X_i) = (1, 0)$. Because of monotonicity such individuals can only be nevertakers. However, consider now individuals with $(Z_i, X_i) = (0, 0)$. Such individuals can be either compliers or alwaysstakers. We cannot infer the type of such individuals from the observed data alone. Hence a control function approach (Blundell and Powell, 2003; Heckman and Robb, 1984; Imbens and Newey, 2002) is not feasible. Even though we cannot identify the type of some units, we can indirectly adjust for differences in types. Imbens and Angrist (1994) show how the mixtures can be decomposed to obtain the average difference between $g(1, \varepsilon)$ and $g(0, \varepsilon)$ for compliers, that is units with η such that $h(1, \eta) = 1$ and $h(0, \eta) = 0$. More generally, we can identify the outcome distributions for units in this subpopulation.

The intuition is as follows. The first step is to see that we can infer the population proportions of the three remaining subpopulations, nevertakers, alwaysstakers and compliers (using the fact that the monotonicity assumption rules out the presence of defiers). Call these population shares P_t , for $t = (0, 0), (0, 1), (1, 1)$. Consider the subpopulation with $Z_i = 0$. Within this subpopulation we observe $X_i = 1$ only for alwaysstakers. Hence the conditional probability of $X_i = 1$ given $Z_i = 0$ is equal to the population share of alwaysstakers: $P_{(1,1)} = \Pr(X = 1|Z = 0)$. Similarly, in the subpopulation with $Z_i = 1$ we observe $X_i = 0$ only for nevertakers. Hence the population share of nevertakers is equal to the conditional probability of $X_i = 0$ given $Z_i = 1$: $P_{(0,0)} = \Pr(X = 0|Z = 1)$. The population share of compliers is then obtained by subtracting the population shares of nevertakers and alwaysstakers from one. The second step uses the distribution of Y given (Z, X) . We can infer the distribution of $Y_i|X_i = 0, T_i = (0, 0)$ from the subpopulation with $(Z_i, X_i) = (1, 0)$ since all these individuals are known to be nevertakers. Then we use the distribution of $Y_i|Z_i = 0, X_i = 0$. This is a mixture of the distribution of $Y_i|X_i = 0, T_i = (0, 0)$ and the distribution of $Y_i|X_i = 0, T_i = (0, 1)$, with mixture probabilities equal to the relative population shares. Since we already inferred the population shares of the nevertakers and compliers as well as the distribution of $Y_i|X_i = 0, T_i = (0, 0)$, we can obtain the conditional distribution of $Y_i|X_i = 0, T_i = (0, 1)$. Similarly we can infer the conditional

distribution of $Y_i|X_i = 1, T_i = (0, 1)$. The average difference between these two conditional distributions is the Local Average Treatment Effect or LATE introduced by Imbens and Angrist (1994):

$$\tau^{\text{LATE}} = \mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon) | h(0, \eta) = 0, h(1, \eta) = 1].$$

This implies that in this setting we can under the independence and monotonicity assumptions identify the conditional average response function $\beta(x, t)$ on the joint support of (X, T) . However, because this support is not the Cartesian product of the support of X and the support of T we cannot necessarily identify the effects of all policies of interest. Specifically, $(X, T) = (0, (1, 1))$ and $(X, T) = (1, (0, 0))$ are not in the joint support of (X, T) . As a consequence, we cannot identify the average effect of the regressor $\mathbb{E}[(g(1, \varepsilon) - g(0, \varepsilon))]$, the average of $\beta(x, t)$ over the distribution of T at all values of x , nor can we generally identify the average effect for those with $X = 1$, the average effect for the treated.

4.2 A Multi-valued Instrument

Now suppose the instrument Z_i takes on values in a set \mathbb{Z} . This improves our ability to identify the causal effects of interest by providing us with additional information to infer the type of a unit. Compared to the examples with a binary instrument one can think of the multiple instrument case as arising in two ways. First, it could be that incentives are allocated more finely. Specifically, in terms of the first example in Section 3 a binary instrument could correspond to participation costs of $Z = 10$ and $Z = 20$. A multi-valued instrument could correspond to $Z \in \{10, 12.5, 15, 17.5, 20\}$. Alternatively we can add more extreme values of the incentives, e.g., $Z \in \{0, 10, 20, 30\}$. This will have different implications for the identification results. With the former we will be able to make finer distinctions between different types of compliers, and estimate the extent of variation in the effect of the regressor for these subpopulations. With the latter, we will be able to expand the subpopulation of compliers and obtain estimates for more representative subpopulations.

First consider the case with finite set of value, or $\mathbb{Z} = \{z_1, \dots, z_K\}$. The type of a unit now consists of the set of values $T_i = (X_i(z_0), \dots, X_i(z_K))$. Without the monotonicity assumption there are now 2^K different types. Monotonicity in η reduces this to $K + 1$ different types, all of the form $T_i = (0, \dots, 0, 1, \dots, 1)$, characterized by a unique value of the instrument where the type switches from $h(z, \eta) = 0$ to $h(z, \eta) = 1$ (this transition can occur after $Z = z_K$, leading to the 1 in the $K + 1$ types). The second form of monotonicity in the instrument (Assumption 2.2) is again equivalent to monotonicity in η . To see this, suppose that monotonicity in z holds, and that there is a η_0 and $z < z'$ such that $h(z, \eta_0) = 0$ and $h(z', \eta_0) = 1$. To show that monotonicity in η holds one just needs to show that there are no η_1 such that $h(z, \eta_1) = 1$ and $h(z', \eta_1) = 1$. This follows directly from monotonicity in z . The assertion in the other direction is trivial.

Let us consider the case with a three-valued instrument under monotonicity. In that case we have four types, $T_i \in \{(0, 0, 0), (0, 0, 1), (0, 1, 1), (1, 1, 1)\}$. A simple extension of the argument for the binary instrument case shows that one can infer the population shares for the four

Table 3: TYPE BY OBSERVED VARIABLES

		Z_i		
		z_1	z_2	z_3
X_i	0	(0,0,0),(0,0,1),(0,1,1)	(0,0,0),(0,0,1)	(0,0,0)
	1	(1,1,1)	(0,1,1),(1,1,1)	(0,0,1),(0,1,1),(1,1,1)

types. Using that we can infer the distribution of $Y(0)$ for the types $(0, 0, 0)$, $(0, 0, 1)$, and $(0, 1, 1)$, and the distribution of $Y(1)$ for the types $(0, 0, 1)$, $(0, 1, 1)$, and $(1, 1, 1)$. Hence we can infer the causal effects for the two types $(0, 0, 1)$ and $(0, 1, 1)$. Both are compliers in the sense that with sufficient incentives they are willing to switch from $X = 0$ to $X = 1$. These two types of compliers differ in the amount of incentives required to participate. Another way of understanding this case with a binary regressor and a multi-valued instrument is as a set of local average treatment effects. Suppose we focus on the subpopulation with $Z \in \{z, z'\}$. Within this subpopulation the basic assumption (independence and monotonicity in the unobserved component) are satisfied if they are satisfied in the overall population. Hence we can estimate the local average treatment effect for this pair of instrument values,

$$\mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon) | h(z, \eta) = 0, h(z', \eta) = 1].$$

This holds for all pairs (z, z') , and thus this argument implies that we can estimate the local average treatment effect for any of the $K \times (K - 1)/2$ pairs of instrument values,

$$\tau_{z, z'}^{\text{LATE}} = \mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon) | h(z, \eta) = 0, h(z', \eta) = 1].$$

These $K \times (K - 1)/2$ local average treatment effects are closely related since there are only $K - 1$ different types of compliers. The remaining local average treatment effects can all be written as linear combinations of the $K - 1$ basic ones. One of the most interesting of these local average treatment effects is the one corresponding to the largest set of compliers. This corresponds to the pair of instrument values with the largest effect on the regressor, (z_1, z_K) if the instrument values are ordered to satisfy monotonicity.

Having a multi-valued instrument helps in two ways if one maintains the independence and monotonicity assumptions. It allows us to estimate separate causal effects for different subpopulations, thus giving the researcher information to assess the amount of heterogeneity in the treatment effect. Second, if the multiple values reflect stronger incentives to participate or not in the activity, they will increase the size of the subpopulation for which we can identify causal effects. With a large number of values for the instrument there are more complier types. One interesting limit is the case with a scalar continuous instrument. If $h(z, \eta)$ is monotone in both z and η , and right continuous in z , we can define the limit of the local average treatment

effect as

$$\mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon) | h(z, \eta) = 1, \lim_{v \uparrow z} h(v, \eta) = 0],$$

the average treatment effect for units who change regressor value when the instrument equals z . In the context of additive latent index models with $h(z, \eta) = 1\{m(z) + \eta \geq 0\}$ this is equal to $\mathbb{E}[g(1, \varepsilon) - g(0, \varepsilon) | \eta = -m(z)]$. Heckman and Vytlacil (2001) refer to this limit of the local average treatment effect as the marginal treatment effect.

5 A Continuous Endogenous Regressor

5.1 A Scalar Unobserved Component

In this section I will discuss the case with a continuous endogenous regressor. The discussion will lean heavily on the work by Imbens and Newey (2002). Initially I will use Assumption 2.5, the strict version of the monotonicity-in-the-unobserved-component assumption: In the binary case (as well as in the general discrete case) strict monotonicity is extremely restrictive. In the binary case strict monotonicity in combination with the independence assumption would imply that η can take on only two values, and thus one can immediately identify the type of a unit because it is perfectly correlated with the endogenous regressor. Conditional on the type there is no variation in the value of the endogenous regressor and so one cannot learn about the effect of the endogenous regressor for any type. In the continuous regressor case this is very different, and strict monotonicity has no testable implications. However, this obviously does not mean that substantively strict monotonicity is not a strong assumption.

In addition to the assumptions discussed in Section 2 I will make a smoothness assumption:

Assumption 5.1 (CONTINUITY)

$h(z, \eta)$ is continuous in η and z .

In the continuous X case the type of a unit is a one-to-one function of the unobserved component η . A convenient normalization of the distribution of the type is to fix it to be uniform on the interval $[0, 1]$. In that case one can identify for each individual the type $T(\eta)$ given the strict monotonicity assumption as a function of the value of the instrument and the value of the regressor as:

$$T_i = F_{X|Z}(X_i | Z_i).$$

One can then use the inferred value of T_i to identify the conditional expectation of Y given X and T :

$$\beta(X, T) = \mathbb{E}[Y | X, T].$$

This identifies $\beta(x, t)$ for all values of x and t in their joint support. However, this support may be very limited. Suppose that the instrument is binary, $Z \in \{0, 1\}$. In that case the argument still works, but now for each value of T there are only two values of X that can be observed,

namely $h(0, \eta(T))$ and $h(1, \eta(T))$. Hence for each type T we can infer the difference between $\mathbb{E}[g(x, \varepsilon)|T]$ at $x = x_0 = h(0, \eta(T))$ and $x = x_1 = h(1, \eta(T))$. These values x_0 and x_1 where we can evaluate this expectation generally differ by type T , so the set of identified effects is very limited.

As the range of values of the instrument increases, there are for each type more and more values of the conditional regression function that can be inferred from the data. In the limit with Z continuous we can infer the value of $\mathbb{E}[Y|X, T]$ over an interval of regressor values. This may be sufficient for some policy questions, but for others there may not be sufficient variation in the instrument. For example, it may be that the instrument generates for each type of agent variation in the endogenous regressor over a different interval, making identification of the population average effect difficult.

If there is sufficient overlap in the joint distribution of X and T one can identify the marginal effect of X on Y by integrating $\beta(x, t)$ back over T . Given that the marginal distribution of T is uniform the average structural function can be written as

$$\mu(x) = \int_t \beta(x, t) dt.$$

Typically the support conditions that allow for identification of the average structural function are strong, and conditions for identification of local effects, either local in the sense of referring to a subpopulation as in the local average treatment effect, or local in the sense of moving agents only marginally away from their current choices, may be more plausible.

5.2 Multiple Unobserved Components in the Choice Equation

Throughout most of the discussion I have assumed that the unobserved component in the choice equation (2.2) can be summarized by the type of a unit, with ordering on the set of types that corresponds to an ordering on $h(z, t)$ for all z . Here I want to discuss some of the issues that arise when there are multiple unobserved components so that η_i is a vector and no such ordering need exist. In that case neither the LATE approach nor the generalized control function approach work. In that case it is still true that conditional on the vector η_i the regressor X_i is independent of the residual in the outcome equation ε_i . Hence, conditioning on η would still eliminate the endogeneity problem. However, even in the case with continuous instruments and regressors one can no longer infer the value of the (vector of) residuals. This can be illustrated in a simple example with a bivariate η .

Suppose

$$X = h(Z, \eta_1, \eta_2) = \eta_1 + \eta_2 \cdot Z,$$

with $Z \geq 0$. Suppose also that η_1 and η_2 are independent of each other and normally distributed with mean zero and unit variance. Finally, suppose that

$$\varepsilon|\eta_1, \eta_2 \sim \mathcal{N}(\eta_1 + \eta_2, 1),$$

and

$$Y_i = \alpha \cdot X_i + \varepsilon_i.$$

Now consider following the Imbens-Newey generalized control function strategy of calculating the generalized residual $\nu = F_{X|Z}(X|Z)$. The conditional distribution of $X|Z$ is normal with mean zero and variance $1 + Z^2$. Hence the generalized residual is

$$\nu_i = F_{X|Z}(X_i|Z_i) = \Phi \left(\frac{X_i}{\sqrt{1 + Z_i^2}} \right).$$

Note that by construction $\nu \perp Z$, and by assumption $\varepsilon \perp Z$. However, it is not necessarily true that $(\varepsilon, \nu) \perp Z$. Rather than work with this residual itself it is easier to work with a strictly monotone transformation, $\nu_i = X_i/\sqrt{1 + Z_i^2}$. Now consider the conditional expectation of Y given X and ν . Previously conditioning on η removed the dependence between ε and X . Here this is not the case. To see this, consider the conditional expectation of Y given $X = x_0$ and $\nu = \nu_0$. Since Z and X are one-to-one given ν , this conditional expectation is identical to the conditional expectation of Y given $Z = z_0 = \sqrt{x_0^2/\nu_0^2 - 1}$ and $\nu = \nu_0$:

$$\begin{aligned} \mathbb{E}[Y|X = x_0, \nu = \nu_0] &= \alpha x_0 + \mathbb{E}[\varepsilon|X = x_0, \nu = \nu_0] = \alpha x_0 + \mathbb{E}[\varepsilon|Z = z_0, \nu = \nu_0] \\ &= \alpha x_0 + \mathbb{E}[\eta_1 + \eta_2|Z = z_0, X/\sqrt{1 + Z^2} = \nu_0] \\ &= \alpha x_0 + \mathbb{E}[\eta_1 + \eta_2|Z = z_0, (\eta_1 + \eta_2 Z)/\sqrt{1 + Z^2} = \nu_0]. \end{aligned}$$

Now evaluate this expectation at $(z_0 = 0, \nu_0, x_0 = \nu_0)$:

$$\mathbb{E}[Y|X = \nu_0, \nu = \nu_0] = \alpha \nu_0 + \mathbb{E}[\eta_1 + \eta_2|Z = z_0, \eta_1 = \nu_0] = \nu_0 \cdot (1 + \alpha).$$

Next evaluate this expectation at $(z_0 = 1, \nu_0, x_0 = \sqrt{2}\nu_0)$:

$$\mathbb{E}[Y|X = \sqrt{2}\nu_0, \nu = \nu_0] = \alpha \sqrt{2}\nu_0 + \mathbb{E}[\eta_1 + \eta_2|Z = z_0, (\eta_1 + \eta_2)/\sqrt{2} = \nu_0] = \nu_0 \cdot \sqrt{2} \cdot (1 + \alpha).$$

Hence the expectation depends on X given ν , and therefore X is not exogenous conditional on ν .

There are two points to this example. First, it shows that the assumption of a scalar unobserved component in the choice equation that enters in a monotone way is very informative. The part of the literature that avoids specification of the choice equation potentially ignores this information if this assumption is plausible. The second point is that often economic theory has much to say about the choice equation. In many cases motivation for endogeneity comes from a specific economic model that articulates the optimization problem that the agents solve as well as specifies the components that lead to the correlation between the unobserved components in the two equations. This includes the examples in Section 3. In that case it seems useful to explore the full identifying power from the theoretical model. Even if theoretical considerations do not determine the exact functional form of the choice equation, it may suggest that it is monotone. Consider for example the two examples in Section 3. In the first case ability is the unobserved component of the outcome equation, and the signal concerning this is the unobserved component in the choice equation. It appears plausible that the choice decision is monotone in this signal.

6 A Discrete Endogenous Regressor and Discrete Instruments

In this section I will focus on the case where both Z and X are discrete. For concreteness I will assume that $\mathbb{Z} = \{0, \dots, M\}$ and $\mathbb{X} = \{0, \dots, L\}$. The discrete case is remarkably different from both the binary and continuous cases. Neither the strategy of inferring the value of η and estimating $\beta(x, \eta)$ directly, as was effective in the continuous regressor case in Section 5, nor the strategy of undoing the mixture of outcome distributions by type as was useful in the binary regressor case analyzed in Section 4 works for this setting in general. To gain insight into, and illustrate some of the difficulties of, these issues, I largely focus on the case where the endogenous regressor takes on three values, $\mathbb{X} = \{0, 1, 2\}$. Initially I consider the case with a binary instrument with $\mathbb{Z} = \{0, 1\}$, and then I will look at the case with a multi-valued and continuous instrument.

Without any monotonicity or separability restrictions the number of distinct types in the case with a binary instrument and three-valued regressor, that is the number of distinct pairs $(h(0, \eta), h(1, \eta))$ is nine: $T_i \in \{(0, 0), (0, 1), (0, 2), (1, 0), \dots, (2, 2)\}$. For some types there is a single outcome distribution under the independence or exclusion restriction. For example, for the type $T_i = (0, 0)$, there is only $f_{Y|X,T}(y|x = 0, t = (0, 0))$. For others there are two outcome distributions, e.g., for type $T_i = (0, 1)$ there are $f_{Y|X,T}(y|x = 0, t = (0, 1))$ and $f_{Y|X,T}(y|x = 1, t = (0, 1))$. This leads to a total number of outcome distributions by regressor x and type t , $f_{Y|X,T}(y|x, t)$, equal to 15. Given that one only observes data in 6 cells defined by observed values of (Z_i, X_i) , $f_{Y|X,Z}(y|x, z)$, it is clear that one cannot identify all potential outcome distributions. In the binary regressor/binary instrument case the same problem arose. Without monotonicity there were six outcome distributions $f_{Y|X,T}(y|x, t)$ (two each for defiers and compliers, and one each for nevertakers and alwaystakers) and four outcome distributions, $f_{Y|X,Z}(y|x, z)$. In that binary regressor case monotonicity ruled out the presence of defiers, reducing the number of outcome distributions by type and regressor $f_{Y|X,T}(y|x, t)$ to four. With the four outcome distributions by regressor and instrument $f_{Y|X,Z}(y|x, z)$ this led to exact identification. In the discrete regressor case it is also useful to consider some restrictions on the number of types.

In the discrete case, as in the binary case, strict monotonicity is extremely restrictive. I therefore return to the weak monotonicity assumption in the unobserved component (Assumption 2.3). This rules out the presence of some types, but there are multiple sets of five types that satisfy this monotonicity assumption. Three of them are $\{(0, 0), (0, 1), (1, 1), (1, 2), (2, 2)\}$, $\{(0, 0), (0, 1), (0, 2), (1, 2), (2, 2)\}$, and $\{(0, 0), (0, 1), (1, 1), (2, 1), (2, 2)\}$. Also requiring weak monotonicity in the instrument (Assumption 2.2) rules out the third of these sets.

To further reduce the sets of different types it may be useful to consider a smoothness assumption. In the discrete regressor case it is not meaningful to assume continuity of the choice function. Instead we assume that changing the instrument by the smallest amount possible does not lead to a jump in the response larger than the smallest jump possible. We formulate this assumption as follows: In the setting with $\mathbb{Z} = \{0, 1, \dots, M\}$ and $\mathbb{X} = \{0, 1, \dots, L\}$, this requires that $|h(z, \eta) - h(z + 1, \eta)| \leq 1$. In the continuous case this assumption is implied by continuity. In the binary case it is automatically satisfied as there can be no jumps larger than size one

in the endogenous regressor. This is potentially a restrictive assumption and it may not be reasonable in all settings. In the three-valued regressor and binary instrument case this implies that we rule out the type $T = (0, 2)$ where an individual changes the value of the regressor by two units in response to a one unit increase in the instrument.

The three assumptions combined lead to the following set of five different types:

$$T_i \in \{(0, 0), (0, 1), (1, 1), (1, 2), (2, 2)\}.$$

From the joint distribution of (X, Z) we can infer the population shares of each of these types. For example,

$$\Pr(T = (0, 0)) = \Pr(X = 0|Z = 1),$$

and

$$\Pr(T = (0, 1)) = \Pr(X = 0|Z = 0) - \Pr(X = 0|Z = 1).$$

Similarly one can identify $\Pr(T = (2, 2))$ and $\Pr(T = (1, 2))$ so that $\Pr(T = (1, 1))$ can be derived from the fact that the population shares add up to unity.

For these five types there are a total of seven outcome distributions, $f(y(0)|T = (0, 0))$, $f(y(0)|T = (0, 1))$, $f(y(1)|T = (0, 1))$, $f(y(1)|T = (1, 1))$, $f(y(1)|T = (1, 2))$, $f(y(2)|T = (1, 2))$ and $f(y(2)|T = (2, 2))$. We observe data in six cells, so we cannot identify all seven distributions. Table 4 illustrates this. It is clear that we can infer the distributions $f(y(0)|T = (0, 0))$ (from

Table 4: TYPE BY OBSERVED VARIABLES

		Z_i	
		0	1
X_i	0	(0,0),(0,1)	(0,0)
	1	(1,1),(1,2)	(0,1),(1,1)
	2	(2,2)	(1,2),(2,2)

the $(Z, X) = (1, 0)$ cell), $f(y(0)|T = (0, 1))$ (from the $(Z, X) = (0, 0)$ and $(Z, X) = (1, 0)$ cells), $f(y(2)|T = (2, 2))$ (from the $(Z, X) = (0, 2)$ cell), $f(y(2)|T = (1, 2))$ (from the $(Z, X) = (0, 2)$ and $(Z, X) = (1, 2)$ cells). However, we cannot infer the distributions of $f(y(1)|T = (0, 1))$, $f(y(1)|T = (1, 1))$, and $f(y(1)|T = (1, 2))$ from the $(Z, X) = (0, 1)$ and $(Z, X) = (1, 1)$ cells. These cells give us some restrictions on these distributions in the form of mixtures with known mixture probabilities, but not point-identification.

As a result we cannot infer the average causal effect of the regressor for any level or for any subpopulation under these assumptions in this simple case. Now let us consider what happens

if there are three values for the instrument as well. Without restrictions there are 27 different types $t = (x_0, x_1, x_2)$, for $x_0, x_1, x_2, \in \{0, 1, 2\}$. Assuming that $h(x, \eta)$ is monotone in both z and x , and making the smoothness assumption $|h(z, \eta) - h(z + 1, \eta)| \leq 1$ this is reduced to seven types:

$$T \in \{(0, 0, 0), (0, 0, 1), (0, 1, 1), (1, 1, 1), (1, 1, 2), (1, 2, 2), (2, 2, 2)\}.$$

This leads to 11 different outcome distributions. There are only 9 cells to estimate these outcome distributions from so as before it will in general still not be possible to estimate the distribution of $Y(1)$ for any of the types $T \in \{(0, 0, 1), (0, 1, 1), (1, 1, 1), (1, 1, 2), (1, 2, 2)\}$.

There are a number of approaches to deal with this problem. First, in the spirit of the work by Manski (1990, 2003), one can focus on identifying bounds for the parameters of interest. Chesher (2005) follows this approach and obtains interval estimates for quantiles of the production function at specific values for the endogenous regressor. Second, following Angrist and Imbens (1995) one can identify weighted average of unit increases in the regressor. Here the weights are partly determined by the joint distribution of the regressor and the instrument and not under the control of the researcher. Thus, one cannot simply estimate the expected value of, say, $\mathbb{E}[g(X, \varepsilon) - g(X - 1, \varepsilon)]$ over the sample distribution of X .

6.1 Bounds on the Conditional Average Response Function

One approach to identification in the case with endogenous regressors is to focus on bounds. With endogenous regressors we cannot determine the exact type of a unit. However, we can determine a range of types consistent with the observed values of the choice and instrument. Hence we can analyze the problem as one with a mismeasured regressor. See Manski and Tamer (2002).

To see how such an approach could work, note that under the assumption of monotonicity in the unobserved component one can normalize η as uniform on the interval $[0, 1]$. Under the smoothness assumption and the monotonicity assumption we can estimate the probabilities of each of the types. Hence we can identify each type with a range of values of η . Now take a particular observation with values (Z_i, X_i) . This implies that the value of η is between $F_{X|Z}(X_i - 1|Z_i)$ and $F_{X|Z}(X_i|Z_i)$. Hence we identify η up to the interval $[F_{X|Z}(X_i - 1|Z_i), F_{X|Z}(X_i|Z_i)]$. If we were to observe η we could regress Y on X and η because conditional on η X is exogenous. Now we can do the same with the provision that η is observed only to lie in an interval, fitting the Manski-Tamer framework. This would lead to an identified region for the regression function $\mathbb{E}[Y|x, \eta]$.

This approach is more likely to be useful when the endogenous regressor and the instrument take on a large number of values. In that case the type can be measured accurately and the bounds are likely to be tight. In the limit one gets to the continuous endogenous regressor case where the type can be measured without error.

6.2 Identification of Weighted Average Causal Effects

Angrist and Imbens (1995) follow a different approach. They show that under the independence assumption the standard IV estimand, the ratio of average effects of the instrument on the

outcome and on the endogenous regressor,

$$\beta^{IV}(z_0, z_1) = \frac{\mathbb{E}[Y|Z = z_1] - \mathbb{E}[Y|Z = z_0]}{\mathbb{E}[X|Z = z_1] - \mathbb{E}[X|Z = z_0]},$$

can be written as a weighted average of unit-level increases in the regressor, with the weights depending on the level of the regressor and the subpopulation:

$$\beta^{IV} = \sum_{l=1}^L \sum_{t|h(z_1,t) \geq l, h(z_0,t) < l} \lambda_{l,t} \cdot \mathbb{E}[g(l, \varepsilon) - g(l-1, \varepsilon)|T = t],$$

with

$$\lambda_{l,t} = \frac{\Pr(T = t|h(z_1, T) \geq l, h(z_0, T) < l)}{\sum_{m=1}^L \Pr(T = t|h(z_1, T) \geq m, h(z_0, T) < m)},$$

so that $\sum_{l,t} \lambda_{l,t} = 1$. If the effect of a unit increase in the regressor is the same for all individuals and the same across all levels of the treatment,

$$g(x, \varepsilon) - g(x-1, \varepsilon) = \beta_0,$$

than $\beta^{IV} = \beta_0$.

Angrist and Imbens show that although the weights $\lambda_{l,t}$ add up to unity, some of them can be negative. They then show that if $h(z, \eta)$ is monotone in z , the weights will be nonnegative. Formally, if $h(z, \eta)$ is nondecreasing in z , then $\lambda_{x,t} \geq 0$.

In this approach one is not limited to a particular pair of instrument values (z_0, z_1) . For each pair one can estimate the corresponding β_{z_0, z_1}^{IV} , each with its own set of weights. One can then combine these β_{z_0, z_1}^{IV} using any set of weights to get

$$\beta_{\omega}^{IV} = \sum_{z < z'} \omega(z, z') \beta^{IV}(z, z').$$

The weights $\omega(z, z')$ can be chosen to make β_{ω}^{IV} as close as possible to the policy parameters of interest. The variation in the weighted local average treatment effects by different choices for the weights also give some indication regarding the variation in the effect of the regressor by individual and level of the regressor. Although such effects may not be representative for large policy changes, they may be relevant for predicting the effects of small policy changes where individuals remain close to their currently optimal level of the regressor.

7 Conclusion

In this discussion I have described some of the identification issues arising in two-equation triangular simultaneous equations systems with endogenous variables. I have discussed the differences between the binary regressor case and the case with a continuous endogenous regressor. Both cases have been analyzed extensively and the requirements for identification are well understood at this point. Somewhat surprisingly the case with a discrete endogenous regressor is

much more difficult than either the discrete or binary case. Although one can identify weighted average effects, it is difficult to identify the average effect of specific changes in the regressor for either the entire population or even for specific subpopulations unless one has instruments that take on a wider range of values than what is typically seen in practice. The identification results highlight the role of monotonicity in various forms, either strict as in the case with a continuous regressor, or weak monotonicity conditions of the type that underlie the local average treatment effect for the binary regressor case.

References

- ABADIE, A., J. ANGRIST, AND G. IMBENS (2002) “Instrumental Variable Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, Vol 70, No 1, 91-117.
- ALTONJI, J., AND R. MATZKIN, (2005) “Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors,” *Econometrica*, Vol. 73, No 4, 1053-1102.
- ANGRIST, J.D., G.W. IMBENS AND D.B. RUBIN (1996), “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 91, 444-472.
- ANGRIST, J., AND G. IMBENS, (1995) “Two-Stage least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity,” *Journal of the American Statistical Association*, Vol 90, No 430, 431-442.
- ANGRIST, J. D. AND A. B. KRUEGER (1999), “Empirical Strategies in Labor Economics,” in A. Ashenfelter and D. Card eds. *Handbook of Labor Economics*, vol. 3. New York: Elsevier Science.
- ANGRIST, J., K. GRADY, AND G. IMBENS, (2000), “The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish,” *The Review of Economic Studies*, Vol. 67, July, 499-527.
- ATHEY, S., (2002), “Monotone Comparative Statics Under Uncertainty” *Quarterly Journal of Economics*, 187-223.
- ATHEY, S., AND G. IMBENS, (2006) “Identification and Inference in Nonlinear Difference-In-Difference Models,” *Econometrica*, Vol 74(2), 431-497.
- ATHEY, S., AND S. STERN, (1998), “An Empirical Framework for Testing Theories About Complementarity in Organizational Design”, NBER working paper 6600.
- BALKE, A., AND J. PEARL, (1994), “Nonparametric Bounds of Causal Effects from Partial Compliance Data,” Technical Report R-199-J, Computer Science Department, University of California, Los Angeles.
- BENKARD, L., AND S. BERRY (2005) “On the Nonparametric Identification of Nonlinear Simultaneous Equations Models: Comment on B. Brown (1983) and Roehrig (1988),” forthcoming *Econometrica*.
- BLUNDELL, R., AND J. POWELL (2003) “Endogeneity in Nonparametric and Semiparametric Regression Models,” Invited lecture at the 2000 World Congress of the econometric society, *Advances in Economics and Econometrics, Theory and Applications* Vol 2, Dewatripont, Hansen and Turnovsky (eds), Cambridge University Press, Cambridge.
- CARD, D. (2001): “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems,” *Econometrica*, Vol. 69, No. 5, 1127-1160.
- CHAMBERLAIN, G. (1983), “Panel Data Models,” in Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, Ch 22, 1247-1318.
- CHERNOZHUKOV, V., AND C. HANSEN. (2005) “An IV Model of Quantile Treatment Effects,” *Econometrica*, Vol. 73, No 1., 245-261.
- CHERNOZHUKOV, V., G. IMBENS AND W. NEWEY, (2005), “Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions,” Unpublished Manuscript.
- CHESHER, A. (2003), “Identification in Nonseparable Models,” *Econometrica*, vol 71(5), 1405-1441.

- CHESHER, A. (2005), "Nonparametric Identification Under Discrete Variation," *Econometrica*, Vol 73(5), 1525-1550.
- DAROLLES, S., J.-P., FLORENS, AND E. RENAULT, (2001), "Nonparametric Instrumental Regression," Invited lecture at the 2000 World Congress of the econometric society, *Advances in Economics and Econometrics, Theory and Applications* Vol 2, Dewatripont, Hansen and Turnovsky (eds), Cambridge University Press, Cambridge.
- DAS, M. (2005): "Instrumental Variable Estimators of Nonparametric Models with Discrete Endogenous Regressors," *Journal of Econometrics*, Vol 124, 335-361.
- DAS, M. (2001): "Monotone Comparative Statics and the Estimation of Behavioral Parameters," Working Paper, Department of Economics, Columbia University.
- GOLDMAN, S., AND H. UZAWA (1964), "A Note on Separability in Demand Analysis," *Econometrica*, Vol 32(3), 387-398.
- HAAVELMO, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, vol 11, No 1, 1-12.
- HALL, P. AND J. HOROWITZ, J., (2003), "Nonparametric Methods for Inference in the Presence of Instrumental Variables," unpublished manuscript.
- HAUSMAN, J. (1983), "Specification and Estimation of Simultaneous Equations Models," in Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, Vol. 1, Ch 7, 391-448.
- HECKMAN, J. (1978), "Sample Selection Bias as a Specification Error," *Econometrica*, Vol. 47, No. 1, 153-162.
- HECKMAN, J., AND R. ROBB, (1984), "Alternative Methods for Evaluating the Impact of Interventions," in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.
- HECKMAN, J., AND E. VYTLACIL, (2001) "Local Instrumental Variables," in Hisao, Morimune, and Powell (eds.), *Nonlinear Statistical Modeling*, Cambridge, Cambridge University Press.
- HENDRY, D., AND M. MORGAN, (1997) *The Foundation of Econometric Analysis*, Cambridge, Cambridge University Press.
- IMBENS, G., AND J. ANGRIST (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 62(2), 467-476.
- IMBENS, G., AND W. NEWEY (2002) "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", NBER Technical Working Paper 285.
- IMBENS, G., AND D. RUBIN (1997) "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, October.
- MILGROM, P., AND C. SHANNON, (1994), "Monotone Comparative Statics," *Econometrica*, Vol 62(1), 1255-1312.
- MUNDLAK, Y., (1963), "Estimation of Production Functions from a Combination of Cross-Section and Time-Series Data," in *Measurement in Economics, Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, C. Christ (ed.), 138-166.
- NEWEY, W., J. POWELL, (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, Vol 71(5), 1565-1578.

- NEWWEY, W., J. POWELL, AND F. VELLA, (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, Vol 67(3), 565-603.
- MANSKI, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.
- MANSKI, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.
- MANSKI, C., AND J. PEPPER, (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, Vol 68(4), 997-1010.
- MANSKI, C., AND E. TAMER, (2002), "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, Vol 70(2), 519 - 546.
- MATZKIN, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, Vol 71(5), 1339-1375.
- MATZKIN, R. (2005), "," Unpublished Manuscript.
- MILGROM, P., AND C. SHANNON, (1994), "Monotone Comparative Statics," *Econometrica*, Vol 62(1), 1255-1312.
- PEARL, J., (2000), *Causality: Models, Reasoning and Inference*, Cambridge, Cambridge University Press.
- PINKSE, J., (2001), "Nonparametric Regression Estimation Using Weak Separability," Unpublished Manuscript, Department of Economics, Pennsylvania State University.
- POWELL, J., J. STOCK, AND T. STOKER (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, Vol 57(6), 1403-1430.
- ROEHRIG, C. (1988), "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, Vol 56(2), 433-447.
- RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.
- ROSENBAUM, P., AND D. RUBIN, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.
- VYTLACIL, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, vol 70, No 1, 331-341.
- VYTLACIL, E. (2005), "Ordered Discrete Choice Selection Models: Equivalence, Nonequivalence, and Representation Results," *Review of Economics and Statistics*, forthcoming.