Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity*

Guido W. Imbens
Harvard University and NBER

Whitney K. Newey
M.I.T.

First Draft: March 2001
This Draft: January 2009

Abstract

This paper uses control variables to identify and estimate models with nonseparable, multidimensional disturbances. Triangular simultaneous equations models are considered, with instruments and disturbances independent and reduced form that is strictly monotonic in a scalar disturbance. Here it is shown that the conditional cumulative distribution function of the endogenous variable given the instruments is a control variable. Also, for any control variable, identification results are given for quantile, average, and policy effects. Bounds are given when a common support assumption is not satisfied. Estimators of identified objects and bounds are provided and a demand analysis empirical example given.

1 Introduction

Models with endogeneity are central in econometrics. An intrinsic feature of many of these models, often generating the endogeneity, is nonseparability in disturbances. In this paper we provide identification and estimation results for such models via control variables that, when conditioned on, make regressors and disturbances independent. We show that in a triangular simultaneous equations model, with instruments independent of disturbances and a scalar disturbance in the equation for a continuous right-hand side endogenous variable, a control variable is given by the conditional distribution function of the endogenous variable given the instruments. We also give identification and estimation results for outcome effects when *any* observable or estimable control variable is present. We focus on models where the dimension of the outcome disturbance is unspecified, allowing for individual heterogeneity and other random effects in a fully flexible way. Since a nonseparable outcome with a general disturbance is equivalent to treatment effects models, some of our identification results apply there.

We give identification and bound results for the outcome quantiles for a fixed value of the endogenous variables. Such quantiles correspond to the outcome at quantiles of the disturbance when the outcome is monotonic in a scalar disturbance. More generally, they can be used to characterize how endogenous variables affect the distribution of outcomes. Differences of these quantiles over values of the endogenous regressors correspond to quantile treatment effects as in Lehman (1974). We give identification and estimation results for these quantile effects under a common support condition. We also derive bounds on quantile treatment effects when the common support condition is not satisfied. Furthermore, we present identification results for averages of linear functionals of the outcome function. Such averages have long been of interest, because they summarize effects for a whole population. Early examples are Chamberlain's (1984) average response probability, Stoker's (1986) average derivative, and Stock's (1988) policy effect.[1] We also give identification results for average and quantile policy effects in the triangular model. In addition we provide a control variable for the triangular model where results of Blundell and Powell (2003, 2004), Wooldridge (2002), Altonji and Matzkin (2005), and Florens et. al. (2008) can be applied to identify various effects.

We employ a multi-step approach to identification and estimation. The first step is construction of the control variable. The second step consists of obtaining the conditional distribution or expectation of the outcome given the endogenous variable and the control variable. Various structural effects are then recovered by averaging over the control variable or the endogenous and control variable together.

---

[1]The average derivative results were developed independently of Altonji and Matzkin (2005), in a 2003 version of our paper.

An important feature of the triangular model is that the joint density of the endogenous variable and the control variable goes to zero at the boundary of the support of the control variable. Consequently, using nonparametric estimators with low sensitivity to edge effects may be important. We describe both locally linear and series estimators, because conventional kernel estimators are known to converge at slower rates in this setting, and give convergence rates for power series estimators. The edge effect also impacts the convergence rate of the estimators. Averaging over the control variable "upweights" the tails relative to the joint distribution. Consequently, unlike the usual results for partial means (e.g. Newey, 1994), such averages do not converge as fast as a smaller dimensional nonparametric regression. Estimators of averages over the joint distribution, do not suffer from this "upweighting," and so will converge faster. Furthermore, the convergence rate of estimators that are affected by the upweighting problem will depend on how fast the joint density goes to zero on the boundary. We find that in a Gaussian model that rate is related to the r-squared of the reduced form. In a Gaussian model this leads to convergence rates that are slower than in the additive nonparametric model of Newey, Powell, and Vella (1999).

We allow for an outcome disturbance of unspecified dimension while Chesher (2003) restricts this disturbance to be at most two dimensional. Allowing any dimension has the advantage that the interpretation of effects does not depend on exactly how many disturbances there are but has the disadvantage that it does not identify effects for particular individuals. Such a tradeoff is familiar from the treatment effects literature (e.g., Imbens and Wooldridge, 2009). To allow for general individual heterogeneity that literature has largely opted for unrestricted dimension. Also, while Chesher (2003) only needs local independence conditions to identify his local effects, we need global ones to identify the global effects we consider.

Our control variable results for the triangular model extend the work of Blundell and Powell (2003), who had extended Newey, Powell, and Vella (1999) and Pinkse (2000a) to allow for a nonseparable structural equation and a separable reduced form, to allow both the structural equation and the reduced form to be nonseparable. Chesher (2002) considers identification under index restrictions with multiple disturbances. Ma and Koenker (2006) consider identification and estimation of parametric nonseparable quantile effects using a parametric, quantile based control variable. Our triangular model results require that the endogenous variable be continuously distributed. Chesher (2005) uses the assumption of a monotonic, scalar outcome disturbance to develop bounds in a triangular model when the endogenous variable is discrete; see also Imbens (2006). Vytlacil and Yildiz (2007) give results on identification with a binary endogenous variable under instrumental variable conditions.

Imbens and Angrist (1994) and Angrist, Graddy, and Imbens (2000) also allow for nonsep-

arable disturbances of any dimension but focus on different effects than the ones we consider. Further, Chernozhukov and Hansen (2005) and Chernozhukov, Imbens and Newey (2007) consider identification and estimation of quantile effects without the triangular structure, but with restrictions on the dimension of the disturbances. Das (2001) also allows for nonseparable disturbances, but considers a single index setting with monotonicity. The independence of disturbances and instruments that we impose is stronger than the conditional mean restriction of Newey and Powell (2003), Das (2004), Darrolles, Florens, and Renault (2003), Hall and Horowitz (2005), and Blundell, Chen, and Kristensen (2007), but they require an additive disturbance.

In Section 2 of the paper we present and motivate our models. Section 3 considers identification. Section 4 describes the estimators and Section 5 gives an empirical example. Some large sample theory is presented in Section 6.

2 The Model

The model we consider has an outcome equation

$$Y = g(X, \varepsilon), \tag{2.1}$$

where X is a vector of observed variables and ε is a general disturbance vector. Here ε often represents individual hetereogeneity, which may be correlated with X because X is chosen by the agent corresponding to ε, or because X is an equilibrium outcome partially determined by ε. We focus on models where ε has unknown dimension, corresponding to a completely flexible specification of heterogeneity.

In a triangular system there is a single endogenous variable X_1 included in X, along with a vector of exogenous variables Z_1, so that $X = (X_1, Z_1')'$. There is also another vector Z_2 and a scalar disturbance η such that for $Z = (Z_1', Z_2')'$ the reduced form for X_1 is given by

$$X_1 = h(Z, \eta), \tag{2.2}$$

where $h(Z, \eta)$ is strictly monotonic in η. Equations (2.1) and (2.2) form a triangular pair of nonparametric, nonseparable, simultaneous equations. We refer to equation (2.2) as the reduced form for X_1, though it could be thought of as a structural equation in a triangular system. This model rules out a nonseparable supply and demand demand model with one disturbance per equation, because that model would generally have a reduced form with two disturbances in both supply and demand equations.

An economic example helps motivate this triangular model. For simplicity suppose Z_1 is absent, so that $X_1 = X$. Let Y denote some outcome such as firm revenue or individual

lifetime earnings, X be chosen by the individual agent, and ε represent inputs at most partially observed by agents or firms. Here $g(x, e)$ is the (educational) production function, with x and e being possible values for X and ε. The agent optimally chooses X by maximizing the expected outcome, minus the costs associated with the value of X, given her information set. Suppose the information set consists of a scalar noisy signal η of the unobserved input ε and a cost shifter Z.[2] The cost function is $c(x, z)$. Then X would be obtained as the solution to the individual choice problem

$$X = \mathrm{argmax}_x \left\{ E[g(x, \varepsilon)|\eta, Z] - c(x, Z) \right\},$$

leading to $X = h(Z, \eta)$. Thus, this economic example leads to a triangular system of the above type.

When X is schooling and Y is earnings this example corresponds to models for educational choices with heterogenous returns such as the one used by Card (2001) and Das (2001). When X is an input and Y is output, this example is a non-additive extension of a classical problem in the estimation of production functions, e.g., Mundlak (1963). Note the importance of allowing the production function $g(x, e)$ to be non-additive in e (and thus allowing the marginal returns $\frac{\partial g}{\partial x}(x, \varepsilon)$ to vary with the unobserved heterogeneity). If the objective function $g(x, e)$ were additively separable in e, so that $g(x, \varepsilon) = g_0(x) + \varepsilon$ the optimal level of x would be $\mathrm{argmax}_x\{g_0(x) + E[\varepsilon|\eta] - c(x, Z)\}$. In that case the solution X would depend on Z, but not on η, and thus X would be exogenous. Hence in these models nonseparability is important for generating endogeneity of choices.

Applying monotone comparative statics results from Milgrom and Shannon (1994) and Athey (2002), Das (2001) discusses a number of examples where monotonicity of the decision rule $h(Z, \eta)$ in the signal η is implied by conditions on the economic primitives. For example, assume that $g(x, e)$ is twice continuously differentiable. Suppose that: (i) The educational production function is strictly increasing in ability e and education x; (ii) the marginal return to formal education is strictly increasing in ability, and decreasing in education, so that $\partial g/\partial e > 0$, $\partial g/\partial x > 0$, $\partial^2 g/\partial x \partial e > 0$, and $\partial^2 g/\partial x^2 < 0$ (this would be implied by a Cobb-Douglas production function); (iii) both the cost function and the marginal cost function are increasing in education, so that $\partial c/\partial x > 0$, $\partial^2 c/\partial x^2 > 0$ and (iv) the signal η and ability ε are affiliated. Under those conditions the decision rule $h(Z, \eta)$ is monotone in η.[3]

The approach we adopt to identification and estimation is based on control variables. For the model $Y = g(X, \varepsilon)$, a control variable is *any* observable or estimable variable V satisfying

---

[2] Although we do not do so in the present example, we could allow the cost to depend on the signal η, if, for example financial aid was partly tied to test scores.

[3] Of course in this case one may wish to exploit these restrictions on the production function, as in, for example, Matzkin, 1993.

the following condition:

ASSUMPTION 1 *(Control Variable)* X *and* ε *are independent conditional on* V.

That is, X is independent of ε once we condition on the control variable V. This assumption makes changes in X causal, once we have conditioned on V, leading to identification of structural effects from the conditional distribution of Y given X and V.

In the triangular model of equations (2.1) and (2.2), it turns out that under independence of (ε, η) and Z, a control variable is the uniformly distributed $V = F_{X_1|Z}(X_1, Z) = F_\eta(\eta)$, where $F_{X_1|Z}(x_1, z)$ is the conditional CDF of X_1 given Z, and $F_\eta(t)$ is the CDF of η. Conditional independence occurs because V is a one-to-one function of η, and conditional on η the variable X_1 will only depend on Z.

THEOREM 1: *In the model of equations (2.1) and (2.2), suppose i) (Independence)* (ε, η) *and* Z *are independent; ii) (Monotonicity)* η *is a continuously distributed scalar with CDF that is strictly increasing on the support of* η *and* $h(Z, t)$ *is strictly monotonic in* t *with probability one. Then* X *and* ε *are independent conditional on* $V = F_{X_1|Z}(X_1, Z) = F_\eta(\eta)$.

In condition i) we require full independence. In the economic example of Section 2 this assumption could be plausible if the value of the instrument was chosen at a more aggregate level rather than at the level of the agents themselves. State or county level regulations could serve as such instruments, as would natural variation in economic environment conditions, in combination with random location of agents. For independence to be plausible in economic models with optimizing agents it is also important that the relation between the outcome of interest and the regressor, $g(x, \varepsilon)$, is distinct from the objective function that is maximized by the economic agent $(g(x, \varepsilon) - c(x, z)$ in the economic example from the previous section), as pointed out in Athey and Stern (1998). To make the instrument correlated with the endogenous regressor it should enter the latter (e.g., through the cost function), but to make the independence assumption plausible the instrument should not enter the former.

A scalar reduced form disturbance η and monotonicity of $h(Z, \eta)$ is essential to $F_{X_1|Z}(X_1, Z)$ being a control variable[4]. Otherwise, all of the endogeneity cannot be corrected by conditioning on identifiable variables, as discussed in Imbens (2006). Condition ii) is trivially satisfied if $h(z, t)$ is additive in t, but allows for general forms of non-additive relations. Matzkin (2003) considers nonparametric estimation of $h(z, t)$ under conditions i) and ii) in a single equation exogenous regressor framework, and Pinkse (2000b) gives a multivariate version. Das (2001)

---

[4]For scalar X we need scalar η. In a systems generalization we would need η to have the same dimension as X.

uses similar conditions to identify parameters in single index models with a single endogenous regressor.

Our identification results that are based on the control variable $V = F_{X_1|Z}(X_1, Z) = F_\eta(\eta)$ are related to the approach to identification in Chesher (2003). For simplicity suppose $z = z_2$ is a scalar, so that $x = x_1$, and let $Q_{Y|X,V}(\tau, x, v)$, $Q_{Y|X,Z}(\tau, x, z)$, and $Q_{X|Z}(\tau, z)$ be conditional quantile functions of Y given X and V, of Y given X and Z, and of X given Z, respectively. Also let ∇_a denote a partial derivative with respect to a variable a.

THEOREM 2: *If $(X, F_{X|Z}(X|Z))$ is a one-to-one transformation of (X, Z) and for z_0, x_0, τ_0, and $v_0 = F_{X|Z}(x_0, z_0)$ it is the case that $Q_{Y|X,Z}(\tau_0, x, z)$ and $F_{X|Z}(x, z)$ are continuously differentiable in (x, z) in a neighborhood of (x_0, z_0), $\nabla_z F_{X|Z}(x_0, z_0) \neq 0$, $\nabla_x F_{X|Z}(x_0, z_0) \neq 0$, then*

$$\nabla_x Q_{Y|X,V}(\tau_0, x_0, v_0) = \nabla_x Q_{Y|X,Z}(\tau_0, x_0, z_0) + \frac{\nabla_z Q_{Y|X,Z}(\tau_0, x_0, z_0)}{\nabla_z Q_{X|Z}(v_0, z_0)}. \qquad (2.3)$$

In the triangular model with two dimensional $\varepsilon = (\eta, \xi)$ for a scalar ξ, Chesher (2003) shows that the righthand side of (2.3) is equal to $\partial g(x, \varepsilon)/\partial x$ under certain local independence conditions. Theorem 2 shows that conditioning on the control variable $V = F_{X|Z}(X, Z)$ leads to the same local derivative, in the absence of the triangular model and without any independence restrictions. In this sense Chesher's (2003) approach to identification is equivalent to using the control variable $V = F_{X|Z}(X, Z)$, but without explicit specification of this variable. Explicit conditioning on V is useful for our results, which involve averaging over V, as discussed below.[5]

Many of our identification results apply more generally than just to the triangular model. They rely on Assumption 1 holding for *any* observed or estimable V rather than on the control variable from Theorem 1 for the triangular model. To emphasize this, we will state some results by referring to Assumption 1 rather than to the reduced form equation (2.2).

Identification of structural effects requires that X varies, while holding the control variable V constant. For identification of some effects we need a strong condition, that the support of the control variable V conditional on X is the same as the marginal support of V.

ASSUMPTION 2: *(Common Support) For all $X \in \mathcal{X}$, the support of V conditional on X equals the support of V.*

To explain, consider the triangular system, where $V = F_{X_1|Z}(X_1, Z)$. Here the control variable conditional on $X = x = (x_1, z_1)$ is $F_{X_1|Z}(x_1, z_1, Z_2)$. Thus, for Assumption 2 to

---

[5]One can obtain an analogous result in a linear quantile model. If the conditional quantile of Y given X and Z is linear in X and Z and the conditional quantile of X given Z is linear in Z, with residual U, then the Chesher (2003) formula equals the coefficient of X in a linear quantile regression of Y on X and U.

be satisfied, the instrumental variable Z_2 must affect $F_{X_1|Z}(x_1, z_1, Z_2)$. This is like the rank condition that is familiar from the linear simultaneous equations model. Also, for Assumption 2 it will be required that Z_2 vary sufficiently. To illustrate, suppose $z = z_2$ is a scalar and that the reduced form is $X_1 = X = \pi Z + \eta$, where η is continuously distributed with CDF $G(u)$. Then

$$F_{X|Z}(x, z) = G(x - \pi z).$$

Assume that the support of $F_{X|Z}(X, Z)$ is $[0, 1]$. Then a necessary condition for Assumption 2 is that $\pi \neq 0$, because otherwise $F_{X|Z}(x, Z)$ would be a constant. This is like the rank condition. Together with $\pi \neq 0$, the support of Z being the entire real line will be sufficient for Assumption 2. This example illustrates that Assumption 2 embodies two types of conditions, one being a rank condition and the other being a full support condition.

3 Identification

In this section we will show identification of several objects and give some bounds. We do this by giving explicit formula for objects of interest in terms of the distribution of observed data. As is well known, such explict formula imply identification in the sense of Hurwicz (1950).

A main contribution of this paper is to give new identification results for quantile, average, and policy effects. Identification results have previously been given for other objects when there is a control variable, including the average structural function (Blundell and Powell, 2003) and the local average response (Altonji and Matzkin, 2005). For these objects a contribution of Theorem 1 above is to show that $V = F_{X|Z}(X, Z)$ serves as a control variable in the triangular model of equations (2.1) and (2.2), and so can be used to identify these other functionals. We focus here on quantile, average, and policy effects.

All the results are based on the fact that for any integrable function $\Lambda(y)$,

$$E[\Lambda(Y)|X = x, V = v] = \int \Lambda(g(x, e)) F_{\varepsilon|X, V}(de|x, v) = \int \Lambda(g(x, e)) F_{\varepsilon|V}(de|v), \qquad (3.4)$$

where the second equality follows from Assumption 1. Thus, changes in x in $E[\Lambda(Y)|X = x, V = v]$ correspond to changes in x in $g(x, \varepsilon)$, i.e. are structural. This equation has an essential role in the identification and bounds results below. This equation is similar in form to equations on p. 1273 in Chamberlain (1984) and equation (2.46) of Blundell and Powell (2003).

3.1 The Quantile Structural Function

We define the quantile structural function (QSF) $q_Y(\tau, x)$ as the τth quantile of $g(x, \varepsilon)$. In this definition x is fixed and ε is what makes $g(x, \varepsilon)$ random. Note that because of the endogeneity

of X, this is in general not equal to the conditional quantile of $g(X, \varepsilon)$ conditional on $X = x$, $q_{Y|X}(\tau|x)$. In treatment effects models, $q_Y(\tau, x'') - q_Y(\tau, x')$ is the quantile treatment effect of a change in x from x' to x''; see Lehman (1974). When ε is a scalar and $g(x, \varepsilon)$ is monotonic increasing in ε, then $q_Y(\tau, x) = g(x, q_\varepsilon(\tau))$, where $q_\varepsilon(\tau)$ is the τth quantile of ε. When ε is a vector then as the value of x changes so may the values of ε that the QSF is associated with. This feature seems essential to distributional effects when the dimension of ε is unrestricted.

To show identification of the QSF, note that equation (3.4) with $\Lambda(Y) = 1(Y \leq y)$ gives

$$F_{Y|X,V}(y|x, v) = \int 1(g(x, e) \leq y) F_{\varepsilon|V}(de|v). \tag{3.5}$$

Then under Assumption 2 we can integrate over the marginal distribution of V and apply iterated expectations to obtain

$$\int F_{Y|X,V}(y|x, v) F_V(dv) = \int 1(g(x, e) \leq y) F_\varepsilon(de) = \Pr(g(x, \varepsilon) \leq y) \overset{def}{=} G(y, x). \tag{3.6}$$

Then by the definition of the QSF we have

$$q_Y(\tau, x) = G^{-1}(\tau, x). \tag{3.7}$$

Thus the QSF is the inverse of $\int F_{Y|X,V}(y|x, v) F_V(dv)$. The role of Assumption 2 is to ensure that $F_{Y|X,V}(y|x, v)$ is identified over the entire support of the marginal distribution of V. We have thus shown the following result:

THEOREM 3: *(Identification of the QSF) In a model where equation (2.1) and Assumptions 1 and 2 are satisfied, $q_Y(\tau, x)$ is identified for all $x \in \mathcal{X}$.*

3.2 Bounds for the QSF and Average Structural Function.

Assumption 2 is a rather strong assumption that may only be satisfied on a small set $\mathcal{X}$. In the empirical example below it does appear to hold but only over part of the range of X. Thus, it would be good to be able to drop Assumption 2.

When Assumption 2 is not satisfied but the structural function $g(x, e)$ is bounded one can bound the average structural function (ASF) $\mu(x) = \int g(x, e) F_\varepsilon(de)$. (Identification of $\mu(x)$ under Assumptions 1 and 2 was shown by Blundell and Powell, 2003). Let $\mathcal{V}$ denote the support of V, $\mathcal{V}(x)$ the support of V conditional on $X = x$, and $P(x) = \int_{\mathcal{V} \cap \mathcal{V}(x)^c} F_V(dV)$. Note that given $X = x$ the conditional expectation function $m(x, v) = E[Y|X = x, V = v]$ is identified for $v \in \mathcal{V}(x)$. Let $\tilde{\mu}(x)$ be the identified object

$$\tilde{\mu}(x) = \int_{\mathcal{V}(x)} m(x, v) F_V(dv).$$

THEOREM 4: *If Assumption 1 is satisfied and $B_\ell \leq g(x, e) \leq B_u$ for all x in the support of X and e in the support of ε then*

$$\mu_\ell(x) \stackrel{def}{=} \tilde{\mu}(x) + B_\ell P(x) \leq \mu(x) \leq \tilde{\mu}(x) + B_u P(x) \stackrel{def}{=} \mu_u(x),$$

and these bounds are sharp.

One example of a setting where there are bounds on $g(x, e)$ is the binary choice model where $g(x, e) \in \{0, 1\}$. In that case $B_\ell = 0$ and $B_u = 1$, so that

$$\tilde{\mu}(x) \leq \mu(x) \leq \tilde{\mu}(x) + P(x).$$

These same bounds apply to the ASF in the example considered below, where $g(x, e)$ is the share of expenditure on a commodity and so is bounded between zero and one.

There are also bounds for the QSF. Replacing Y by $1(Y \leq y)$ in the bounds for the ASF and setting $B_\ell = 0$ and $B_u = 1$ gives a lower bound $G_\ell(y, x)$ and an upper bound $G_u(y, x)$ on the integral of equation (3.6),

$$G_\ell(y, x) = \int_{\mathcal{V}(x)} \Pr(Y \leq y | X = x, V) F_V(dV), \ G_u(y, x) = G_\ell(y, x) + P(x). \qquad (3.8)$$

Assuming that Y is continuously distributed and inverting these bounds G leads to the bounds for the QSF, given by

$$q_Y^\ell(\tau, x) = \begin{cases} -\infty, \tau \leq P(x) \\ G_u^{-1}(\tau, x), \tau > P(x) \end{cases}, q_Y^u(\tau, x) = \begin{cases} G_\ell^{-1}(\tau, x), \tau < 1 - P(x), \\ +\infty, \tau \geq 1 - P(x) \end{cases}. \qquad (3.9)$$

THEOREM 5: *(Bounds for the QSF) If Assumption 1 is satisfied, then*

$$q_Y^\ell(\tau, x) \leq q_Y(\tau, x) \leq q_Y^u(\tau, x).$$

These bounds on the QSF imply bounds on the quantile treatment effects in the usual way. For values x' and x'' we have

$$q_Y^\ell(\tau, x'') - q_Y^u(\tau, x') \leq q_Y(\tau, x'') - q_Y(\tau, x') \leq q_Y^u(\tau, x'') - q_Y^\ell(\tau, x').$$

These bounds are essentially continuous versions of selection bounds in Manski (1994) and are similar to . See also Heckman and Vytlacil (2000) and Manski (2007). Blundell, Gosling, Ichimura, and Meghir (2007) have refined the Manski (1994) bounds using monotonicity and other restrictions. It should also be possible to refine the bounds here under similar conditions, although that is beyond the scope of this paper.

3.3 Average Effects

Assumption 2 is not required for identification of averages over the joint distribution of (X, ε). For example, consider the policy effect

$$\gamma = E[g(\ell(X), \varepsilon) - Y],$$

where $\ell(X)$ is some known function of X. This object is analogous to the policy effect studied by Stock (1988) in the exogenous X case. For example, one might consider a policy that imposes an upper limit $\bar{x}$ on the choice variable X in the economic model described above. Then, for a single peaked objective function it follows that the optimal choice will be $\ell(X) = \min\{X, \bar{x}\}$. Assuming there are no general equilibrium effects, the average difference of the outcome with and without the constraint will be $E[g(\ell(X), \varepsilon) - Y]$.

For this example, rather than Assumption 2, we can assume that the support of (X, V) includes the support of $(\ell(X), V)$. Then for $m(x, v) = E[Y|X = x, V = v]$, equation (3.4) with $\Lambda(Y) = Y$ gives

$$E[g(\ell(X), \varepsilon)] = E[E[g(\ell(X), \varepsilon)|X, V]] = E\left[\int g(\ell(X), e) F_{\varepsilon|V}(de|V)\right] = E[m(\ell(X), V)], \tag{3.10}$$

Then $\gamma = E[m(\ell(X), V)] - E[Y]$.

Another example is the average derivative

$$\delta = E[\partial g(X, \varepsilon)/\partial x].$$

This object is like that studied by Stoker (1986) and Powell, Stock and Stoker (1989) in the context of exogenous regressors. It summarizes the marginal effect of x on g over the population of X and ε. In a linear random coefficients model $Y = \alpha(\varepsilon) + X'\beta(\varepsilon)$, the average derivative is $\delta = E[\beta(\varepsilon)]$. If the structural function satisfies a single index restriction, with $g(x, \varepsilon) = \tilde{g}(x'\beta_0, \varepsilon)$, then δ will be proportional to β_0.

For this example we assume that the derivative of $m(x, v)$ and $g(x, \varepsilon)$ with respect to x are well defined objects, implying that X and Y are continuous random variables. Then differentiating equation (3.4) with $\Lambda(Y) = Y$ gives

$$\partial m(X, V)/\partial x = \int g_x(X, e) F_{\varepsilon|V}(de|V), \tag{3.11}$$

for $g_x(x, \varepsilon) = \partial g(x, \varepsilon)/\partial x$. Then by Assumption 1

$$\begin{aligned} \delta &= E[g_x(X, \varepsilon)] = E\left[\int g_x(X, e) F_{\varepsilon|X,V}(de|X, V)\right] \tag{3.12}\\ &= E\left[\int g_x(X, e) F_{\varepsilon|V}(de|V)\right] = \mathbb{E}\left[\frac{\partial}{\partial x} m(X, V)\right]. \end{aligned}$$

[10]

We give precise identification results for the policy function and average derivative in the following result:

THEOREM 6: *Consider a model where Assumption 1 is satisfied. If the support of $(\ell(X), V)$ is a subset of the support of (X, V) then $\gamma = E[g(\ell(X), \varepsilon) - Y]$ is identified. If (i) X has a continuous conditional distribution given V, (ii) with probability one $g(x, \varepsilon)$ is continuously differentiable in x at $x = X$; (iii) for all x and some $\Delta > 0$, $E[\int \sup_{\|x-X\| \leq \Delta} \|g_x(x, \varepsilon)\| F_{\varepsilon|V}(d\varepsilon|V)]$ exists, then $\delta = E[g_x(X, \varepsilon)]$ is identified.*

Analogous identification results can be formulated for expectations of other linear transformations of $g(x, \varepsilon)$. Let $h(x)$ denote a function of x and $T(h(\cdot), x)$ be a transformation that is linear in h. Then, assuming that the order of integration and transformation can be interchanged we obtain, from equation (3.4),

$$
\begin{aligned}
T(m(\cdot, v), x) &= T(\int g(\cdot, \varepsilon) F_{\varepsilon|V}(d\varepsilon|v), x) = \int T(g(\cdot, \varepsilon), x) F_{\varepsilon|V}(d\varepsilon|v) \\
&= \int T(g(\cdot, \varepsilon), x) F_{\varepsilon|X,V}(d\varepsilon|x, v) = E[T(g(\cdot, \varepsilon), X)|X = x, V = v].
\end{aligned}
$$

Taking expectations of both sides we find that

$$
E[T(m(\cdot, V), X)] = E[E[T(g(\cdot, \varepsilon), X)|X, V]] = E[T(g(\cdot, \varepsilon), X)].
$$

This formula leads to the following general identification result:

THEOREM 7: *In a model where Assumption 1 is satisfied, $T(m(\cdot, V), X)$ is a well defined random variable, $E[T(m(\cdot, V), X)]$ exists, and $T(\int g(\cdot, \varepsilon) F_{\varepsilon|V}(d\varepsilon|V), X) = \int T(g(\cdot, \varepsilon), X) F_{\varepsilon|V}(d\varepsilon|V)$, the object $E[T(g(\cdot, \varepsilon), X)]$ is identified.*

Theorem 6 is a special case of this result with $T(h(\cdot), x) = \partial h(x)/\partial x$ and $T(h(\cdot), x) = h(\ell(x))$.

3.4 Policy Effects in the Triangular Model

In the triangular model one can consider the effects of changes in the X equation $h(z, v)$ for X, where X is a scalar and we use the normalization $\eta = V$.[6] Let $\tilde{h}(z, v)$ denote a new function. Assuming that the change has no effect on the distribution of (ε, V) the average outcome given $Z = z$ after the change to $\tilde{h}$ would be

$$
\tilde{\theta}(z) = \int g(\tilde{h}(z, v), e) F_{\varepsilon, V}(de, dv) = \int [\int g(\tilde{h}(z, v), e) F_{\varepsilon|V}(de|v)] F_V(dv).
$$

---

[6]Steven Berry suggested the subject of this subsection. The policy effects and cost identification considered here are similar in motivation to those of Heckman and Vytlacil (2005, 2008) for their models.

From equation (3.4) with $\Lambda(Y) = Y$ we obtain

$$\tilde{\theta}(z) = \int m(\tilde{h}(z, v), v) F_V(dv).$$

An average, conditional policy effect of changing the Y equation from $h(z, v)$ to $\tilde{h}(z, v)$ is

$$\tilde{\rho}(z) = \tilde{\theta}(z) - E[Y|Z = z].$$

An unconditional policy effect of changing both h to $\tilde{h}$ and the distribution of Z to $\tilde{F}$ is

$$\tilde{\rho} = \int \tilde{\theta}(z) \tilde{F}_Z(dz) - E[Y].$$

THEOREM 8: *Consider a model where the conditions of Theorem 1 are satisfied and expectations exist. If the support of $(\tilde{h}(z, V), V)$ is contained in the suport of (X, V) then $\tilde{\rho}(z)$ is identified. Also if the support of $(\tilde{h}(z, V), V)$ is contained in the support of (X, V) for all z in the support of $\tilde{F}_Z$ then $\tilde{\rho}$ is identified.*

The previous policy effect γ is a special case of $\tilde{\rho}$ where $\tilde{h}(z, v) = \ell(h(z, v))$. Here γ is obtained by integrating over the product of the marginal distributions of (ε, V) and $F_Z(z)$, while above it is obtained by integrating over the joint distribution of (X, V, ε). This difference could lead to different estimators in practice, although it is beyond the scope of this paper to compare their properties.

One can also consider analogous quantile effects. Define the conditional CDF of Y after a change to $\tilde{h}(z, v)$, at a given z, to be

$$\tilde{J}(y, z) = \int 1(g(\tilde{h}(z, v), \varepsilon) \leq y) F_{\varepsilon, V}(d\varepsilon, dv).$$

It follows similarly to previous results that this object is identified from

$$\tilde{J}(y, z) = \int F_{Y|X, V}(y|\tilde{h}(z, v), v) F_V(dv).$$

The τth conditional quantile of Y following the change is

$$\tilde{Q}_{Y|Z}(\tau, z) = \tilde{J}^{-1}(\tau, z).$$

A quantile policy effect is

$$\tilde{Q}_{Y|Z}(\tau, z) - Q_{Y|Z}(\tau, z).$$

An unconditional policy effect that includes a change in the CDF of Z to $\tilde{F}$ is

$$\tilde{Q}_Y(\tau) - Q_Y(\tau), \tilde{Q}_Y(\tau) = \tilde{J}^{-1}(y), \tilde{J}(y) = \int \tilde{J}(y, z) \tilde{F}_Z(dz),$$

[12]

where $Q_Y(\tau)$ is the τth quantile of Y.

THEOREM 9: *Consider a model where the conditions of Theorem 1 are satisfied. If the support of $(\tilde{h}(z, V), V)$ is contained in the suport of (X, V) then $\tilde{Q}_{Y|Z}(\tau, z)$ is identified. Also if the support of $(\tilde{h}(z, V), V)$ is contained in the support of (X, V) for all z in the support of $\tilde{F}_Z$ then $\tilde{Q}_Y(\tau)$ is identified.*

In the economic model of Section 2 a possible choice of a changed $\tilde{h}(z, v)$ corresponds to a shift in the cost function. Note that for a given x we have

$$E[g(x, \varepsilon)|V = v, Z = z] = E[g(x, \varepsilon)|V = v] = m(x, v).$$

Then for an alternative cost function $\tilde{c}(x, z)$ the value $\tilde{h}(z, v)$ of x that maximizes the objective function would be

$$\tilde{h}(z, v) = \arg\max_x \{m(x, v) - \tilde{c}(x, z)\}.$$

Also, it may be desireable to specify $\tilde{c}$ relative to the cost function $c(x, z)$ identified from the data. The cost function is identified, up to an additve function of Z, by the first-order conditions

$$\frac{\partial c(X, Z)}{\partial x} = \frac{\partial m(X, h^{-1}(X, Z))}{\partial x}.$$

4   Estimation

We follow a multistep approach to estimation from i.i.d. data $(Y_i, X_i, Z_i), (i = 1, ..., n)$. The first step is estimation of the control variable observations V_i by $\hat{V}_i$. Details of this step depend on the form of the control variable. For the triangular simultaneous equations system we can form

$$\hat{V}_i = \hat{F}_{X_1|Z}(X_{1i}, Z_i),$$

where $\hat{F}_{X_1|Z}(x_1, z)$ is an estimator of the conditional CDF of X_1 given Z. These estimates can then be used to construct an estimator $\hat{F}_{Y|X,V}(y|x, v)$ of $F_{Y|X,V}(y|x, v)$ or an estimator $\hat{m}(x, v)$ of $E[Y|X, V]$ where $\hat{V}_i$ is used in place of V_i.

Estimators of objects of interest can then be formed by plugging these estimators into the formulae of Section 3, replacing integrals with sample averages. An estimator of the QSF is given by

$$\hat{q}_Y(\tau, x) = \hat{G}^{-1}(y, x); \ \hat{G}(y, x) = \sum_{i=1}^n \hat{F}_{Y|X,V}(y|x, \hat{V}_i)/n.$$

In the triangular simultaneous equations model, where V_i is known to be uniformly distributed, the sample averages can be replaced by integrals over the uniform distribution (or simulation

[13]

estimators of these integrals). Estimators of the policy effect and average derivative can be constructed by plugging in the formulae and replacing the expectation over (X, V) with a sample average, as in

$$\hat{\gamma} = \sum_{i=1}^{n} [\hat{m}(\ell(X_i), \hat{V}_i) - Y_i]/n, \hat{\delta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{m}(X_i, \hat{V}_i)}{\partial x}.$$

When Assumption 2 is not satisfied the bounds for the ASF and QSF can be estimated in a similar way. An estimator $\hat{\mathcal{V}}(x)$ of the support of V conditional on X is needed for these bounds. One can form that as

$$\hat{\mathcal{V}}(x) = \{V : \hat{f}_{V|X}(v|x) \geq \delta_n, V \in \hat{\mathcal{V}}\},$$

where δ_n is a trimming parameter and $\hat{\mathcal{V}}$ is an estimator of the support $\mathcal{V}$ of V containing all $\hat{V}_i$. In some cases $\mathcal{V}$ may be known, as for the triangular model where $\mathcal{V} = [0, 1]$. Estimates of the ASF bounds can then be formed as sample analogs,

$$\hat{\mu}_\ell(x) = \widehat{\tilde{\mu}}(x) + B_\ell \hat{P}(x), \hat{\mu}_u(x) = \widehat{\tilde{\mu}}(x) + B_u \hat{P}(x),$$

$$\widehat{\tilde{\mu}}(x) = \sum_{i=1}^{n} 1(\hat{V}_i \in \hat{\mathcal{V}}(x)) \hat{m}(x, \hat{V}_i)/n, \hat{P}(x) = \sum_{i=1}^{n} 1(\hat{V}_i \notin \hat{\mathcal{V}}(x))/n.$$

Bounds for the QSF can be formed in an analogous way. Estimates of the upper and lower bounds on $G(y, x)$ can be constructed as

$$\hat{G}_\ell(y, x) = \sum_{i=1}^{n} 1(\hat{V}_i \in \hat{\mathcal{V}}(x)) \hat{F}_{Y|X,V}(y|x, \hat{V}_i)/n, \hat{G}_u(y, x) = \hat{G}_\ell(y, x) + \hat{P}(x).$$

Assuming that $\hat{G}_\ell(y, x)$ is strictly increasing in y we then can compute the bounds for the QSF by plugging $\hat{G}_\ell(y, x)$ and $\hat{P}(x)$ into equation (3.9) to obtain

$$\hat{q}_Y^\ell(\tau, x) = \begin{cases} -\infty, \tau \leq \hat{P}(x) \\ \hat{G}_u^{-1}(\tau, x), \tau > \hat{P}(x) \end{cases}, \hat{q}_Y^u(\tau, x) = \begin{cases} \hat{G}_\ell^{-1}(\tau, x), \tau < 1 - \hat{P}(x), \\ +\infty, \tau \geq 1 - \hat{P}(x) \end{cases}.$$

To implement these estimators we need to be specific about each of their components, including the needed nonparametric regression estimators. Our choice of regression estimators is influenced by the potential importance of edge effects. For example an important feature of the triangular model is that the joint density of (X, V) may go to zero on the boundary of the support of V. For example this can easily be seen when the reduced form is linear. Suppose that $X_1 = X = Z + \eta$ and that the support of Z and η is the entire real line. Let $f_Z(z)$ and $F_\eta(t)$ be the marginal pdf and CDF of Z and η, respectively. The joint pdf of (X, V) is

$$f_{X,V}(x, v) = f_Z(x - F_\eta^{-1}(v)), 0 < t < 1.$$

[14]

Although V has a uniform marginal distribution, the joint pdf goes to zero as v goes to zero or one. In the Gaussian Z and η case, we can be specific about the rate of decrease of the joint density, as shown by the following result:

LEMMA 10: *If $X = Z + \eta$ where Z and η are normally distributed and independent, then for $R^2 = Var(Z)/[Var(X)]$ and $\bar\alpha = (1 - R^2)/R^2$, for any $B, \delta > 0$ there exists C such that for all $|x| \leq B$, $v \in [0,1]$,*

$$C[v(1-v)]^{\bar\alpha - \delta} \geq f_{X,V}(x,v) \geq C^{-1}[v(1-v)]^{\bar\alpha + \delta}.$$

Here the rate at which the joint density goes to zero at the boundary is a power of v, that increases as the reduced form r-squared falls. Thus, the lower the r-squared of the reduced form, the less tail information there is about the control variable V.

Locally linear regression estimators and series estimators are known to be less sensitive to edge effects than kernel estimators, so we focus on these. For instance, Hengarter and Linton (1996) showed that locally linear estimators have optimal convergence rates when regressor densities can go to zero, and kernel estimators do not. We will consider estimators that use the same method in both first and second stages. We also smooth out the indicator functions that appear as the left-hand side variables in these estimators, as has been suggested by ??.

To facilitate describing both steps of each estimator we establish some additional notation. For a random variable Y and a $r \times 1$ random vector W let $(Y_i, \hat{W}_i)$ denote a sample of observations where the observations on W may be estimated. We will let $\hat{a}_Y^h(w)$ denote the locally linear estimator with bandwidth h, of $E[Y|W = w]$. For a kernel function $K(u)$ let $\hat{K}_i^h(w) = K((w - \hat{W}_i)/h)$ and

$$\hat{S}_0^w = \sum_{i=1}^n \hat{K}_i^h(w), \hat{S}_1^w = \sum_{i=1}^n \hat{K}_i^h(w)(w - \hat{W}_i), \hat{S}_2^w = \sum_{i=1}^n \hat{K}_i^h(w)(w - \hat{W}_i)(w - \hat{W}_i)'.$$

Then

$$\hat{a}_Y^h(w) = (\hat{S}_0^w - \hat{S}_1^{w\prime}(\hat{S}_2^w)^{-1}\hat{S}_1^w)^{-1}[\sum_{i=1}^n \hat{K}_i^h(w)Y_i - \hat{S}_1^{w\prime}(\hat{S}_0^w)^{-1}\sum_{i=1}^n \hat{K}_i^h(w)(w - \hat{W}_i)Y_i],$$

For the first stage of the locally linear estimator we also smooth the indicator function in $F_{X_1|Z}(x|z) = E[1(X_{1i} \leq x)|Z_i = z]$. Let b_1 be a positive scalar bandwidth and $\Phi(x)$ be a CDF for a scalar x, so that $\Phi(x/b_1)$ is a smooth approximation to the indicator function. The estimator is a locally linear estimator where $w = z$ and $Y = \Phi((x - X_1)/b_1)$. For observations $(X_{1i}, Z_i), i = 1, ..., n$ on X_1 and Z and a positive bandwidth h_1 an estimator of $F_{X_1|Z}(x|z)$ is

$$\hat{F}_{X_1|Z}(x, z) = \hat{a}_{\Phi((x - X_1)/b_1)}^{h_1}(z).$$

[15]

Then $\hat{V}_i = \hat{F}_{X_1|Z}(X_i, Z_i)$, $(i = 1, ..., n)$. For the second step let $w = (x, v)$, $\hat{W}_i = (X_i, \hat{V}_i)$, b_2, and h_2 be bandwidths. We also use $\Phi(x/b_2)$ to approximate the indicator function for the conditional CDF estimator. The estimators will be locally linear estimators where $Y = \Phi((y - Y)/b_2)$ or just $Y = Y$. These are given by

$$\hat{F}_{Y|X,V}(y|x, v) = \hat{a}^{h_2}_{\Phi((y-Y)/b_2)}(x, v), \hat{m}(x, v) = \hat{a}^{h_2}_Y(x, v).$$

Evidently these estimators depend on the bandwidths b_1, h_1, b_2, and h_2. Derivation of optimal bandwidths is beyond the scope of this paper but we will consider sensitivity to their choice in the application.

To describe a series estimator of $E[Y|W = w]$ for any random vector W, let $p^K(w) = (p_{1K}(w), ..., p_{KK}(w))'$ denote an $K \times 1$ vector of approximating functions, such as power series or splines, and $p_i = p(\hat{W}_i)$. Let $\tilde{a}^K_Y(w)$ denote the series estimator obtained as the predicted value from regressing Y_i on p_i, that is

$$\tilde{a}^K_Y(w) = p^K(w)' \left(\sum_{i=1}^n p_i p_i' \right)^- \sum_{i=1}^n p_i Y_i,$$

where A^- denotes any generalized inverse of the matrix A. Let $\tau(u)$ denote the CDF for a uniform distribution. Then a series estimator of the observations on the control variable is given choosing $w = z$ and calculating

$$\hat{F}_{X_1|Z}(x_1, z) = \tau(\tilde{a}^{K_1}_{1(X_1 \leq x)}(z)).$$

Then $\hat{V}_i = \hat{F}_{X_1|Z}(X_i, Z_i)$, $(i = 1, ..., n)$. For the second stage let $w = (X, V)$, $\hat{W}_i = (X_i, \hat{V}_i)$, b_2 be a bandwidth and K_2 be a number of terms to be used in approximating functions of $w = (X, V)$. Then series estimators of the conditional CDF $F_{Y|X,V}(y|x, v)$ and the conditional expectation $E[Y|X, V]$ are given by

$$\hat{F}_{Y|X,V}(y|x, v) = \tilde{a}^{K_2}_{\Phi((y-Y)/b_2)}(x, v), \hat{m}(x, v) = \tilde{a}^{K_2}_Y(x, v).$$

Evidently these estimators depend on the bandwidth b_2 and number of approximating functions K_1 and K_2. Derivation of optimal values for these tuning parameters is beyond the scope of this paper.

5   An Application

In this Section we consider an application to estimation of a triangular simultaneous equations model for Engel curves. Here Y will be the share of expenditure on a commodity and X will be the log of total expenditure. We use as an instrument Z gross earnings of the head

of household. This instrument is also used by Blundell, Chen, and Kristensen (2007), who motivate it by separability of household saving and consumption decisions. In the application we estimate the QSF and ASF when Y is the share of expenditure on either food or leisure. Here we may interpret the QSF as giving quantiles, across heterogenous individuals, of individual Engel curves. This interpretation depends on ε solely representing heterogeneity and no other source of randomness, such as measurement error.

The data (and this description) is similar to that considered in Blundell, Chen, and Kristensen (2007). The data is taken from the British Family Expenditure Survey for the year 1995. To keep some demographic homogeneity the data is a subset of married and cohabitating couples where the head of the household is aged between 20 and 55 and those with three or more children are excluded. Unlike Blundell et. al. (2007), we do not include number of children as covariates. In this application we exclude households where the head of household is unemployed in order to have the instrument Z available. This earnings variable is the amount that the male of the household earned in the chosen year before taxes. This leaves us with 1655 observations.

In this application we use locally linear estimators as described earlier. We use Silverman's (1986) density bandwidth throughout and carry out some sensitivity checks. We also check sensitivity of the results to the choice δ_n used in the bounds.

As previously discussed, an important identification concern is over what values of X the common support condition might be satisfied. Similarly to the rank condition in linear models, the common support condition can be checked by examining the data. We do so in Figure 1, that gives a graph of level sets of a joint kernel density estimator for (X, V) based on X_i and the control variable estimates $\hat{V}_i = \hat{F}_{X|Z}(X_i|Z_i)$. This figure suggests that Assumption 2 may be satisfied only over a narrow range of X values, so that it may be important to allow for bounds.

For comparison purposes we first give graphs of the QSF and ASF for food and leisure expenditure respectively assuming that the common support condition is satisfied.[7] Figure 2 and 3 report graphs of these functions for the quartiles. These graphs have the shape one has come to expect of Engel curves for these commodities. In comparing the curves it is interesting to note that there is evidence of substantial asymmetry for the leisure expenditure. The QSF for $\tau = 1/2$ (i.e. the median) is quite different from the ASF and there is more of a shift towards leisure at the upper quantiles of the expenditure. There is less evidence of asymmetry for food expenditure.

Turning now to the bounds, we chose δ_n so the probability that a Gaussian pdf (with mean

---

[7]These graphs were initially derived by Richard Blundell.

equal to the sample mean $\hat{\mu}$ of X and variance equal to the sample variance $\hat{\sigma}^2$) exceeds δ_n is .975. This δ_n satisfies the equation

$$\int_{\phi((t-\hat{\mu})/\hat{\sigma})\geq\delta_n} \phi((t-\hat{\mu})/\hat{\sigma})dt = .975.$$

Figure 4 graphs the $\hat{P}(x)$ for this δ_n. The bounds coincide when $\hat{P}(x) = 0$ but differ when it is nonzero. Here we find that the bounds will coincide only over a small interval of x values.

Figures 5 and 6 graph bounds for the median QSF for food and leisure, along with an estimator of the marginal pdf of total expenditure X. Here we find that even though the upper and lower bounds coincide only over a small range, they are quite informative.

We also carried out some sensitivity analysis. We found that the ASF and QSF estimates are not very sensitive to the choice of bandwidth. Also, increasing δ_n does widen the bounds appreciably, although δ_n does not have to increase much before $\hat{P}(x)$ is nonzero for all x.

6 Asymptotic Theory

We have presented two kinds of estimators for a variety of functionals. A full account of asymptotic theory for all these cases is beyond the scope of this paper. As an example here we give asymptotic theory for a power series estimator of the ASF in the triangular model. Here we assume that the order of the approximating functions, i.e. the sum of the exponents of the powers in $p_{kK}(w)$, are increasing in K, with all terms of a given order included before increasing the order.

Results for the power series estimators are used to highlight two important features of the estimation problem that arise from the fact that the joint density of x and V goes to zero on the boundary of the control variable. One feature is that the rate of convergence of the ASF will depend on the how fast the density goes to zero, since the ASF integrates over the control variable. The other feature is that the ASF does not necessarily converge at the same rate as a regression of Y on just X. In other words, unlike e.g. in Newey (1994), integrating over a conditioning variable does not lead to a rate that is the same as if that variable was not present.

The convergence rates of the estimators will depend on certain smoothness restrictions. The next Assumption imposes smoothness conditions on the control variable.

ASSUMPTION 6.1: *$Z_i \in \Re^{r_1}$ has compact support and $F_{X_1|Z}(x_1, z)$ is continuously differentiable of order d_1 on the support with derivatives uniformly bounded in x and z.*

This condition implies an approximation rate of $K_1^{-d_1/r_1}$ for the CDF that is uniform in both its arguments; see Lorentz (1986). The following result gives a convergence rate for the first step:

LEMMA 11: *If the conditions of Theorem 2.1 and Assumption 6.1 are satisfied then*

$$E\left[\sum_{i=1}^{n}(\hat{V}_i - V_i)^2/n\right] = O(K_1/n + K_1^{1-2d_1/r_1}).$$

The two terms in this rate result are variance (K_1/n) and squared bias $(K_1^{1-2d_1/r_1})$ terms respectively. In comparison with previous results for series estimators, this convergence result has $K_1^{1-2d_1/r}$ for the squared bias term as a rate rather than $K_1^{-2d_1/r}$. The extra K_1 arises from the predicted values $\hat{V}_i$ being based on regressions with the dependent variables varying over the observations.

To obtain convergence rates for series estimators it is necessary to restrict the rate at which the density goes to zero as V approaches zero or one. The next condition fulfills this purpose. Let $w = (x, v)$ and $\mathcal{X}$ denote the support of $\mathcal{X}$.

ASSUMPTION 6.2: *$\mathcal{X}$ is a Cartesian product of compact intervals, $p^{K_2}(w) = p^{K_x}(x) \otimes p^{K_V}(v)$, and there exist constants $C, \alpha > 0$ such that*

$$\inf_{x \in \mathcal{X}} f_{X,V}(x, v) \geq C[v(1-v)]^\alpha.$$

The next condition imposes smoothness of $m(w)$, in order to obtain an approximation rate for the second step.

ASSUMPTION 6.3: *$m(w)$ is continuously differentiable of order d_2 on $\mathcal{X} \times [0,1] \subset \Re^{r_2}$.*

Note that w is an $r_2 \times 1$ vector, so that x is a $(r_2 - 1) \times 1$ vector. Next, we bound the conditional variance of Y, as is often done for series estimators.

ASSUMPTION 6.4: *$Var(Y|X_1, Z)$ is bounded.*

With these conditions in place we can obtain a convergence rate bound for the second-step estimator.

THEOREM 12: *If the conditions of Theorem 1 and Assumptions 6.1 - 6.4 are satisfied and $K_2^2 K_V^{\alpha+2}(K_1/n + K_1^{1-2d_1/r}) \to 0$ then*

$$\int [\hat{m}(w) - m(w)]^2 \, dF(w) = O_p(K_2/n + K_2^{-2d_2/r_2} + K_1/n + K_1^{1-2d_1/r_1})$$

$$\sup_{w \in W} |\hat{m}(w) - m(w)| = O_p(K_V^\alpha K_2[K_2/n + K_2^{-2d_2/r_2} + K_1/n + K_1^{1-2d_1/r_1}]^{1/2}).$$

This result gives both mean-square and uniform convergence rates for $\hat{m}(x, V)$. It is interesting to note that the mean-square rate is the sum of the first step convergence rate and the rate that would obtain for the second step if the first step was known. This result is similar to that of Newey, Powell, and Vella (1999), and results from conditioning on the first step in the second step regression. Also, the first step and second step rates are each the sum of a variance term and a squared bias term.

The following result gives an upper bound on the rate of convergence for the ASF estimator $\hat{\mu}(x) = \int_0^1 \hat{m}(x, v) dv$.

THEOREM 13: *If the conditions of Theorem 2.1 and Assumptions 6.1 - 6.4 are satisfied and* $K_2^2 K_V^{2+2\alpha}(K_1/n + K_1^{1-2d_1/r_1}) \longrightarrow 0$ *then*

$$\int [\hat{\mu}(x) - \mu(x)]^2 F_X(dx) = O_p(K_V^{2+2\alpha}(K_x/n + K_2^{-2d_2/r_2} + K_1/n + K_1^{1-2d_1/r_1})).$$

To interpret this result, we can use the fact that all terms of a given order are added before increasing the order to say that there is a constant C with $K_2 \geq K_V^{r_2}/C$ and $K_x \leq CK_V^{r_2-1}$. In that case we will have

$$\int [\hat{\mu}(x) - \mu(x)]^2 F_X(dx) = O_p(K_V^{r_2+1+2\alpha}/n + K_V^{2+2\alpha}(K_V^{-2d_2} + K_1/n + K_1^{1-2d_1/r_1})).$$

The choice of K_V and K_1 minimizing this expression is proportional to $n^{1/(2d_2+r_2-1)}$ and $n^{r_1/2d_1}$ respectively. For this choice of K_V and K_1 the rate hypothesis and the convergence rate are given by

$$\frac{2(r_2 + \alpha + 1)}{2d_2 + r_2 + 1} + \frac{r_1}{2d_1} < 1,$$
$$\int [\hat{\mu}(x) - \mu(x)]^2 F_X(dx) = O_p(n^{2(2+\alpha-d_2)/(2d_2+r_2-1)} + n^{[2(1+\alpha)/(2d_2+r_2-1)]+(r_1/2d_1)-1}).$$

The inequality requires that $m(w)$ have more than $(1 + 2\alpha + r_2)/2$ derivatives and that $F_{X_1|Z}(x_1|z)$ have more than $r_1/2$ derivatives.

One can compare convergence rates of estimators in a model where several estimators are consistent. One such model is the additive disturbance model

$$Y = g(X) + \varepsilon, \ X = Z + \eta, Z \sim N(0, 1), \eta \sim N(0, (1 - R^2)/R^2),$$

where X, Z, ε, and η are scalars and we normalize $E[\varepsilon] = 0$. Here additive triangular and nonparametric instrumental variable estimators will be consistent, in addition to the triangular nonseparable estimators given here. Suppose that the object of estimation is the ASF $g(x)$. Under regularity conditions like those give above the estimator will converge at a rate that is

a power of n, but slower than the optimal one-dimensional rate. In contrast, the estimator of Newey, Powell, and Vella (1999), which imposes additivity, does converge at the optimal one-dimensional rate. Also, estimators of $g(x)$ that only use $E[\varepsilon|Z] = 0$, will converge at a rate that is slower than any power of n (e.g. see Severini and Tripathi, 2003). Thus, the convergence rate we have obtained here is intermediate between that of an estimator that imposes additivity and one that is based just on the conditional mean restriction.

7   Conclusion

The identification and bounds results for the QSF, ASF, and policy effects also apply to settings with an observable control variable V, in addition to the triangular model. For example, the set up with $Y = g(X, \varepsilon)$ for $X \in \{0, 1\}$ and unrestricted ε, and Assumption 1 for an observable V is a well known treatment effects model, where Assumption 1 is referred to as unconfoundedness or selection on observables (e.g., Imbens, 2004; Heckman, Ichimura, Smith and Todd, 1998). The QSF identification and bounds apply to this model, and to generalizations where X takes on more than two values.

8   Appendix

PROOF OF THEOREM 1: Let $h^{-1}(z, x)$ denote the inverse function for $h(z, \eta)$ in its second argument, which exists by condition ii). Then, as shown in the proof of Lemma 1 of Matzkin (2003),

$$
\begin{aligned}
F_{X_1|Z}(x, z) &= Pr(X_1 \le x|Z = z) = Pr(h(z, \eta) \le x|Z = z) = Pr(\eta \le h^{-1}(z, x)|Z = z) \\
&= Pr(\eta \le h^{-1}(z, x)) = F_\eta(h^{-1}(z, x)),
\end{aligned}
$$

By condition ii), $\eta = h^{-1}(X_1, Z)$, so that plugging in gives

$$
V = F_{X_1|Z}(X_1, Z) = F_\eta(h^{-1}(X_1, Z)) = F_\eta(\eta).
$$

By F_η strictly monotonic on the support of η, the sigma algebra associated with η is equal to the one associated with $V = F_\eta(\eta)$, so that conditional expectations given η are identical to those given V. Also, for any bounded function $a(X)$, by independence of Z and (ε, η),

$$
E[a(X)|\eta, \varepsilon] = \int a(h(z, \eta))F_Z(dz) = E[a(X)|\eta]
$$

Therefore, for any bounded function $b(\varepsilon)$ we have

$$
E[a(X)b(\varepsilon)|V] = E[b(\varepsilon)E[a(X)|\eta, \varepsilon]|\eta] = E[b(\varepsilon)E[a(X)|\eta]|\eta] = E[b(\varepsilon)|\eta]E[a(X)|\eta].
$$

[21]

Q.E.D.

PROOF OF THEOREM 2: Define $V = F_{X|Z}(X, Z)$ and let $(X, k(V, X))$ denote the inverse of $(X, F_{X|Z}(X|Z))$, so that $Z = k(V, X)$. It then follows by (X, V) and (X, Z) being one-to-one transformations of each other that

$$Q_{Y|X,V}(\tau, X, V) = Q_{Y|X,Z}(\tau, X, Z) = Q_{Y|X,Z}(\tau, X, k(V, X)).$$

Also, by the inverse function theorem, $Q_{X|Z}(\tau, z)$ is differentiable at (v_0, z_0) and $k(v, x)$ is differentiable at (v_0, x_0) with

$$\nabla_x k(v_0, x_0) = -\nabla_x F_{X|Z}(x_0, z_0)/\nabla_z F_{X|Z}(x_0, z_0) = 1/\nabla_z Q_{X|Z}(v_0, z_0).$$

Then by the chain rule

$$
\begin{aligned}
\nabla_x Q_{Y|X,V}(\tau_0, x_0, v_0) &= \frac{\partial}{\partial x} Q_{Y|X,Z}(\tau_0, x, k(v_0, x))|_{x=x_0} \\
&= \nabla_x Q_{Y|X,Z}(\tau_0, x_0, z_0) + \nabla_z Q_{Y|X,Z}(\tau_0, x_0, z_0)\nabla_x k(v_0, x_0) \\
&= \nabla_x Q_{Y|X,Z}(\tau_0, x_0, z_0) + \nabla_z Q_{Y|X,Z}(\tau_0, x_0, z_0)/\nabla_z Q_{X|Z}(v_0, z_0). Q.E.D.
\end{aligned}
$$

PROOF OF THEOREM 3: By Assumption 2 the support of V conditional on $X = x$ equals the support $\mathcal{V}$ of V, so that $\Pr(Y \leq y|X = x, V)$ is unique with probability one on $\mathcal{X} \times \mathcal{V}$. The conclusion then follows by the derivation in the text. Q.E.D.

PROOF OF THEOREM 4: By the definition of $\mathcal{V}(x)$ and Assumption 1, on a set of x with probability 1, integrating eq. (3.4) gives

$$\int_{\mathcal{V}(x)} m(x, v)F_V(dv) = \int_{\mathcal{V}(x)} \int g(x, e)F_{\varepsilon|V}(de|V)F_V(dV).$$

Also by $B_\ell \leq g(x, e) \leq B_u$ it follows that $B_\ell \leq \int g(x, e)F_{\varepsilon|V}(de|V) \leq B_u$, so that

$$B_\ell P(x) \leq \int_{\mathcal{V} \cap \mathcal{V}(x)^c} \int g(x, e)F_{\varepsilon|V}(de|V)F_V(dV) \leq B_u P(x).$$

Summing up these two equations and applying iterated expectations gives

$$\mu_\ell(x) \leq \int \int_{\mathcal{V}} g(x, e)F_{\varepsilon|V}(de|V)F_V(dV) = \mu(x) \leq \mu_u(x).$$

To see that the bound is sharp, let $\varepsilon = V$ and

$$g^u(x, \varepsilon) = \begin{cases} m(x, V), V \in \mathcal{V}(x), \\ B_u, V \notin \mathcal{V}(x). \end{cases}$$

[22]

By ε constant given V, ε is independent of X conditional on V. Then $\mu(x) = \mu_u(x)$. Defining $g^\ell(x, \varepsilon)$ similarly with B_ℓ replacing B_u, gives $\mu(x) = \mu_\ell(x)$. Q.E.D.

PROOF OF THEOREM 5: Note first that by Assumption 1,

$$G_\ell(y, x) = \int_{\mathcal{V}(x)} \Pr(Y \leq y | X = x, V = v) F_V(dv) = \int_{\mathcal{V}(x)} \Pr(g(x, \varepsilon) \leq y | X = x, V = v) F_V(dv)$$

$$= \int_{\mathcal{V}(x)} \Pr(g(x, \varepsilon) \leq y | V = v) F_V(dv).$$

Then by $\Pr(g(x, \varepsilon) \leq y | V) \geq 0$ we have

$$G_\ell(y, x) \leq \int \Pr(g(x, \varepsilon) \leq y | V = v) F_V(dv) = G(y, x).$$

Also by $\Pr(g(x, \varepsilon) \leq y | V) \leq 1$ we have

$$G(y, x) = G_\ell(y, x) + \int_{\mathcal{V} \cap \mathcal{V}(x)^c} \Pr(g(x, \varepsilon) \leq y | V = v) F_V(dv) \leq G_\ell(y, x) + \int_{\mathcal{V} \cap \mathcal{V}(x)^c} F_V(dv) = G_u(y, x).$$

The conclusion then follows by inverting. Q.E.D.

PROOF OF THEOREM 6: By the fact that $g(x, \varepsilon)$ continuously differentiable and the integrability condition, it follows that $m(x, v)$ is differentiable and eq. (3.11) is satisfied. Then by eq. (3.12) the average derivative is an explicit functional of the data distribution, and so is identified. For the policy effect by the assumption about it follows that $\beta(x, t)$ is well defined, with probability one, at $(x, t) = (\ell(X), V)$, so that the conclusion follows as in equation (3.10) Q.E.D.

PROOF OF THEOREM 7: By eq. (3.4) $m(X, V) = E[Y|X, V] = \int g(X, e) F_{\varepsilon|V}(de|V)$. Then by $T(\int g(\cdot, e) F_{\varepsilon|V}(de|V), X) = \int T(g(\cdot, e), X) F_{\varepsilon|V}(de|V)$ and iterated expectations,

$$E[T(m(\cdot, V), X)] = E[T(\int g(\cdot, e) F_{\varepsilon|V}(de|V), X)] = E[\int T(g(\cdot, e), X) F_{\varepsilon|V}(de|V)]$$

$$= E[\mathbb{E}[T(g(\cdot, \varepsilon), X)|V, X]] = E[T(g(\cdot, \varepsilon), X)].$$

Since $E[T(g(\cdot, \varepsilon), X)]$ is equal to an explicit function of the data distribution, it is identified. Q.E.D.

PROOF OF THEOREM 8: See text.

PROOF OF THEOREM 9: See text.

[23]

The proof of Lemmas 10 and 11 and Theorems 12 and 13 are given in the supplementary material for this paper.

REFERENCES

Altonji, J., and R. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Econometrica* 73, 1053-1102.

Angrist, J., G.W. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association 91, 444-472.*

Angrist, J., K. Graddy, and G.W. Imbens (2000): "The Interpretation of Instrumental Variable Estimators in Simultaneous Equations Models with An Application to the Demand for Fish," *Review of Economic Studies* 67, 499-527.

Athey, S. (2002), "Monotone Comparative Statics Under Uncertainty" *Quarterly Journal of Economics*, 187-223.

Athey, S., and P. Haile (2002), "Identification of Standard Auction Models", *Econometrica* 70, 2107-2140.

Athey, S., and S. Stern, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design", NBER working paper 6600.

Blundell, R., and J.L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in M. Dewatripont, L. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics,* Ch. 8, 312-357.

Blundell, R., and J.L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies* 71, 581-913.

Blundell, R., A. Gosling, H. Ichimura, C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Unemployment Using Bounds," *Econometrica* 75, 323-363.

Blundell, R., X. Chen, and D. Christensen (2007), "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica* 75, 1613-1669.

Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69, 1127-1160.

Chamberlain, G., (1984): "Panel Data," in Griliches and Intrilligator (eds.), *Handbook of Econometrics*, Vol 2.

Chernozhukov, V. and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica* 73, 1127-1160.

Chernozhukov, V., G. Imbens, and W. Newey (2007), "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics* 139, 4-14.

Chesher, A. (2002), "Semiparametric Identification in Duration Models," Cemmap working paper CWP20/02.

Chesher, A. (2003), "Identification in Nonseparable Models," *Econometrica* 71, 1405-1441.

Chesher, A. (2005), "Nonparametric Identification Under Discrete Variation," *Econometrica* 73, 1525-1550.

Chesher, A. (2006), "Identification of Nonadditive Structural Functions," Invited Lecture at the 2005 World Congress of the Econometric Society, *Advances in Economic and Econometrics, Theory and Applications,* Blundell, Newey, and Persson eds., Cambridge: Cambridge University Press.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2008): "Dealing with Limited Overlap in Estimation of Average Treatment Effects,"forthcoming *Biometrika.*

Darolles, S., J.-P., Florens, and E. Renault, (2001), "Nonparametric Instrumental Regression," working paper.

Das, M. (2004): "Instrumental Variable Estimators for Nonparametric Models with Discrete Endogenous Regressors," *Journal of Econometrics* 124, 335-361.

Das, M. (2001): "Monotone Comparative Statics and the Estimation of Behavioral Parameters," Working Paper, Department of Economics, Columbia University.

de Jong, R. (2002): "A Note on Convergence Rates and Asymptotic Normality for Series Estimators': Uniform Convergence Rates," *Journal of Econometrics* 111, 1-9.

Doss, H. and R.D. Gill (1992): "An Elementary Approach to Weak Convergence for Quantile Processes, With Applications to Censored Survival Data," *Journal of the American Statistical Association* 87, 869-877.

Florens, J.P., J.J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica,* forthcoming.

Hall P., and J.L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics,* 33,2904-2929.

Heckman, J.J., H. Ichimura, J. Smith, P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.

Heckman, J.J., and E.J. Vytlacil (2000): "Local Instrumental Variables," in Hsiao, C., K.Morimune, and J.L. Powell, eds., *Nonlinear Statististical Modeling*; Cambridge: Cambridge University Press.

Heckman, J.J., and E.J. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation", *Econometrica* 73, 669-738.

Heckman, J.J., and E.J. Vytlacil (2008): "Generalized Roy Model and Cost Benefit Analysis of Social Programs," working paper, Yale University.

Hengartner, N.W. and O.B. Linton (1996): "Nonparametric Regression Estimation at Design Poles and Zeros," *Canadian Journal of Statistics* 24, 583-591.

Hurwicz, L. (1950): "Generalization of the Concept of Identification," in *Statistical Inference in Dynamic Ecnomic Models*, Cowles Comission Monograph 10, New York: John Wiuley.

Imbens, G. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G.W. (2006): "Nonadditive Models with Endogenous Regressors," Invited Lecture at the 2005 World Congress of the Econometric Society, *Advances in Economic and Econometrics, Theory and Applications,* Blundell, Newey, and Persson eds., Cambridge: Cambridge University Press.

Imbens, G.W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62, 467-476.

Imbens, G.W. and J. Wooldridge (2009): "Recent Developments in the Econometrics of Program

Evaluation," *Journal of Economic Literature*, forthcoming.

Lehmann, E. L. (1974): Nonparametrics: Statistical Methods Based on Ranks. San Francisco, CA: Holden-Day.

Ma, L. and R. Koenker (2006): "Quantile Regression Methods for Recursive Structural Equation Models," *Journal of Econometrics* 134, 471-506.

Manski, C. (1994): "The Selection Problem," in C. Sims (ed.), *Advances in Economics and Econometrics*, Cambridge University Press, Cambridge, England.

Manski, C. (2007), *Identification for Prediction and Decision*, Princeton, Princeton University Press.

Matzkin, R. (1993), "Restrictions of Economic Theory in Nonparametric Models" *Handbook of Econometrics*, Vol IV, Engle, R. and D. McFadden (eds.), Amsterdam: North-Holland.

Matzkin, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions", *Econometrica* 71, 1339-1375.

Matzkin, R. (2006): "Heterogenous Choice," Invited Lecture at the 2005 World Congress of the Econometric Society, *Advances in Economic and Econometrics, Theory and Applications,* Blundell, Newey, and Persson eds., Cambridge: Cambridge University Press.

Milgrom, P., and C. Shannon, (1994), "Monotone Comparative Statics," *Econometrica*, 58, 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.K. (1994), "Kernel Estimation of Partial Means and a Variance Estimator", *Econometric Theory* 10, 233-253.

Newey, W.K. and J.L. Powell (2003): "Nonparametric Instrumental Variables Estimation," *Econometrica* 71, 1565-1578.

Newey, W.K., J.L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica* 67, 565-603.

Pinkse, J., (2000a): "Nonparametric Two-step Regression Functions when Regressors and Error are Dependent," *Canadian Journal of Statistics* 28, 289-300.

Pinkse, J. (2000b): "Nonparametric Regression Estimation Using Weak Separability", University of British Columbia.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

Roehrig, C. (1988): "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica* 55, 875-891.

Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

Stock, J. (1988): "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," in W. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics*, Cambridge University Press, Ch. 3, 77-98.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica* 54, 1461-1481.

Vytlacil, E.J. and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica* 75, 757 - 779.

Wooldridge, J. (2002): *Econometric Analysis of Cross Section and Panel Data* MIT Press.