

Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity*

Guido W. Imbens
Department of Economics,
Harvard University, and NBER

Whitney K. Newey
Department of Economics
M.I.T.

First Draft: March 2001
This Draft: October 2006

Abstract

This paper investigates identification and estimation of models with nonseparable, multi-dimensional disturbances using control variables. Triangular simultaneous equations models are considered, with instruments and disturbances independent and reduced form that is strictly monotonic in a scalar disturbance. It is shown that in that setting 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, policy, and derivative effects. Bounds for average and quantile effects 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.

JEL Classification: C21, C23, C31, C33

Keywords: *Nonseparable Models, Control Variables, Quantile Effects, Bounds, Average Derivative, Policy Effect, Nonparametric Estimation, Demand Analysis.*

*This research was partially completed while the second author was a fellow at the Center for Advanced Study in the Behavioral Sciences during the year 2000/2001. The NSF provided partial financial support through grants SES 0136789 (Imbens) and SES 0136869 (Newey). This version (absent the bounds) was presented at the December 2003 meeting of EC2. We are grateful for comments by S. Athey, L. Benkard, G. Chamberlain, A. Chesher, J. Heckman, O. Linton, A. Nevo, A. Pakes, J. Powell, R. Blundell and participants at seminars at Stanford University, University College London, Harvard University, MIT, and Northwestern University. We especially thank R. Blundell for providing the data and initial empirical results.

1 Introduction

Models with endogeneity are central in econometrics. A key feature of many of these models, motivated by economics, is nonseparability in disturbances. In this paper we provide identification and estimation results for such models via control variables, that is 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 reduced form strictly monotonic in a scalar disturbance, a control variable is given by the conditional distribution function of the endogenous variables given the instruments. We also give identification and estimation results for several structural effects when *any* control variable is present.

We focus on models where the structural equation disturbance can be a vector of any dimension. We relate these results to other, more conventional, cases where the disturbance is a scalar or a pair of random variables. By focusing on the general disturbance case we allow for individual heterogeneity and other random effects in a completely flexible way. Also, since a nonseparable model with a general disturbance is isomorphic to treatment effects models (with possibly continuous treatment), our identification results for quantile, policy, and derivative effects apply to treatment effect models.

The quantile effect we consider is the quantile of the structural outcome for a fixed value of the endogenous variables. Such quantiles correspond to the value of the structural function at quantiles of the disturbance when the structural function is monotonic in a scalar disturbance. More generally, they can be used to characterize how endogenous variables affect the distribution of structural 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 effects when the common support condition is not satisfied.

We also consider identification and estimation of average features of the structural function. Such averages have long been of interest, because they summarize structural effects for a whole population. Early examples are the average response probability in Chamberlain (1984) and the average derivative in Stoker (1986). We give identification and estimation results for certain policy functions and the average derivative.¹ We also consider several effects that have been previously analyzed, where we provide a control function in the triangular model. These are the average structural function of Blundell and Powell (2003) and Wooldridge (2002), which is the average effect in a treatment effects model, and the local average response of Altonji and Matzkin (2005), which was characterized by Florens et. al. (2004) as the average effect for the

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

treated when there is continuous treatment.

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 of interest given the endogenous variable and the control function. Various structural effects are then recovered by averaging over the control function or the endogenous variable and control function together.

An important feature of the triangular model is that the joint density of the endogenous variable and the control function goes to zero at the boundary of the support of the control function. Consequently, using nonparametric estimators with low sensitivity to edge effects may be important. We consider both locally linear and series estimators, because kernel estimators are known to converge at slower rates in this setting.

The edge effect also impacts the convergence rate of the estimators. Averaging over the control function "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 the average derivative, and other 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) but faster than the rates when only a conditional mean condition is satisfied.

Our control variable results for the triangular model extend the work of Blundell and Powell (2003) to a nonseparable reduced form, who had extended Newey, Powell, and Vella (1999) and Pinske (2000b) to a nonseparable structural equation. Chesher (2003, 2005) considered local identification of levels and derivatives of a structural function in a triangular nonseparable system with one disturbance per equation and 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 function.

The independence of disturbances and instruments that we impose is stronger than the conditional mean restriction of Newey and Powell (1988, 2003), Das (2004), Darrolles, Florens, and Renault (2003), and Hall and Horowitz (2004) and the local independence of Chesher (2003). With one or two structural disturbances, the triangular model is a special case of that of Roehrig (1988), although our identification results are not. Imbens and Angrist (1994) and

Angrist, Graddy, and Imbens (2000) also allow for nonseparable disturbances but focus on effects other than the ones we consider. Das (2001) also allows for nonseparable disturbances, but considers a single index setting with monotonicity. 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.

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 heterogeneity, 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 one 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. Equation (2.2) can be thought of as a reduced form equation for X_1 .

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 x given her information set. The information set consists of a scalar noisy signal η of the unobserved input ε and a cost shifter Z .² The cost

²Although 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.

function is $c(x, z)$. Then X would be obtained as the solution to the individual choice problem

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

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 $\operatorname{argmax}_x \{ g_0(x) + \mathbb{E}[\varepsilon | \eta] - c(x, Z) \}$. In that case the solution 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.

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 using monotone comparative statics results (e.g., Milgrom and Shannon, 1994; Athey, 2002). 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) the cost function and the marginal cost 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 an observable or estimable variable v satisfying 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 mean or CDF of Y given X and v .

In the model of equations (2.1) and (2.2), it turns out that under independence of (ε, η) and Z , a control variable is $v = F_{X_1|Z}(X_1, Z)$, where $F_{X_1|Z}(x_1, z)$ is the conditional CDF of X_1 given

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

Z . The conditional independence of Assumption 1 arises because conditional on η the variable X_1 will only depend on Z and because the scalar nature of η and independence of Z and η lead to the $F_{X_1|Z}(X_1, Z)$ being a one-to-one function of η . 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 function⁴. Otherwise, all of the endogeneity cannot be absorbed into identifiable variables, as discussed in Imbens (2005).

All of our identification results for the outcome function $g(X, \varepsilon)$ will only depend on existence of some control variable v , rather than on the particular structure of the triangular model. To emphasize this, we will state our identification results for g by referring only to Assumption 1 rather than to the reduced form equation (2.2). For example, $Y = g(X, \varepsilon)$ and Assumption 1 with an observable v are isomorphic to the well known treatment effects model, where X is the (possibly continuous) treatment variable, Y is the outcome variable, and Assumption 1 is selection on observables. Thus, the identification results given below apply to treatment effects models. Also, Altonji and Matzkin (2005) suggest some potential control variables for nonseparable panel data and Matzkin (2004) some control variables based on unobservable instruments, and the results below could be applied for identification of quantile and policy effects for those control variables.

Another example is a nonseparable sample selection model, where $Y = g(X, \varepsilon)$ but (Y, X) are only observed if some selection indicator variable $S \in \{0, 1\}$ is equal to 1. In Newey (2006) it is shown that if $S = 1(\Pi(Z) \leq \eta)$ and (ε, η) are independent of (X, Z) , then the selection probability is a control variable conditional on selection. Our identification results, along with previous ones, then apply to give identification of various effects conditional on selection.

The effects for which we give identification results are quantiles, policy, average derivatives. For expositional purposes and to help make our contribution clear we here list these effects, along with some references to previous appearances in the literature.

Quantile structural function (QSF):

$$q_Y(\tau, x) \text{ is the } \tau^{th} \text{ quantile of } g(x, \varepsilon).$$

In this definition x is fixed and ε is what makes $g(x, \varepsilon)$ random. 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), Chernozhukov and Hansen (2005) and Firpo (2007). 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 τ^{th} quantile of ε . The QSF is one generalization of the value of g to the case where ε is a vector, e.g. where there are multiple sources of heterogeneity.

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

Policy effect:

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

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 $\mathbb{E}[g(\ell(X), \varepsilon) - Y]$.

Average derivative:

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

This object is analogous to the average derivative studied in 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 $\delta = \mathbb{E}[\beta(\varepsilon)]$. 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 .

One contribution of this paper is to give control variable identification results for the QSF, policy effects, the average derivative, and a certain linear class of functionals discussed below. Control variable identification results have previously been given for other functionals. For these a contribution is to show that 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. For this reason, and to help place our contribution in the context of the literature, we give here a brief account of other functionals that have previously been shown to be identified under Assumption 1.

Local Average Response (LAR):

$$\beta(x) = \int [\partial g(x, \varepsilon)/\partial x] F_{\varepsilon|X}(d\varepsilon|x).$$

This object was considered by Altonji and Matzkin (2005). It was characterized as the effect of treatment on the treated for continuous treatment by Florens, Heckman, Meghir, and Vytlacil (2004), who considered how restrictions on g help with identification of this and other objects.

Average Structural Function (ASF):

$$\mu(x) = \int g(x, \varepsilon) F_{\varepsilon}(d\varepsilon).$$

This object was considered by Chamberlain (1984), Blundell and Powell (2003), and Wooldridge (2002) as especially useful for binary choice models. In the treatment effects model it is the average treatment effect.

In addition to these, Chesher (2003, 2005) has considered identification of interesting local effects in the triangular simultaneous equations model of equations (2.1) and (2.2). This work focuses mainly on models where ε is known to be one or two dimensional, though see Chesher (2002). In particular, Chesher (2003) considers the triangular system when $\varepsilon = (\eta, \xi)$, giving sufficient local conditions for identification of $\partial g(x, \varepsilon)/\partial x = \partial g(x, \eta, \xi)/\partial x$. An advantage of these objects is that their identification only requires local independence of the instruments and disturbances. A disadvantage is that they require correct specification of the number of elements of ε .

The control variable $v = F_{X_1|Z}(X_1, Z)$ is related to Chesher's (2003) results. 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 . Then we show in the Appendix, where all our results are proved, that under certainly regularity conditions, that do not include equations (2.1) or (2.2),

$$\nabla_x Q_{Y|X,v}(\tau, x, v) = \nabla_x Q_{Y|X,Z}(\tau, x, z) + \frac{\nabla_z Q_{Y|X,Z}(\tau, x, z)}{\nabla_z Q_{X|Z}(v, z)}. \quad (2.3)$$

Chesher (2003) uses the object on the right of the equality to identify $\partial g(x, \eta, \xi)/\partial x$ under certain local independence conditions. Equation (2.3) shows that conditioning on the control variable $v = F_{X|Z}(X|Z)$ leads to the same local, derivative effect, 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 $v = F_{X_1|Z}(X_1, Z)$, but without explicit specification of this variable. Explicit conditioning on v is more useful for our results, which involve averaging over v , as discussed below.⁵

Altonji and Matzkin (2005) also consider identification of the function $g(x, \varepsilon)$ and the distribution of ε when g is strictly monotonic in a scalar ε , up to a specific transformation of ε . For some fixed \bar{x} they give conditions for identification of $g(x, g^{-1}(\bar{x}, u))$ where $u = g(\bar{x}, \varepsilon)$, when both \bar{x} and x are included in the support of X conditional on a value of v . Identification of $g(x, g^{-1}(\bar{x}, u))$ suffices for identification of the QSF, since the τ^{th} quantile of $g(x, g^{-1}(\bar{x}, u))$, which is identified under the conditions of Altonji and Matzkin (2005), is

$$g(x, g^{-1}(\bar{x}, q_u(\tau))) = g(x, q_\varepsilon(\tau)) = q(x, \tau).$$

We innovate by giving identification results for the QSF when the dimension of ε is unknown..

⁵One 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 .

3 Identification

In this section we give precise identification results. The first result gives conditions for $F_{X_1|Z}(X_1, Z)$ to be a control function in the triangular model.

THEOREM 1: *In the model of equations (2.1) and (2.2), if i) (Independence) (ε, η) and Z are independent; ii) (Monotonicity) η is continuously distributed 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 $v = F_{X_1|Z}(X_1, Z)$.*

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.

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 (2000) gives a multivariate version. Das (2001) uses a stochastic version of this assumption to identify parameters in single index models with a single endogenous regressor.

We now turn to identification of functionals of $g(X, \varepsilon)$ based on a control variable. Identification of structural effects requires that X varies while holding the control variable v constant. For identification of the QSF 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 .*

For example, 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 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.

To show identification we give explicit formulae in terms of identified objects. These formulae will later be used to construct estimators. For the QSF note first that by Assumption 1,

$$F_{Y|X,v}(y|x, v) = \int \mathbf{1}(g(x, e) \leq y) F_{\varepsilon|X,v}(de|x, v) = \int \mathbf{1}(g(x, e) \leq y) F_{\varepsilon|v}(de|v). \quad (3.4)$$

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

$$\int F_{Y|X,v}(y|x, v) F_v(dv) = \int \mathbf{1}(g(x, e) \leq y) F_{\varepsilon}(de) = \Pr(g(x, \varepsilon) \leq y) \stackrel{def}{=} G(y, x). \quad (3.5)$$

Then by the definition of the QSF we have

$$q_Y(\tau, x) = G^{-1}(\tau, x). \quad (3.6)$$

Thus the QSF is the inverse of the integral over the marginal distribution of v of the conditional CDF of Y given X and v . 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 .

The following theorem gives a precise statement of this identification result.

THEOREM 2: (*Identification of the QSF*) *If Assumptions 1 and 2 are satisfied then the QSF is identified on the set \mathcal{X} .*

As mentioned in Section 2, identification of the QSF also implies identification of the structural function in the one disturbance case. However, Altonji and Matzkin (2005) give weaker conditions for identification with one disturbance, so we do not pursue that case here.

For identification of the policy effect γ we can replace Assumption 2 by the assumption that the support of (X, v) includes the support of $(\ell(X), v)$. Then

$$\mathbb{E}[g(\ell(X), \varepsilon)] = \mathbb{E}[\mathbb{E}[g(\ell(X), \varepsilon)|X, v]] = \mathbb{E} \left[\int g(\ell(X), \varepsilon) F_{\varepsilon|v}(d\varepsilon|v) \right] = \mathbb{E}[m(\ell(X), v)], \quad (3.7)$$

where the second equality follows by Assumption 1.

For identification of the average derivative, all that we require is that X is continuously distributed given v , so that the derivative of $m(x, v) = \mathbb{E}[Y|X = x, v = v]$ with respect to x is a well defined object. Note that by Assumption 1,

$$m(X, v) = \int g(X, \varepsilon) F_{\varepsilon|X, v}(d\varepsilon|X, v) = \int g(X, \varepsilon) F_{\varepsilon|v}(d\varepsilon|v). \quad (3.8)$$

Assuming that we can differentiate under the integral, it then follows that

$$\partial m(X, v)/\partial x = \int g_x(X, \varepsilon) F_{\varepsilon|v}(d\varepsilon|v), \quad (3.9)$$

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

$$\begin{aligned} \delta &= \mathbb{E}[g_x(X, \varepsilon)] = \mathbb{E} \left[\int g_x(X, \varepsilon) F_{\varepsilon|X, v}(d\varepsilon|X, v) \right] \\ &= \mathbb{E} \left[\int g_x(X, \varepsilon) F_{\varepsilon|v}(d\varepsilon|v) \right] = \mathbb{E} \left[\frac{\partial}{\partial x} m(X, v) \right], \end{aligned} \quad (3.10)$$

where the third equality follows by Assumption 1.

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

THEOREM 3: (*Identification of the limit policy and average derivative*). *Suppose that Assumption 1 is satisfied. If the support of $(\ell(X), v)$ is a subset of the support of (X, v) then $\gamma = \mathbb{E}[g(\ell(X), \varepsilon) - Y]$ 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 $\Delta > 0$, $\mathbb{E}[\int \sup_{\|x-X\| \leq \Delta} \|g_x(x, \varepsilon)\| F_{\varepsilon|v}(d\varepsilon|v)]$ exists, then $\delta = \mathbb{E}[g_x(X, \varepsilon)]$ 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, by linearity of T the order of integration and transformation can be interchanged to obtain, from equation (3.8),

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

Taking expectations of both sides we find that

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

This formula leads to the following general identification result:

THEOREM 4: (*Identification of expectations of linear functions*): Suppose that Assumption 1 is satisfied, that $T(m(\cdot, v), X)$ is a well defined random variable, $\mathbb{E}[T(m(\cdot, v), X)]$ exists, and $T(\int g(\cdot, \varepsilon)F_{\varepsilon|v}(d\varepsilon|v), X) = \int T(g(\cdot, \varepsilon), X)F_{\varepsilon|v}(d\varepsilon|v)$. Then $\mathbb{E}[T(g(\cdot, \varepsilon), X)]$ is identified.

Theorem 4 is a special case of this result with $T(h(\cdot), x) = \partial h(x)/\partial x$ and $T(h(\cdot), x) = h(\ell(x))$. Another example would be an integral of g over elements of x .

Identification results for other objects were previously given in the literature. Altonji and Matzkin (2005) did so for the LAR and Chamberlain (1984), Blundell and Powell (2003), and Wooldridge (2002) for the ASF. These results are potentially useful for the triangular nonparametric simultaneous equations model, where we have shown that $v = F_{X_1|Z}(X_1, Z)$ is a control variable. Also, we give bounds for the ASF when Assumption 2 is not satisfied. For these reasons we present a brief review of their identification here, giving (previously known) explicit equations for their identification in terms of observables.

If Assumptions 1 and 2 are satisfied then for all $x \in \mathcal{X}$, the average structural function is given by

$$\mu(x) = \int g(x, \varepsilon)F_{\varepsilon}(d\varepsilon) = \int [\int g(x, \varepsilon)F_{\varepsilon|v}(d\varepsilon|v)]F_v(dv) = \int m(x, v)F_v(dv).$$

where the second equality follows by iterated expectations, the third by equation (3.8), and identification of $m(x, v)$ inside the last integral by Assumption 2. Also, the local average response is given by

$$\begin{aligned} \beta(x) &= \int g_x(x, \varepsilon)F_{\varepsilon|X}(d\varepsilon|x) = \int \int g_x(x, \varepsilon)F_{\varepsilon|v, X}(d\varepsilon|v, x)F_{v|X}(dv|x) \\ &= \int \int g_x(x, \varepsilon)F_{\varepsilon|v}(d\varepsilon|v)F_{v|X}(dv|x) = \int [\partial \int g(x, \varepsilon)F_{\varepsilon|v}(d\varepsilon|v)/\partial x]F_{v|X}(dv|x) \\ &= \int [\partial m(x, v)/\partial x]F_{v|X}(dv|x), \end{aligned}$$

where the second equality holds by iterated expectations, the third by Assumption 1, and the fourth by interchanging differentiation and integration.

Assumption 2 is a rather strong assumption that may only be satisfied on a small set \mathcal{X} . In treatment effects models, where X is binary, it is often not satisfied at all; e.g. see Heckman, Ichimura, Smith, and Todd (1998). 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 ASF. Let \mathcal{V} denote the support of v , $\mathcal{V}(x)$ the support of v conditional on $X = x$, and $P(x) = \int_{\mathcal{V} \cap \mathcal{V}(x)^c} F_v(dv)$. Also let

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

THEOREM 5: *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*

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

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 Y is the share of expenditure on a commodity and so is bounded between zero and one.

Without Assumption 2 it is also possible to derive bounds for the QSF. Replacing Y by $1(Y \leq y)$ in the bounds for the ASF and setting $B_\ell = 0$ and $B_u = 0$ gives a lower bound $G_\ell(y, x)$ and an upper bound $G_u(y, x)$ on the integral of equation (3.5),

$$G_\ell(y, x) = \int_{\mathcal{V}(x)} \Pr(Y \leq y | X = x, v) F_v(dv), \quad G_u(y, x) = G_\ell(y, x) + P(x). \quad (3.11)$$

Inverting these bounds G leads to the bounds for the QSF, given by

$$q_Y^\ell(\tau, x) = \begin{cases} -\infty, & \tau \leq P(x) \\ G_u^{-1}(\tau, x), & \tau > P(x) \end{cases}, \quad q_Y^u(\tau, x) = \begin{cases} G_\ell^{-1}(\tau, x), & \tau < 1 - P(x) \\ +\infty, & \tau \geq 1 - P(x) \end{cases}. \quad (3.12)$$

THEOREM 6: *(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). Blundell, Gosling, Ichimura, and Meghir (2004) have refined the Manski (1994) bounds using monotonicity and other restrictions.

4 Estimation

We follow a multistep approach to estimation from i.i.d. data (Y_i, X_i, Z_i) , $(i = 1, \dots, n)$. 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 $\mathbb{E}[Y|X, v]$ 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

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

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 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

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

Similarly to Blundell and Powell (2002) and Altonji and Matzkin (2005), the average structural function and local average reponse can be estimated by

$$\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n \hat{m}(x, \hat{v}_i), \quad \hat{\beta}(x) = \int \frac{\partial \hat{m}(x, v)}{\partial x} \hat{f}_{v|X}(v|x) dv.$$

where $\hat{f}_{v|X}(v|x)$ is a nonparametric conditional density estimator.

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,

$$\begin{aligned} \hat{\mu}_\ell(x) &= \hat{\mu}(x) + B_\ell \hat{P}(x), \quad \hat{\mu}_u(x) = \hat{\mu}(x) + B_u \hat{P}(x), \\ \hat{\mu}(x) &= \sum_{i=1}^n 1(\hat{v}_i \in \hat{\mathcal{V}}(x)) \hat{m}(x, \hat{v}_i)/n, \quad \hat{P}(x) = \sum_{i=1}^n 1(\hat{v}_i \notin \hat{\mathcal{V}}(x))/n. \end{aligned}$$

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 (3.12) 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) goes 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 $X = Z + \eta$ 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

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

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 7: *If $X = Z + \eta$ where Z and η are normally distributed and independent, then for $R^2 = \text{Var}(Z) / [\text{Var}(X)]$ 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, which seems to give better results in practice.

To facilitate describing both steps of each estimator we establish a little 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

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

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 $(X_{1i}, Z_i), i = 1, \dots, n$ on X_1 and Z and a positive bandwidth h_1 an estimator of $F_{X_1|Z}(x|z)$ is

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

Then $\hat{v}_i, (i = 1, \dots, n)$ can be calculated as described above. 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

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

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 $p^K(w) = (p_{1K}(w), \dots, p_{KK}(w))'$ denote an $K \times 1$ vector of approximating functions, such as power series or splines, and $p_i = p(\hat{W}_i)$. Let $\tilde{a}_Y^K(w)$ denote the series estimator obtained as the predicted value from regressing Y_i on p_i , that is

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

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 control function is given choosing $w = z$ and calculating

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

Then $\hat{v}_i, (i = 1, \dots, n)$ can be calculated as described above. 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

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

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. 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 (2004). 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. (2004), 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 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 function estimates $\hat{v}_i = \hat{F}_{X|Z}(X_i|Z_i)$. This figure suggests that Assumption 2 may be satisfied 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. 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 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 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 we present here the asymptotic theory for a series estimator of the ASF in the triangular model. This result is 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

function. 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 function. 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 function.

ASSUMPTION 6.1: $Z_i \in \mathfrak{R}^{r_1}$ has compact support and $F_{X|Z}(x|z)$ is continuously differentiable with respect to x 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 8: *If Assumption 6.1 is satisfied,*

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

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:

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 $\beta(w)$, in order to obtain an approximation rate for the second step.

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

We also 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. Let $w = (x, v)$

THEOREM 9: If Assumptions 6.1 - 6.4 are satisfied and $K_2^2 K_v^{\alpha+2} (K_1/n + K_1^{1-2d_1/r}) \rightarrow 0$ then

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

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.

THEOREM 10: If Assumptions 6.1 - 6.5 are satisfied and $K_2^2 K_v^{2+2\alpha} (K_1/n + K_1^{1-2d_1/r}) \rightarrow 0$,

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

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 with $K \geq K_v^s/C$ and $K_x \leq CK_v^{s-1}$. In that case we will have

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

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

$$\begin{aligned} n^{[(1+\alpha+s)/d_2] + (r/2d_1) - 1} &\longrightarrow 0, \\ \int [\hat{\mu}(x) - \mu(x)]^2 F_X(dx) &= O_p(n^{-2(d_2-\alpha-1)/(2d_2+s-1)} + n^{[(1+\alpha)/d_2] + (r/2d_1) - 1}). \end{aligned}$$

The rate hypothesis means that $[(1 + \alpha + s)/d_2] + (r/2d_1) < 1$, which requires that $m(w)$ have more than $1 + \alpha + s$ derivatives and that $F_{X_1|Z}(x_1|z)$ have more than $r/2$ derivatives.

In the Gaussian example of Lemma 8, $d_2 - s > 1 + \alpha$ is equivalent to $d_2 - s > 1/R^2$. In microeconomic applications, where reduced form r-squareds are low, this condition would require many derivatives to exist.

We emphasize that these results only provide a bound on the convergence rate of the estimators. Some improvements may be possible, although currently it seems not to be known whether series estimators can attain optimal uniform rates, see deJong (2003). It would be good to know what the best attainable rate is, but that does not yet seem to be known when the density goes to zero at the boundary, and derivation of such rates is beyond the scope of this paper.

For comparison purposes consider 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 the ASF is $g(X)$. 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, the estimator of Severini and Tripathi (2003), which only uses the conditional mean restriction $E[\varepsilon|Z] = 0$, will converge at a rate that is slower than any power of n (when the approximation error declines as a power of K , as we have assumed here). 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. Using the triangular structure allows one to get a rate that is a power of n , but not quite as fast if one knew and imposed additivity in X and ε .

7 Conclusion

In this paper we presented several identification results for a triangular simultaneous equations model without additivity. Relaxing the additivity assumption is important because such assumptions rarely follow from economic theory. Moreover, economic theory often implies that unless models are non-additive in unobserved components, regressors will be exogenous. Exploiting these identification results we develop estimators for the effects of policies of interest and for the underlying structural functions themselves.

A Appendix

PROOF OF EQ. (2.3): Here we show this equation at z_0, x_0, τ_0 , and $v_0 = F_{X|Z}(x_0|z_0)$ when $Q_{Y|XZ}(\tau, x, z)$ and $F_{X|Z}(x|z)$ are continuously differentiable in a neighborhood of (x_0, z_0) ,

$\nabla_z F_{X|Z}(x_0|z_0) \neq 0$, $\nabla_x F_{X|Z}(x_0|z_0) \neq 0$, and $(X, F_{X|Z}(X|Z))$ is a one-to-one transformation of (X, Z) . 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

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

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,

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

where the third equality follows by condition ii) and the fourth by independence of Z and η . . 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 monontonic 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

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

Q.E.D.

PROOF OF THEOREM 2: 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 3: By the fact that $g(x, \varepsilon)$ continuously differentiable and the integrability condition, it follows that $m(x, v)$ is differentiable and eq. (3.9) is satisfied. Then by eq. (3.10) 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 (3.7) Q.E.D.

PROOF OF THEOREM 4: By eq. (3.8) $m(X, v) = \mathbb{E}[Y|X, v] = \int g(X, \varepsilon)F_{\varepsilon|v}(d\varepsilon|v)$. Then by $T(\int g(\cdot, \varepsilon)F_{\varepsilon|v}(d\varepsilon|v), X) = \int T(g(\cdot, \varepsilon), X)F_{\varepsilon|v}(d\varepsilon|v)$ and iterated expectations,

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

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

PROOF OF THEOREM 5: By the definition of $\mathcal{V}(x)$ and Assumption 1, on a set of x with probability 1, integrating eq. (3.8) 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

$$\mu_\ell(x) \leq \int \int_{\mathcal{V}} g(x, e)F_{\varepsilon|v}(de|v)F_v(dv) = \mu(x) \leq \mu_u(x). Q.E.D.$$

PROOF OF THEOREM 6: Note first that by Assumption 1,

$$\begin{aligned} G_\ell(y, x) &= \int_{\mathcal{V}(x)} \Pr(Y \leq y|X = x, v)F_v(dv) = \int_{\mathcal{V}(x)} \Pr(g(x, \varepsilon) \leq y|X = x, v)F_v(dv) \\ &= \int_{\mathcal{V}(x)} \Pr(g(x, \varepsilon) \leq y|v)F_v(dv). \end{aligned}$$

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

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

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

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

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

C Proofs of Consistency

Throughout the remainder of the Appendix, C will denote a generic positive constant that may be different in different uses. Also, with probability approaching one will be abbreviated as w.p.a.1, positive semi-definite as p.s.d., positive definite as p.d., $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, and $A^{1/2}$ will denote the minimum and maximum eigenvalues, and square root, of respectively of a symmetric matrix A . Let \sum_i denote $\sum_{i=1}^n$. Also, let CS, M, and T refer to the Cauchy-Schwartz, Markov, and triangle inequalities, respectively. Also, let CM refer to the following well known result: *If $\mathbb{E}[|Y_n||Z_n] = O_p(r_n)$ then $Y_n = O_p(r_n)$.*

Before proving Theorem 8, we prove some preliminary results. Let $q_i = q^L(Z_i)$, $\eta_{ij} = 1(X_{1j} \leq X_{1i}) - F_{X_1|Z}(X_{1i}|Z_j)$.

LEMMA B1: *For $Z = (Z_1, \dots, Z_n)$ and $L \times 1$ vectors of functions $b_i(Z)$, ($i = 1, \dots, n$), if $\sum_{i=1}^n b_i(Z)' \hat{Q} b_i(Z)/n = O_p(r_n)$ then*

$$\sum_{i=1}^n \{b_i(Z)' \sum_{j=1}^n q_j \eta_{ij} / \sqrt{n}\}^2 / n = O_p(r_n).$$

Proof: Note that $|\eta_{ij}| \leq 1$. Consider $j \neq k$ and suppose without loss of generality that $j \neq i$ (otherwise reverse the role of j and k because we cannot have $i = j$ and $i = k$). By independence of the observations,

$$\begin{aligned} \mathbb{E}[\eta_{ij}\eta_{ik}|Z] &= \mathbb{E}[\mathbb{E}[\eta_{ij}\eta_{ik}|Z, X_i, X_k]|Z] = \mathbb{E}[\eta_{ik}\mathbb{E}[\eta_{ij}|Z, X_i, X_k]|Z] = \mathbb{E}[\eta_{ik}\mathbb{E}[\eta_{ij}|Z_j, Z_i, X_i]|Z] \\ &= \mathbb{E}[\eta_{ik}\{\mathbb{E}[1(X_{1j} \leq X_{1i})|Z_j, Z_i, X_i] - F_{X_1|Z}(X_{1i}|Z_j)\}|Z] = 0. \end{aligned}$$

Therefore, it follows that

$$\begin{aligned} \mathbb{E}[\sum_{i=1}^n \{b_i(Z)' \sum_{j=1}^n q_j \eta_{ij} / \sqrt{n}\}^2 / n | Z] &\leq \sum_{i=1}^n b_i(Z)' \{ \sum_{j,k=1}^n q_j \mathbb{E}[\eta_{ij}\eta_{ik}|Z] q'_k / n \} b_i(Z) / n = \\ \sum_{i=1}^n b_i(Z)' \{ \sum_{j=1}^n q_j \mathbb{E}[\eta_{ij}^2 | Z] q'_j / n \} b_i(Z) / n &\leq \sum_{i=1}^n b_i(Z)' \hat{Q} b_i(Z) / n, \end{aligned}$$

so the conclusion follows by CM. Q.E.D.

LEMMA B2: (Lorentz, 1986, p. 90, Theorem 8). If Assumption 6.1 is satisfied then there exists C such that for each x there is $\gamma(x)$ with $\sup_{z \in Z} |F_{X_1|Z}(x|z) - p^{K_1}(z)' \gamma(x)| \leq CK_1^{-d_1/r_1}$.

LEMMA B3: If Assumption 6.2 is satisfied then for each K there exists a nonsingular constant matrix B such that $\hat{p}^{K_2}(w) = Bp^{K_2}(w)$ satisfies $E[\hat{p}^{K_2}(w_i)\hat{p}^{K_2}(w_i)'] = I_K$, $\sup_{w \in \mathcal{W}} \|\hat{p}^{K_2}(w)\| \leq CK_v^\alpha K_2$, $\sup_{w \in \mathcal{W}} \|\partial \hat{p}^{K_2}(w)/\partial v\| \leq CK_v^{\alpha+2} K_2$, and $\sup_{[0,1]} \|\hat{p}^{K_v}(t)\| \leq CK_v^{1+\alpha}$.

Proof: For $u \in [0, 1]$, let $P_j^\alpha(u)$ be the j^{th} orthonormal polynomial with respect to the weight $u^\alpha(1-u)^\alpha$. Denote $\mathcal{X} = \Pi_{\ell=1}^{s-1}[\underline{x}_\ell, \bar{x}_\ell]$. By the fact that the order of the power series is increasing and that all terms of a given order are included before a term of higher order, for each k and $\lambda(k, \ell)$ with $p_k(w) = \Pi_{\ell=1}^s w_\ell^{\lambda(k, \ell)}$, there exists b_{kj} , ($j \leq k$), such that

$$\sum_{j=1}^k b_{kj} p_j(w) = \Pi_{\ell=1}^{s-1} P_{\lambda(k, \ell)}^0([x_\ell - \underline{x}_\ell]/[\bar{x}_\ell - \underline{x}_\ell]) P_{\lambda(k, s)}^\alpha(t).$$

Let B_k denote a $K \times 1$ vector $B_k = (b_{k1}, \dots, b_{kk}, 0)'$, $b_{kk} \neq 0$ where 0 is a $K - k$ dimensional zero vector and let \bar{B} be the $K \times K$ matrix with k^{th} row B_k' . Then \bar{B} is a lower triangular matrix with nonzero diagonal elements and so is nonsingular. As shown in Andrews (1991) there is C such that $|P_j^\alpha(u)| \leq C(j^{\alpha+1/2} + 1) \leq Cj^{\alpha+1/2}$ and $|dP_j^\alpha(u)/du| \leq Cj^{\alpha+5/2}$ for all $u \in [0, 1]$ and $j \in \{1, 2, \dots\}$. Then for $\bar{p}^K(w) = \bar{B}p^K(w)$, it follows that $|\bar{p}_k(w)| \leq C\lambda(k, s)^{\alpha+1/2} \Pi_{\ell=1}^{s-1} \lambda(k, \ell)^{1/2}$, so that $\|\bar{p}^K(w)\| \leq CK_\eta^\alpha K$, and $\sup_{w \in \mathcal{W}} \|\partial \bar{p}^K(w)/\partial t\| \leq CK_\eta^{\alpha+2} K$. Then by Assumption 6.2, it follows that $\Omega_K = E[\bar{p}^K(w_i)\bar{p}^K(w_i)'] \geq CI_K$. Let $\tilde{B} = \Omega_K^{-1/2}$, and define $\tilde{p}^K(w) = \tilde{B}\bar{p}^K(w)$. Then $\|\tilde{p}^K(w)\| = \sqrt{\tilde{p}^K(w)' \tilde{p}^K(w)} \leq \sqrt{\bar{p}^K(w)' \Omega^{-1} \bar{p}^K(w)} \leq C\|\bar{p}^K(w)\|$ and an analogous inequality holds for $\|\partial \tilde{p}^K(w)/\partial t\|$, giving the conclusion. Q.E.D.

Henceforth define $\zeta = CK_v^\alpha K_2$ and $\zeta_1 = CK_v^{\alpha+2} K_2$. Also, since the estimator is invariant to nonsingular linear transformations of $p^{K_2}(w)$, we can assume that the conclusion of Lemma B3 is satisfied with $p^{K_2}(w)$ replacing $\hat{p}^{K_2}(w)$.

Proof of Theorem 4: Let $\delta_{ij} = F_{X_1|Z}(X_{1i}|Z_j) - q_j' \gamma^{K_1}(X_{1i})$, with $|\delta_{ij}| \leq K_1^{-d_1/r_1}$ by Lemma B2. Then for $\tilde{v}_i = \tilde{F}(X_{1i}|Z_i)$

$$\tilde{v}_i - \bar{v}_i = \Delta_i^I + \Delta_i^{II} + \Delta_i^{III},$$

where

$$\Delta_i^I = q_i' \hat{Q}^- \sum_{j=1}^n q_j v_{ij} / n, \Delta_i^{II} = q_i' \hat{Q}^- \sum_{j=1}^n q_j \delta_{ij} / n, \Delta_i^{III} = -\delta_{ii}.$$

Note that $|\Delta_i^{III}| \leq CK_1^{-d_1/r}$ by Lemm B2. Also, by \hat{Q} p.s.d. and symmetric there exists a diagonal matrix of eigenvalues Λ and an orthonormal matrix B such that $\hat{Q} = B\Lambda B'$. Let Λ^- denote the diagonal matrix of inverse of nonzero eigenvalues and zeros and $\hat{Q}^- = B\Lambda^-B'$. Then $\sum_i q'_i \hat{Q}^- q_i = \text{tr}(\hat{Q}^- \hat{Q}) \leq CL$. By CS and Assumption 6.1,

$$\begin{aligned} \sum_{i=1}^n (\Delta_i^{II})^2/n &\leq \sum_{i=1}^n (q'_i \hat{Q}^- q_i \sum_{j=1}^n \delta_{ij}^2/n)/n \leq C \sum_{i=1}^n (q'_i \hat{Q}^- q_i) L^{-2d_1}/n \\ &= CK_1^{-2d_1/r} \text{tr}(\hat{Q}^- \hat{Q}) \leq CK_1^{1-2d_1/r}. \end{aligned}$$

Note that for $b_i(Z) = q'_i \hat{Q}^- / \sqrt{n}$ we have

$$\sum_{i=1}^n b_i(Z)' \hat{Q} b_i(Z)/n = \text{tr}(\hat{Q} \hat{Q}^- \hat{Q} \hat{Q}^-)/n = \text{tr}(\hat{Q} \hat{Q}^-)/n \leq CK_1/n = O_p(K_1/n),$$

so it follows by Lemma A1 that $\sum_{i=1}^n (\Delta_i^I)^2/n = O_p(L/n)$. The conclusion then follows by T and by $|\tau(\tilde{v}) - \tau(v)| \leq |\tilde{v} - v|$, which gives $\sum_i (\tilde{v}_i - \bar{v}_i)^2/n \leq \sum_i (\tilde{v}_i - \bar{v}_i)^2/n$. Q.E.D.

Before proving other results we give some useful lemmas. For these results let $p_i = p^{K_2}(w_i)$, $\hat{p}_i = p^{K_2}(\hat{w}_i)$, $p = [p_1, \dots, p_n]$, $\hat{p} = [\hat{p}_1, \dots, \hat{p}_n]$, $\tilde{P} = \hat{p}'\hat{p}/n$, $\tilde{P} = p'p/n$, $P = \mathbb{E}[p_i p'_i]$. Note that in the statement of these results we allow \hat{v}_i and \bar{v}_i to be vectors. Also, as in Newey (1997) it can be shown that without loss of generality we can set $P = I_{K_2}$.

LEMMA B4: *If the hypotheses of Theorem 1 are satisfied then $\mathbb{E}[Y|X, Z] = m(X, v)$.*

Proof: By the proof of Theorem 1, $v = F_{X_1|Z}(X_1|Z)$ is a function of X_1 and Z that is invertible in X_1 with inverse $X_1 = \tilde{h}(Z, v)$ where $\tilde{h}(z, v)$ is the inverse of $F_{X_1|Z}(x|z)$ in its first argument. By independence of Z and (ε, η) , ε is independent of Z conditional on v , so that by eq. (3.8),

$$\begin{aligned} \mathbb{E}[Y|X, Z] &= \mathbb{E}[Y|Z, v] = \mathbb{E}[g(\tilde{h}(Z, v), \varepsilon)|Z, v] = \int g(\tilde{h}(Z, v), e) F_{\varepsilon|Z, v}(de|Z, v) \\ &= \int g(\tilde{h}(Z, v), e) F_{\varepsilon|v}(de|v) = m(X, v). \text{Q.E.D.} \end{aligned}$$

Let $u_i = Y_i - m(X_i, v_i)$, and let $u = (u_1, \dots, u_n)'$.

LEMMA B5: *If $\sum_i \|\hat{v}_i - v_i\|^2/n = O_p(\Delta_n^2)$ and Assumptions 6.1 - 6.4 are satisfied then*

$$\begin{aligned} (i), \|\tilde{P} - P\| &= O_p(\zeta \sqrt{K_2/n}); (ii) \|p'u/n\| = O_p(\sqrt{K_2/n}), (iii) \|\hat{p} - p\|^2/n = O_p(\zeta_1^2 \Delta_n^2), \\ (iv), \|\hat{P} - \tilde{P}\| &= O_p(\zeta_1^2 \Delta_n^2 + \sqrt{K_2} \zeta_1 \Delta_n); (v) \|(\hat{p} - p)'u/n\| = O_p(\zeta_1 \Delta_n / \sqrt{n}). \end{aligned}$$

Proof: The first two results follow as in eq. A.1 and p. 162 of Newey (1997). For (iii) a mean value expansion gives $\hat{p}_i = p_i + [\partial p^{K_2}(\tilde{w}_i)/\partial v](\hat{v}_i - \bar{v}_i)$, where $\tilde{w}_i = (x_i, \hat{v}_i)$ and \hat{v}_i lies in

between \hat{v}_i and \bar{v}_i . Since \hat{v}_i and \bar{v}_i lie in $[0, 1]$, it follows that $\hat{v}_i \in [0, 1]$ so that by Lemma B3, $\|\partial p^{K_2}(\tilde{w}_i)/\partial v\| \leq C\zeta_1$. Then by CS, $\|\hat{p}_i - p_i\| \leq C\zeta_1|\hat{v}_i - \bar{v}_i|$. Summing up gives

$$\|\hat{p} - p\|^2/n = \sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n = O_p(\zeta_1^2 \Delta_n^2). \quad (\text{C.1})$$

For (iv), by Lemma B3, $\sum_{i=1}^n \|p_i\|^2/n = O_p(\mathbb{E}[\|p_i\|^2]) = \text{tr}(I_{K_2}) = K_2$. Then by T, CS, and M,

$$\begin{aligned} \|\hat{P} - \tilde{P}\| &\leq \sum_{i=1}^n \|\hat{p}_i \hat{p}'_i - p_i p'_i\|/n \leq \sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n + 2\left(\sum_{i=1}^n \|\hat{p}_i - p_i\|^2/n\right)^{1/2} \left(\sum_{i=1}^n \|p_i\|^2/n\right)^{1/2}. \\ &= O_p(\zeta_1^2 \Delta_n^2 + \sqrt{K_2} \zeta_1 \Delta_n). \end{aligned}$$

Finally, for (v), for $Z = (Z_1, \dots, Z_n)$ and $X = (X_1, \dots, X_n)$, it follows from Lemma B4, Assumption 6.4, and independence of the observations that $\mathbb{E}[uu'|X, Z] \leq CI_n$, so that by p and \hat{p} depending only on Z and X ,

$$\begin{aligned} \mathbb{E}[\|(\hat{p} - p)'u/n\|^2|X, Z] &= \text{tr}\{(\hat{p} - p)' \mathbb{E}[uu'|X, Z](\hat{p} - p)/n^2\} \\ &\leq C\|\hat{p} - p\|^2/n^2 = O_p(\zeta_1^2 \Delta_n^2/n). \end{aligned}$$

Q.E.D.

LEMMA B6: *If Assumptions 6.1-6.4 are satisfied and $K_2 \zeta_1^2 \Delta_n^2 \rightarrow 0$, then w.p.a.1, $\lambda_{\min}(\hat{P}) \geq C$, $\lambda_{\min}(\tilde{P}) \geq C$.*

Proof: By Lemma B3 and $\zeta_1^2 K_2/n \leq CK_2 \zeta_1^2 K_1/n$, we have $\|\hat{P} - \tilde{P}\| \xrightarrow{p} 0$ and $\|\tilde{P} - P\| \xrightarrow{p} 0$, so the conclusion follows as on p. 162 of Newey (1997). Q.E.D.

Let $m = (m(w_1), \dots, m(w_n))'$, and $\hat{m} = (m(\hat{w}_1), \dots, m(\hat{w}_n))'$.

LEMMA B7: *If $\sum_i \|\hat{v}_i - v_i\|^2/n = O_p(\Delta_n^2)$, Assumptions 6.1 - 6.4 are satisfied, $\sqrt{K_2} \zeta_1 \Delta_n \rightarrow 0$, and $K_2 \zeta^2/n \rightarrow 0$ then for $\tilde{\alpha} = \hat{P}^{-1} \hat{p}' \hat{m}/n$, $\bar{\alpha} = \hat{P}^{-1} \hat{p}' m/n$,*

$$(i) \|\tilde{\alpha} - \bar{\alpha}\| = O_p(\sqrt{K_2/n}), (ii) \|\tilde{\alpha} - \bar{\alpha}\| = O_p(\Delta_n), (iii) \|\tilde{\alpha} - \alpha^{K_2}\| = O_p(K_2^{-d_2/s}).$$

Proof: For (i)

$$\begin{aligned} \mathbb{E}[\|\hat{P}^{1/2}(\tilde{\alpha} - \bar{\alpha})\|^2|X, Z] &= \mathbb{E}[u' \hat{p} \hat{P}^{-1} \hat{p}' u/n^2|X, Z] = \text{tr}\{\hat{P}^{-1/2} \hat{p}' \mathbb{E}[uu'|X, Z] \hat{p} \hat{P}^{-1/2}\}/n^2 \\ &\leq C \text{tr}\{\hat{p} \hat{P}^{-1} \hat{p}'\}/n^2 \leq C \text{tr}(I_K)/n = CK/n. \end{aligned}$$

Since by Lemma B6, $\lambda_{\min}(\hat{P}) \geq C$ w.p.a.1, this implies that $\mathbb{E}[\|\tilde{\alpha} - \bar{\alpha}\|^2|X, Z] \leq CK/n$. Similarly, for (ii),

$$\|\hat{P}^{1/2}(\tilde{\alpha} - \bar{\alpha})\|^2 \leq C(\hat{m} - m)' \hat{p} \hat{P}^{-1} \hat{p}' (\hat{m} - m)/n^2 \leq C\|\hat{m} - m\|^2/n = O_p(\Delta_n^2),$$

which follows from $m(w)$ being Lipschitz in η , so that also $\|\tilde{\alpha} - \bar{\alpha}\|^2 = O_p(\Delta_n^2)$. Finally for (iii),

$$\begin{aligned} \|\hat{P}^{1/2}(\tilde{\alpha} - \alpha^{K_2})\|^2 &= \|\tilde{\alpha} - \hat{P}^{-1}\hat{p}'\hat{p}\alpha^{K_2}/n\|^2 \leq C(\hat{m} - \hat{p}'\alpha^{K_2})'\hat{P}^{-1}\hat{p}'(\hat{m} - \hat{p}'\alpha^{K_2})/n^2 \\ &\leq \|\hat{m} - \hat{p}\alpha^{K_2}\|^2/n \leq C \sup_{w \in \mathcal{W}} |m_0(w) - p^K(w)'\alpha^{K_2}|^2 = O_p(K_2^{-2d_2/s}), \end{aligned}$$

so that $\|\hat{P}^{1/2}(\tilde{\alpha} - \alpha^{K_2})\|^2 = O_p(K_2^{-2d_2/s})$. Q.E.D.

Proof of Theorem 9: Note that by Theorem 4, for $\Delta_n^2 = K_1/n + K_1^{1-2d_1/r_1}$, we have $\sum_i \|\hat{v}_i - v_i\|^2/n = O_p(\Delta_n^2)$, so by $K_2\zeta^2/n \leq CK_2\zeta_1^2K_1/n$ the hypotheses of Lemma B7 are satisfied. Also by Lemma B7 and T, $\|\hat{\alpha} - \alpha^K\|^2 = O_p(K_2/n + K_2^{-2d_2/r_2} + \Delta_n^2)$. Then

$$\begin{aligned} \int [\hat{m}(w) - m(w)]^2 F_w(dw) &= \int [p^{K_2}(w)'(\hat{\alpha} - \alpha^{K_2}) + p^{K_2}(w)'\alpha^{K_2} - m(w)]^2 F_w(dw) \\ &\leq C\|\hat{\alpha} - \alpha^{K_2}\|^2 + CK_2^{-2d_2/r_2} = O_p(K_2/n + K_2^{-2d_2/r_2} + \Delta_n^2). \end{aligned}$$

For the second part of Theorem 9

$$\begin{aligned} \sup_{w \in \mathcal{W}} |\hat{m}(w) - m(w)| &= \sup_{w \in \mathcal{W}} |p^K(w)'(\hat{\alpha} - \alpha^K) + p^K(w)'\alpha^K - \beta(w)| \\ &= O_p(\zeta(K_2/n + K_2^{-2d_2/r_2} + \Delta_n^2)^{1/2}) + O_p(K^{-d_2/r_2}) \\ &= O_p(\zeta(K_2/n + K_2^{-2d_2/r_2} + \Delta_n^2)^{1/2}). \end{aligned}$$

Q.E.D.

Proof of Theorem 10: Let $\bar{p} = \int_0^1 p^{K_v}(t)dt$ and note that by Lemma B3, $\bar{p}'\bar{p} \leq CK_v^{2+2\alpha}$. Also,

$$\bar{p}(x) \stackrel{\text{def}}{=} \int_0^1 p^K(w)dt = p^{K_x}(x) \otimes \bar{p}. \quad (\text{C.2})$$

As above, $\mathbb{E}[uu'|X, Z] \leq CI_n$, so that by Fubini's Theorem,

$$\begin{aligned} \mathbb{E}\left[\int \{\bar{p}(x)'(\hat{\alpha} - \bar{\alpha})\}^2 F_X(dx) | X, Z\right] &= \int \{\bar{p}(x)'\hat{P}^{-1}\hat{p}'\mathbb{E}[uu'|X, Z]\hat{p}\hat{P}^{-1}\bar{p}(x)\} F_X(dx)/n^2 \\ &\leq C \int \bar{p}(x)'\hat{P}^{-1}\bar{p}(x) F_X(dx)/n \leq C\mathbb{E}[\bar{p}(X)'\bar{p}(X)]/n \\ &= C\{\mathbb{E}[p^{K_x}(X)'\bar{p}^{K_x}(X)](\bar{p}'\bar{p})\}/n = K_x K_v^{2+2\alpha}/n. \end{aligned}$$

It then follows by CM that $\int \{\bar{p}(x)'(\hat{\alpha} - \bar{\alpha})\}^2 F_X(dx) = O_p(K_x K_v^{2+2\alpha}/n)$.

$$\int \bar{p}(x)\bar{p}(x)' F_X(dx) = I_{K_x} \otimes \bar{p}\bar{p}' \leq CI_{K_x}\bar{p}'\bar{p} \leq CI_{K_x}K_v^{2+2\alpha},$$

so that by Lemma B7 and T,

$$\begin{aligned} \int \{\bar{p}(x)'(\bar{\alpha} - \alpha^K)\}^2 F_X(dx) &\leq (\bar{\alpha} - \alpha^K)' \int \bar{p}(x)\bar{p}(x)' F_X(dx) (\bar{\alpha} - \alpha^K) \\ &\leq CK_\eta^{2+2a} \|\bar{\alpha} - \alpha^K\|^2 = O_p(K_\eta^{2+2a}(K^{-2d/s} + \Delta_n^2)). \end{aligned}$$

Also, by CS,

$$\int \{\bar{p}(x)' \alpha^K - \mu(x)\}^2 F_X(dx) \leq \int \int_0^1 \{p^K(w)' \alpha - \beta(w)\}^2 d\eta F_X(dx) = O(K^{-2d/s}).$$

Then the conclusion follows by T and

$$\begin{aligned} \int [\hat{\mu}(x) - \mu(x)]^2 F_0(dx) &= \int \{\bar{p}(x)'(\hat{\alpha} - \alpha^K) + \bar{p}(x)' \alpha^K - \mu(x)\}^2 F_X(dx) \\ &= O_p(K_x/n + K_x^{-4d} + \Delta_n^2) + O_p(K_x^{-4d}). \quad Q.E.D. \end{aligned}$$

REFERENCES

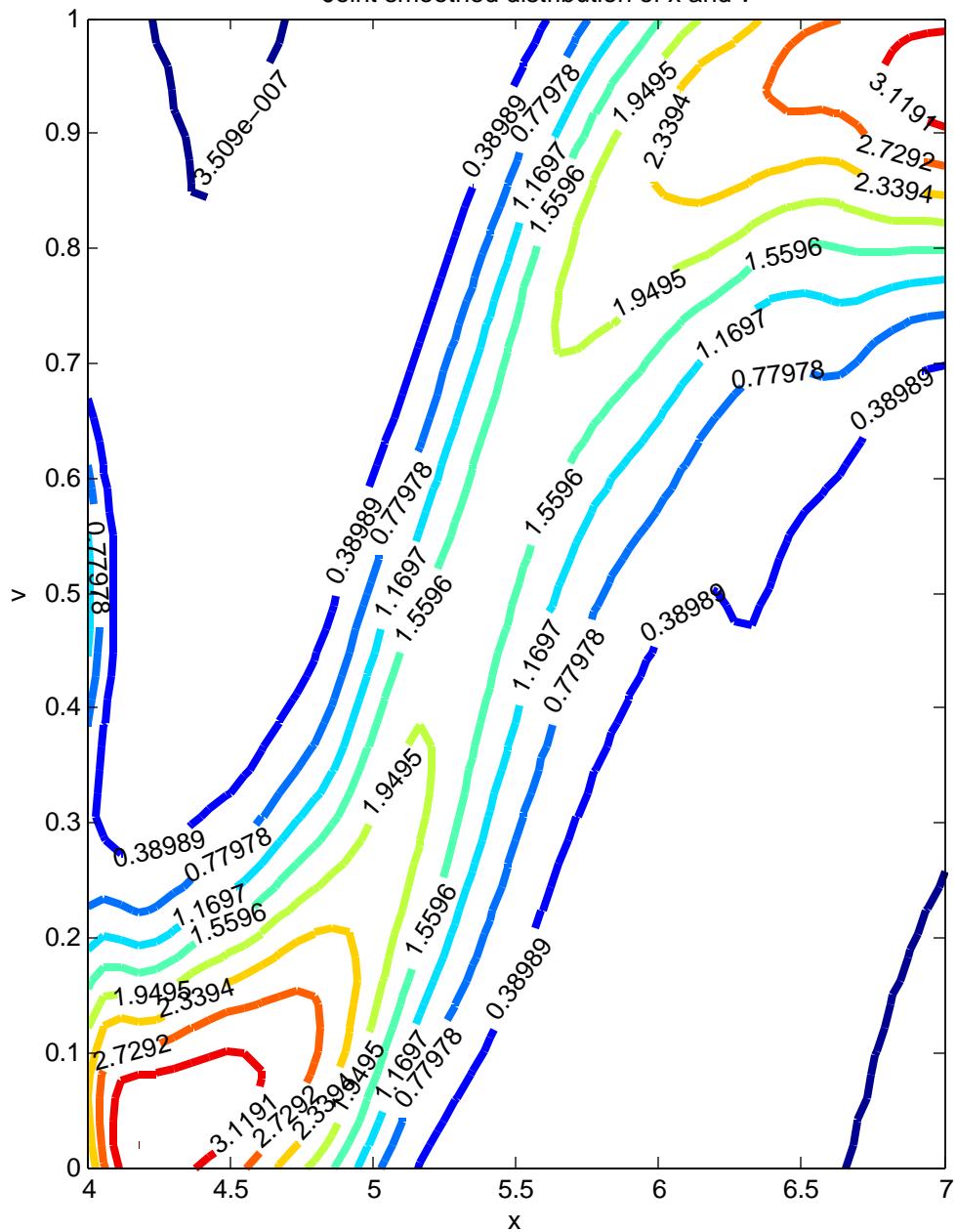
- Altonji, J., and R. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Econometrica* 73, 1053-1102.
- Amemiya, T. (1982), "Two Stage Least Absolute Deviations Estimators," *Econometrica* 50, 689-712.
- Angrist, J., G.W. Imbens, and D. Rubin (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.
- Bajari, P., and L. Benkard (2001), "Demand Estimation with Heterogenous Consumers and Unobserved Product Characteristics: A Hedonic Approach," unpublished paper, Department of Economics, Stanford University.
- Benkard, L., and S. Berry (2006), "On the Nonparametric Identification of Nonlinear Simultaneous Equations Models: Comment on Brown (1983) and Roehrig (1988)," *Econometrica*, 74(5).

- 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 (2004): "Changes in the Distribution of Male and Female Wages Accounting for Unemployment Using Bounds," Institute of Fiscal Studies Working Paper 4/25.
- Brown, D., and R. Matzkin, (1996): "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," mimeo, Northwestern University.
- 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 Intriligator (eds.), *Handbook of Econometrics*, Vol 2.
- Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics* 34, 305-334.
- Chesher, A. (2002), "Semiparametric Identification in Duration Models," Cemmap working paper CWP20/02.
- Chesher, A. (2003), "Identification in Nonseparable Models," *Econometrica* 71(5), 1405-1441.
- Chesher, A. (2005), "Nonparametric Identification Under Discrete Variation," *Econometrica*, 73(5), 1525-1550.
- Chernozhukov, V., and C. Hansen, (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261.
- Chernozhukov, V., G. Imbens, and W. Newey (2007), "Instrumental Variables Estimation of Nonseparable Models," *Journal of Econometrics*, forthcoming.
- Darolles, S., J.-P., Florens, and E. Renault, (2001), "Nonparametric Instrumental Regression," working paper.
- Das, M. (2000): "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.

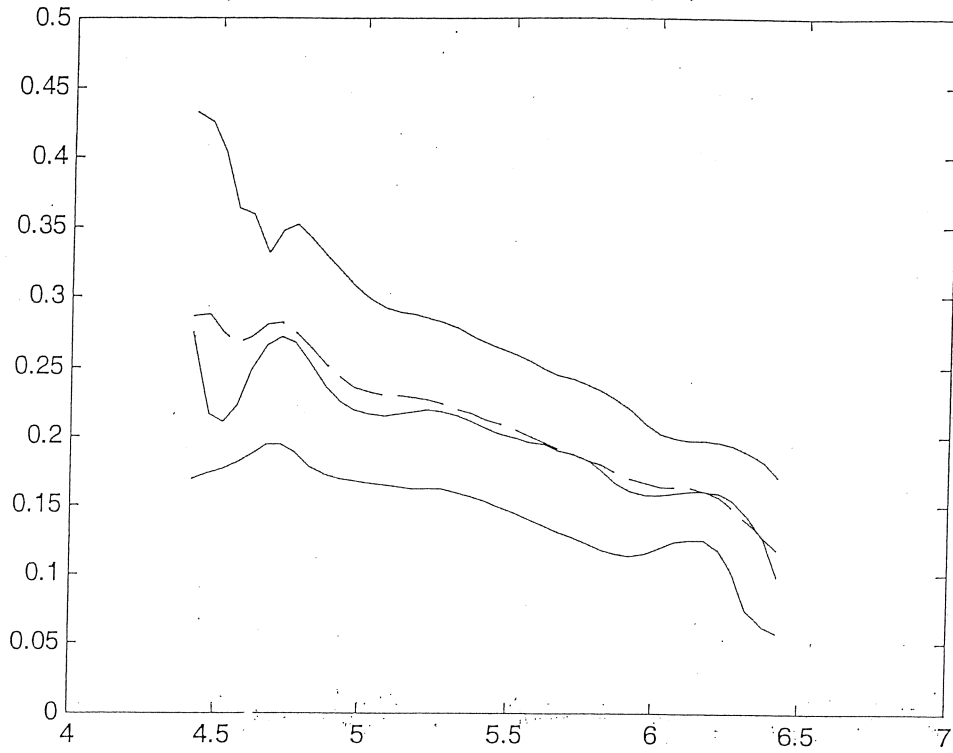
- Doss, H. and R.D. Gill (1992): "An Elementary Approach to Weak Convergence for Quantile Processes, With Applications to Censored Survival Data," *Journal of the American Statistical Association* 87, 869-877.
- Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica* 75(1), 259-276.
- Florens, J.P., J.J. Heckman, C. Meghir, and E. Vytlačil (2004): "Instrumental Variables, Local Instrumental Variables and Control Functions," working paper, UCL.
- Hall, P., J.S. Marron, M.H. Neuman, and D.M. Titterington (1997): "Curve Estimation When the Design Density is Low," *Annals of Statistics* 25, 756-770.
- Hausman, J.A. and W.K. Newey (1995), "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," *Econometrica* 63, 1445-1476.
- Heckman, J., H. Ichimura, J. Smith, P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098. , Cambridge.
- Hengartner, N.W. and O.B. Linton (1996): "Nonparametric Regression Estimation at Design Poles and Zeros," *Canadian Journal of Statistics* 24, 583-591.
- Imbens, G.W. (2005): "Nonadditive Models with Endogenous Regressors," World Congress of the Econometric Society, London.
- Imbens, G.W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62, 467-476.
- Lorentz, G., (1986), *Approximation of Functions*, New York: Chelsea Publishing Company.
- Ma, L. and R. Koenker (2006): "Quantile Regression Methods for Recursive Structural Equation Models," *Journal of Econometrics*, 134(2), 471-506.
- Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80:2, 319-323.
- Manski, C. (1994): "The Selection Problem," in C. Sims (ed.), *Advances in Economics and Econometrics*, Cambridge University Press.
- Manski, C. (1995): *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA.
- Manski, C. (1997): "The Mixing Problem in Program Evaluation," *Review of Economic Studies* 64, 537-553.
- Mark, S, and J. Robins, "Estimating the Causal Effect of Smoking Cessation in the Presence of Confounding Factors Using a Rank-Preserving Structural Failure Time Model," *Statistics in Medicine* 12, 1605-1628.

- Matzkin, R. (1993), "Restrictions of Economic Theory in Nonparametric Models" *Handbook of Econometrics*, Vol IV, Engle and McFadden (eds.)
- Matzkin, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions", *Econometrica* 71, 1339-1375.
- 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. (2006): "Nonparametric Discrete Continuous Choice,"
- 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. (2000): "Nonparametric Regression Estimation Using Weak Separability", University of British Columbia.
- Powell, J., J. Stock, and T. Stoker, "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.
- Robins, J. (1995): "An Analytic Method for Randomized Trials with Informative Censoring: Part 1, *Lifetime Data Analysis* 1, 241-254.
- Roehrig, C. (1988): "Conditions for Identification in Nonparametric and Parametric Models", *Econometrica* 55, 875-891.
- 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.
- Wooldridge, J. (2002): *Econometric Analysis of Cross Section and Panel Data* MIT Press.

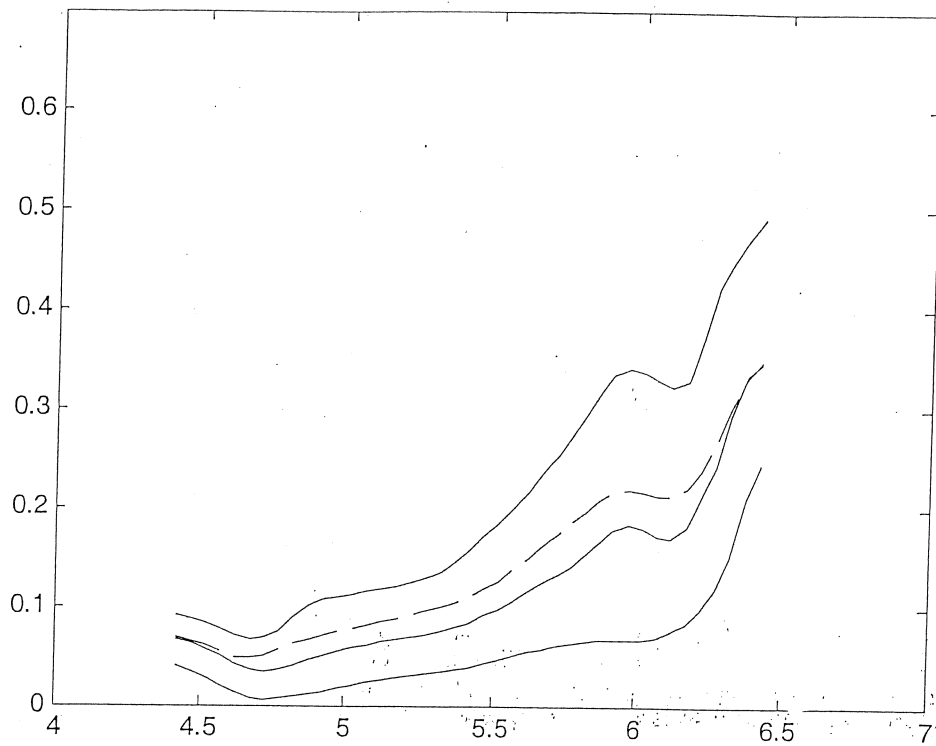
Joint smoothed distribution of x and v



Food Expenditure - ASF and QSF for $\tau = .25, .5, .75$ - Local Linear



Leisure Expenditure - ASF and QSF for $\tau = .25, .5, .75$ - Local Linear



P(x) for $\delta = 0.072005$

