Endogeneity and Instruments in Nonparametric Models

A Discussion of the Papers by Jean-Pierre Florens and by Richard Blundell and James L. Powell

Manuel Arellano

The chapters by Jean-Pierre Florens and by Richard Blundell and James Powell consider the extension of nonlinear models with endogenous regressors to semiparametric and nonparameteric contexts. In my comments, I focus on two aspects. First, I refer to nonlinear implicit structural equations. Second, I provide some comments on the connections between the control variable and the instrumental variable approaches.

1. NONLINEAR IMPLICIT STRUCTURAL EQUATIONS

The literature on parametric nonlinear structural models, beginning with the work of Amemiya (1977), considered the implicit equation formulation

$$f(y, x) \equiv f(w) = u, \tag{1.1}$$

$$E(u|z) = 0, \tag{1.2}$$

where $w = (y, x')$ is a data point and u is an unobservable structural error that is mean independent of a vector z of instrumental variables.

In a semiparametric structural context, this formulation may also provide a useful setup. The model $y = g(x) + u$ in which $g(x)$ is treated as a nonparametric function is a special case of (1.1) with $f(y, x) = y - g(x)$. Such a model is reminiscent of nonparametric regression, but in structural applications other specializations of (1.1) and (1.2) may be relevant.

An example is a time-series consumption Euler equation of the form

$$U'(c_{t+1})r_{t+1} - U'(c_t) = u_{t+1}, \tag{1.3}$$

where $U'(\cdot)$ denotes the marginal utility of consumption and r_{t+1} is the stochastic return on a financial asset (Hansen and Singleton, 1982). Parameters of interest here could be the coefficients of relative risk aversion for different values of c estimated in a nonparametric way (semiparametric models of this type were considered by Gallant and Tauchen, 1989).

In this example there is no left-hand-side variable, but the model imposes a particular structure in the implicit function, namely additivity and monotonicity in $U'(\cdot)$. Structural models often impose not only instrumental variable (IV) conditions but also restrictions on the shape of functions, which may give rise to alternative economically based types of regularization for specific models.

Another situation of interest is when the starting point is an invertible response function,

$$y = H(x, u), \tag{1.4}$$

which can be represented as $f(y, x) = u$, together with the assumption

$$E[c(u)|z] = 0$$

for some function $c(\cdot)$. Identification of $f^*(y, x) = c\,[f(y, x)]$ up to scale affords calculation of the following derivative effects:

$$m\,(y, x) = -\left[\frac{\partial f(y, x)}{\partial y}\right]^{-1} \frac{\partial f(y, x)}{\partial x}. \tag{1.5}$$

The average structural function is given by

$$G\,(x) = E_u\{m[H(x, u), x]\}$$

and may or may not be identified, depending on the nature of the restrictions imposed on $f(y, x)$.

1.1. Discrete Case

If (w, z) is a discrete random vector with finite support, the analysis of the implicit equation IV model is straightforward. The model specifies

$$\sum_{j=1}^{J} f(\xi_j)\,\Pr(w = \xi_j | z = \zeta_\ell) = 0 \quad (\ell = 1, \ldots, L), \tag{1.6}$$

where the supports of w and z are $\{\xi_1, \ldots, \xi_J\}$ and $\{\zeta_1, \ldots, \zeta_L\}$, respectively.

In matrix form we have

$$P\theta = 0, \tag{1.7}$$

where P is an $L \times J$ matrix of conditional probabilities, and θ is the $J \times 1$ vector with elements $\theta_j = f(\xi_j)$. The order condition for identification of θ up to scale is $L \geq J - 1$, and the rank condition is $\operatorname{rank}(P) = J - 1$.

This is a standard generalized method of moments (GMM) problem: Let $r_j = \mathbf{1}(w = \xi_j)$ and $m_\ell = \mathbf{1}(z = \zeta_\ell)$. Then we can write

$$E[m_\ell\,(\theta_1 r_1 + \cdots + \theta_J r_J)] = 0 \quad (\ell = 1, \ldots, L), \tag{1.8}$$

which is in the form of a system of L simultaneous equations with instruments m_ℓ in equation ℓ.

The discreteness of endogenous and conditioning variables plays fundamentally different roles in this context. As another example, in a model that includes a subset of z in f so that $f(w, z_1) = u$, if the endogenous variables w are discrete with finite support but $z = (z_1, z_2)$ are continuous, this is equivalent to considering the following semiparametric conditional moment restriction:

$$E\left[\sum_{j=1}^{J} \theta_j(z_1) r_j | z\right] = 0, \tag{1.9}$$

where $w \in \{\xi_1, \ldots, \xi_J\}$, $\theta_j(z_1) = f(\xi_j, z_1)$, and $r_j = \mathbf{1}(w = \xi_j)$.

1.2. Testing for Overidentification and Underidentification

Models (1.1) and (1.2) can be regarded as a restriction on the distribution of w given z,

$$\int f(w) dF(w|z) = 0.$$

Sometimes the focus is in testing the restrictions on $F(w|z)$ rather than in the estimation of $f(w)$ or other average effects. From this point of view, $f(w)$ becomes a nuisance parameter function.

Clearly, in the discrete case, an invariant chi-square test statistic of the overidentifying restrictions (with $L - J + 1$ degrees of freedom) is readily available – but of no use in the continuous case. This is given by

$$\min_{\theta} n\hat{p}'(I \otimes \theta)[(I \otimes \theta')\hat{V}(I \otimes \theta)]^{-1}(I \otimes \theta')\hat{p}, \tag{1.10}$$

where $\hat{p} = \text{vec}(\hat{P})$ denotes a vector of sample frequencies, and $\hat{V}$ is the estimated sampling variance of $\hat{p}$.

Testing for underidentification in the discrete case is also straightforward. One would test the null hypothesis of underidentification, $\text{rank}(P) < J - 1$, against the alternative of identification, $\text{rank}(P) = J - 1$. A test statistic of this kind provides a natural diagnostic of the extent to which structural parameter estimates are well identified.

2. CONTROL FUNCTIONS AND INSTRUMENTAL VARIABLES

2.1. Additive Errors

Newey, Powell, and Vella (1999) considered a nonparametric structural equation together with an explicit reduced form for the endogenous explanatory variables:

$$y = g(x) + u, \tag{2.1}$$

$$x = \pi(z) + v, \tag{2.2}$$

and the assumptions

$$E(u|z, v) = E(u|v), \tag{2.3}$$

$$E(v|z) = 0. \tag{2.4}$$

These assumptions were chosen for convenience. In effect, they imply

$$\begin{aligned} E(y|x, v) &= g(x) + E(u|x, v) \\ &= g(x) + E(u|z, v) = g(x) + E(u|v) = g(x) + h(v). \end{aligned} \tag{2.5}$$

In this way, the problem of nonparametric estimation of $g(x)$ is assimilated to the problem of estimating the regression function $E(y|x, v)$ subject to an additive structure.

Assumptions (2.3) and (2.4) do not imply that $E(u|z) = 0$. The situation is that

$$E(u|z) = E[E(u|z, v)\,|z] = E[E(u|v)\,|z] = E[h(v)|z].$$

A sufficient condition for $E[h(v)|z] = 0$ is that v is independent of z. The mean independence condition does not guarantee that $E[h(v)|z] = 0$ unless $h(v)$ is linear in v.

Alternatively, suppose we begin with the assumptions $E(u|z) = 0$ and $E(v|z) = 0$. Then, in general, (2.3) or (2.5) is not satisfied.

Expression (2.5) makes it clear that the control function assumption can be very useful in applied work, but one should insist that the IV condition $E(u|z) = 0$ also holds. Having a structural equation in which the instruments are correlated with the errors because of a simplifying assumption will, in general, jeopardize the interpretability of the structural parameters.

From the point of view of econometric practice, it is better to regard the control function assumption as a specialization of the IV assumption than to pretend that one is no more or no less general than the other. I regard the control function approach as one in which estimation of the structural function $g(x)$ is *helped* by an explicit semiparametric modeling of the reduced form for x. In practice this will typically require aiming for a reduced form with errors that are statistically independent of instruments.

For example, suppose that for a scalar x, v is heteroskedastic (and hence not independent of z) with $\sigma^2(z) = E(v^2|z)$, but $v^\dagger = \sigma^{-1}(z)v$ is independent of z. In such case, the assumption

$$E(u|z, v^\dagger) = E(u|v^\dagger) \tag{2.6}$$

will be compatible with $E(u|z) = 0$, but (2.3) will imply, in general, correlation between u and z.

The control v can be generalized further, for example to include some kind of Box–Cox transformation of x. The general idea is that the approach works well when there is a reduced-form equation for x or some transformation of x with errors that are independent of z.

The IV assumption in the model with additive errors constrains only the marginal distributions of y and x given z, whereas the control variable assumption places restrictions on their joint distribution. Suppose we have two independent samples on (y, z) and (x, z), respectively, but the joint distribution of y and x is not observed (as in Angrist and Krueger, 1992, or Arellano and Meghir, 1992). It is interesting to compare the IV and control function assumptions in the two-sample estimation context to highlight the different data requirements in the two approaches. In the IV approach, only the marginal distributions of y and x given z are needed for identification, at least conceptually. Thus $y|z$ can be obtained from one sample and $x|z$ from the other. In the control function approach, however, (y, x, z) have to be observed in the same sample to ensure the identification of the nonparametric regression of y on x and v.

2.2. Discrete Choice

Blundell and Powell (1999) (BP) show how the control function approach can be particularly helpful in models with nonadditive errors. They consider a discrete choice model of the form

$$y = \mathbf{1}(x\beta + u > 0), \tag{2.7}$$

$$x = \pi(z) + v, \tag{2.8}$$

$$E(v|z) = 0, \tag{2.9}$$

together with the assumption

$$u|x, v \sim u|v. \tag{2.10}$$

In this way,

$$\Pr(y = 1|x, v) = \Pr(-u \leq x\beta|x, v) = \Pr(-u \leq x\beta|v).$$

Thus

$$E(y|x, v) = F(x\beta, v),$$

where $F(\cdot, v)$ is the conditional cumulative distribution function (CDF) of $-u$ given v.

As in the additive-error case of Newey, Powell, and Vella (NPV), the problem of estimating a structural equation is assimilated to the problem of estimating the regression function $E(y|x, v)$ subject to restrictions. In the case of NPV, it was sufficient to assume that u was mean independent of x given v, and $E(y|x, v)$ had an additive structure. In the discrete choice case, full independence of x given v is required, and $E(y|x, v)$ has a multiple index structure. The difference between the two models exists because (2.7) is not additive or invertible in u.

An interesting feature of the BP method is that the marginal CDF of u evaluated at $x\beta$ can be obtained by averaging $F(x\beta, v)$ over v whose CDF is identified:

$$\Pr(-u \leq x\beta) \equiv G(x\beta) = \int F(x\beta, v)\, dF_v. \tag{2.11}$$

This is useful because the function $G(x\beta)$ is arguably a parameter of interest for policy evaluation in this context.

Turning to a discussion of the assumptions, if (u, v) are independent of z, then

$$u|z, v \sim u|v. \tag{2.12}$$

Moreover, in view of (2.8), $u|x, v \sim u|z, v \sim u|v$. However, (2.9), (2.10), and (2.12) by themselves do not imply independence or even lack of correlation between u and the instruments z. Note that if the BP assumption (2.12) holds, in general u will not be independent of z unless v is independent of z:

$$\begin{aligned} F(u|z) &= \int F(u|z, v) dF_v(v|z) = \int F(u|v) dF_v(v|z) \\ &\neq \int F(u|v) dF_v(v) = F(u). \end{aligned}$$

So, the same remarks as for NPV apply in this context. If $x\beta + u$ represents a latent structural equation, one would expect to select instruments on a priori grounds that suggest some form of independence with the unobservable structural disturbance u.

In particular, in the BP empirical data, the conditional variance of the log other income variable may vary with the education of the spouse. If so, it would be nice to obtain a (possibly nonparametric) estimate of $\mathrm{var}(y_2|z) \equiv \sigma^2(z)$. Then consider $v^\dagger = v/\sigma(z)$ as an alternative control function. In this way we could assess the impact of the correction for heteroskedasticity in the β coefficients and the estimated average structural function.

The conclusion is that the control function approach is best regarded not as a competing identification strategy to IV assumptions but as a complementary modeling strategy for the reduced form of the model. This strategy is specially useful in discrete choice and related models, for which IV assumptions by themselves do not appear to be sufficient to identify parameters of interest.

References

Amemiya, T. (1977), "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model," *Econometrica*, 45, 955–968.

Angrist, J. and A. Krueger (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87, 328–336.

Arellano, M. and C. Meghir (1992), "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets," *Review of Economic Studies*, 59, 537–559.

Blundell, R. W. and J. L. Powell (1999), "Endogeneity in Single Index Models," July, unpublished manuscript, Berkeley.

Gallant, A. R. and G. Tauchen (1989), "Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications," *Econometrica*, 57, 1091–1120.

Hansen, L. P. and K. J. Singleton (1982), "Generalized Instrumental Variables Estimators of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286.

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