**New Developments in Econometrics** **Cemmap, UCL, June 2009**
**Lecture 14, Thursday, June 18th, 10:45-11;45**

# Control Function and Related Methods

These notes review the control function approach to handling endogeneity in models linear in parameters, and draws comparisons with standard methods such as 2SLS and maximum likelihood methods. Certain nonlinear models with endogenous explanatory variables are most easily estimated using the CF method, and the recent focus on average marginal effects suggests some simple, flexible strategies. Recent advances in semiparametric and nonparametric control function method are covered, and an example for how one can apply CF methods to nonlinear panel data models is provided.

## 1. Linear-in-Parameters Models: IV versus Control Functions

Most models that are linear in parameters are estimated using standard instrumental variables methods – either two stage least squares (2SLS) or generalized method of moments (GMM). An alternative, the control function (CF) approach, relies on the same kinds of identification conditions. In the standard case where a endogenous explanatory variables appear linearly, the CF approach leads to the usual 2SLS estimator. But there are differences for models nonlinear in endogenous variables even if they are linear in parameters. And, for models nonlinear in parameters, the CF approach offers some distinct advantages.

To illustrate the CF approach, let $y_1$ denote the response variable, $y_2$ the endogenous explanatory variable (a scalar for simplicity), and $\mathbf{z}$ the $1 \times L$ vector of exogenous variables (which includes unity as its first element). Consider the model

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{1.1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of $\mathbf{z}$ that also includes a constant. The sense in which $\mathbf{z}$ is exogenous is given by the $L$ orthogonality (zero covariance) conditions

$$\mathrm{E}(\mathbf{z}'u_1) = \mathbf{0}. \tag{1.2}$$

Of course, this is the same exogeneity condition that we use for consistency of the 2SLS estimator, and we can consistently estimate $\boldsymbol{\delta}_1$ and $\alpha_1$ by 2SLS under (1.2) and the rank condition, which reduces to *rank* $E(\mathbf{z}'\mathbf{x}_1) = K_1$, where $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$ is a $1 \times K_1$ vector. (We also need to assume $E(\mathbf{z}'\mathbf{z})$ is nonsingular, but this assumption is rarely a concern.)

Just as with 2SLS, the reduced form of $y_2$ – that is, the linear projection of $y_2$ onto the exogenous variables – plays a critical role. Write the reduced form with an error term as

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{1.3}$$

$$\mathrm{E}(\mathbf{z}'v_2) = \mathbf{0} \tag{1.4}$$

where $\boldsymbol{\pi}_2$ is $L \times 1$. Endogeneity of $y_2$ arises if and only if $u_1$ is correlated with $v_2$. Write the linear projection of $u_1$ on $v_2$, in error form, as

$$u_1 = \rho_1 v_2 + e_1, \tag{1.5}$$

where $\rho_1 = \mathrm{E}(v_2u_1)/\mathrm{E}(v_2^2)$ is the population regression coefficient. By definition, $\mathrm{E}(v_2e_1) = 0$, and $\mathrm{E}(\mathbf{z}'e_1) = \mathbf{0}$ because $u_1$ and $v_2$ are both uncorrelated with $\mathbf{z}$.

Plugging (1.5) into equation (1.1) gives

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1, \tag{1.6}$$

where we now view $v_2$ as an explanatory variable in the equation. As just noted, $e_1$, is uncorrelated with $v_2$ and $\mathbf{z}$. Plus, $y_2$ is a linear function of $\mathbf{z}$ and $v_2$, and so $e_1$ is also uncorrelated with $y_2$.

Because $e_1$ is uncorrelated with $\mathbf{z}_1$, $y_2$, and $v_2$, (1.6) suggests a simple procedure for

consistently estimating $\delta_1$ and $\alpha_1$ (as well as $\rho_1$): run the OLS regression of $y_1$ on $\mathbf{z}_1, y_2$, and $v_2$ using a random sample. (Remember, OLS consistently estimates the parameters in any equation where the error term is uncorrelated with the right hand side variables.) The only problem with this suggestion is that we do not observe $v_2$; it is the error in the reduced form equation for $y_2$. Nevertheless, we can write $v_2 = y_2 - \mathbf{z}\boldsymbol{\pi}_2$ and, because we collect data on $y_2$ and $\mathbf{z}$, we can consistently estimate $\boldsymbol{\pi}_2$ by OLS. Therefore, we can replace $v_2$ with $\hat{v}_2$, the OLS residuals from the first-stage regression of $y_2$ on $\mathbf{z}$. Simple substitution gives

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 \hat{v}_2 + error, \tag{1.7}$$

where, for each $i$, $error_i = e_{i1} + \rho_1 \mathbf{z}_i(\hat{\boldsymbol{\pi}}_2 - \boldsymbol{\pi}_2)$, which depends on the sampling error in $\hat{\boldsymbol{\pi}}_2$ unless $\rho_1 = 0$. Standard results on two-step estimation imply the OLS estimators from (1.7) will be consistent for $\delta_1, \alpha_1$, and $\rho_1$.

The OLS estimates from (1.7) are control function estimates. The inclusion of the residuals $\hat{v}_2$ "controls" for the endogeneity of $y_2$ in the original equation (although it does so with sampling error because $\hat{\boldsymbol{\pi}}_2 \neq \boldsymbol{\pi}_2$).

It is a simple exercise in the algebra of least squares to show that the OLS estimates of $\boldsymbol{\delta}_1$ and $\alpha_1$ from (1.7) are *identical* to the 2SLS estimates starting from (1.1) and using $\mathbf{z}$ as the vector of instruments. [Standard errors from (1.7) must adjust for the generated regressor.]

It is trivial to use (1.7) to test $H_0 : \rho_1 = 0$, as the usual $t$ statistic is asymptotically valid under homoskedasticity ($Var(u_1|\mathbf{z}, y_2) = \sigma_1^2$ under $H_0$); or use the heteroskedasticity-robust version (which does *not* account for the first-stage estimation of $\boldsymbol{\pi}_2$).

An estimator that can be different from the CF and 2SLS estimators is the limited information (quasi-) maximum likelihood (LIML) estimator. The LIML estimator is obtained from equations (1.1) and (1.3) under the assumption that $(u_1, v_2)$ is independent of $\mathbf{z}$ with a

mean-zero bivariate normal distribution. In fact, we can work off of (1.3) and (1.6) and use the relationship $f(y_1, y_2|\mathbf{z}) = f(y_1|y_2, \mathbf{z})f(y_2|\mathbf{z})$. If $\eta_1^2 = Var(e_1)$ and $\tau_2^2 = Var(v_2)$, the quasi-log-likelihood for observation $i$ is

$$\begin{aligned} &-\log(\eta_1^2)/2 - [(y_{i1} - \mathbf{z}_{i1}\boldsymbol{\delta}_1 - \alpha_1 y_{i2} - \rho_1(y_{i2} - \mathbf{z}_i\boldsymbol{\pi}_2)]^2/(2\eta_1^2) \\ &\quad -\log(\tau_2^2)/2 - (y_{i2} - \mathbf{z}_i\boldsymbol{\pi}_2)^2/(2\tau_2^2), \end{aligned} \tag{1.8}$$

and all parameters are estimated simultaneously. When (1.1) is overidentified, LIML is generally different from CF (2SLS). And, as the weak instruments notes document, LIML typically has better statistical properties than 2SLS in situations with overidentification. The CF approach can be seen to be a two-step version of LIML, where $\boldsymbol{\pi}_2$ is obtained in a first step and then $\boldsymbol{\delta}_1, \alpha_1$, and $\rho_1$ are estimated in a second step. (The variance parameters can be estimated in the two-step procedure, too.) Fortunately, while LIML is derived under joint normality, it is just as robust as the CF estimator: independence between the errors and $\mathbf{z}$ and normality are not needed.

[Incidentally, full information maximum likelihood (FIML) arises in systems with true simultaneity when interest lies in estimating all structural equations. In these notes, we assume that one equation is of particular interest. This could be because it is the main equation in a truly simultaneous system or because the endogeneity we are worried about is due to omitted variables.]

Now extend the model to include a quadratic:

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1 \tag{1.9}$$

$$\mathrm{E}(u_1|\mathbf{z}) = 0. \tag{1.10}$$

For simplicity, assume that we have a scalar, $z_2$, that is not also in $\mathbf{z}_1$. Then, under (1.10) – which is stronger than (1.2), and is essentially needed to identify nonlinear models – we can

use, say, $z_2^2$ (if $z_2$ is not binary) as an instrument for $y_2^2$ because any function of $z_2$ is uncorrelated with $u_1$. In other words, we can apply the standard IV estimator with explanatory variables $(\mathbf{z}_1, y_2, y_2^2)$ and instruments $(\mathbf{z}_1, z_2, z_2^2)$; note that we have two endogenous explanatory variables, $y_2$ and $y_2^2$.

What would the CF approach entail in this case? To implement the CF approach in (1.9), we obtain the conditional expectation $\mathrm{E}(y_1|\mathbf{z}, y_2)$ – a linear projection argument no longer works because of the nonlinearity – and that requires an assumption about $\mathrm{E}(u_1|\mathbf{z}, y_2)$. A standard assumption is

$$\mathrm{E}(u_1|\mathbf{z}, y_2) = \mathrm{E}(u_1|\mathbf{z}, v_2) = \mathrm{E}(u_1|v_2) = \rho_1 v_2, \tag{1.11}$$

where the first equality follows because $y_2$ and $v_2$ are one-to-one functions of each other (given $\mathbf{z}$) and the second would hold if $(u_1, v_2)$ is independent of $\mathbf{z}$ – a nontrivial restriction on the reduced form error in (1.3), not to mention the structural error $u_1$.. The final assumption is linearity of the conditional expectation $\mathrm{E}(u_1|v_2)$, which is more restrictive than simply defining a linear projection. Under (1.11),

$$\begin{aligned} \mathrm{E}(y_1|\mathbf{z}, y_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1(y_2 - \mathbf{z}\boldsymbol{\pi}_2) \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + \rho_1 v_2. \end{aligned} \tag{1.12}$$

Implementing the CF approach means running the OLS regression $y_1$ on $\mathbf{z}_1, y_2, y_2^2, \hat{v}_2$,where $\hat{v}_2$ still represents the reduced form residuals. The CF estimates are *not* the same as the 2SLS estimates using any choice of instruments for $(y_2, y_2^2)$.

The CF approach, while likely more efficient than a direct IV approach, is less robust. For example, it is easily seen that (1.10) and (1.11) imply that $\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$. A linear conditional expectation for $y_2$ is a substantive restriction on the conditional distribution of $y_2$. Therefore, the CF estimator will be inconsistent in cases where the 2SLS estimator will be consistent. On

the other hand, because the CF estimator solves the endogeneity of $y_2$ and $y_2^2$ by adding the scalar $\hat{v}_2$ to the regression, it will generally be more precise – perhaps much more precise – than the IV estimator. [I do not know of a systematic analysis comparing the two approaches in models such as (1.9).]

The equivalence between CF approaches and IV methods is broken even in the simple model (1.1) if we allow $y_2$ to have discreteness in its distribution and we use a distributional assumption to exploit that discreteness. For example, suppose $y_2$ is a binary response. The standard CF approach involves estimating

$$E(y_1|\mathbf{z},y_2) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + E(u_1|\mathbf{z},y_2), \tag{1.13}$$

and so we must be able to estimate $E(u_1|\mathbf{z},y_2)$. If $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + e_2 \geq 0]$, $(u_1, e_2)$ is independent of $\mathbf{z}$, $E(u_1|e_2) = \rho_1 e_2$, and $e_2$ ~Normal(0, 1), then

$$\begin{aligned} E(u_1|\mathbf{z},y_2) &= E[E(u_1|\mathbf{z},e_2)|\mathbf{z},y_2] = \rho_1 E(v_2|\mathbf{z},y_2) \\ &= \rho_1[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)], \end{aligned} \tag{1.14}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio (IMR). A simple two-step estimator is to obtain the probit estimate $\hat{\boldsymbol{\delta}}_2$ and then to add the "generalized residual," $\widehat{gr}_{i2} \equiv y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1-y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as a regressor:

$$y_{i1} \text{ on } \mathbf{z}_{i1},\ y_{i2},\ \widehat{gr}_{i2},\ i = 1,\ldots,N. \tag{1.15}$$

The estimators from this regression are consistent and $\sqrt{N}$-asymptotically normal provided $D(y_2|\mathbf{z})$ follows a probit, $E(u_1|v_2)$ is linear, and $E(u_1|\mathbf{z},v_2) = E(u_1|v_2)$. (Standard errors need to be adjusted for the two-step estimation, except when $\rho_1 = 0$. A simple $t$ test on $\widehat{gr}_{i2}$ is valid as a test of $H_0 : \rho_1 = 0$.)

Of course, if we just apply 2SLS directly to $y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1$, we make no distinction

among discrete, continuous, or some mixture for $y_2$. 2SLS is consistent if $L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ actually depends on $\mathbf{z}_2$ and (1.2) holds. So, while estimating (1.1) using CF methods when $y_2$ is binary is somewhat popular (Stata's "treatreg" even has the option of full MLE, where $(u_1, e_2)$ is bivariate normal), one should remember that it is less robust than standard IV approaches. In principal, it is much less robust, but whether estimates obtained from (1.15) differ substantially from 2SLS estimates is an empirical issue.

Often researchers look to exploit the binary nature of the endogenous explanatory variable, and there may even be some confusion about the properties of 2SLS in such contexts. Again, it is important to understand that 2SLS is consistent, $\sqrt{N}$-asymptotically normal, and inference is standard. But it could be asymptotically inefficient. Therefore, a natural question is: How might one use the binary nature of $y_2$ in IV estimation [as opposed to the CF approach in (1.15)]? We need to assume $E(u_1|\mathbf{z}) = 0$ to exploit nonlinear functions $\mathbf{z}$ as IVs. Nominally, the same probit model for $D(y_2|\mathbf{z})$ that is used in the CF approach. Then, after estimating the probit model, obtain the fitted probabilities, $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$. These fitted probabilities are then used as IVs for $y_{i2}$ in estimating (1.1). This method has several attractive features: it is fully robust to misspecification of the probit model, provided one uses $\Phi(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as an IV for $y_{i2}$, not as a regressor in place of $y_{i2}$; the standard errors need not be adjusted for the first-stage probit (asymptotically); and it is the efficient IV estimator if $P(y_2 = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\delta}_2)$ and $Var(u_1|\mathbf{z}) = \sigma_1^2$. Probably it is less efficient than the CF estimator if the additional assumptions needed for CF consistency hold; a careful study could shed light on the tradeoffs. See Wooldridge (2002, Chapter) 18 for further discussion.

We can briefly summarize the main points of this section. In the model (1.1), CF methods based on $E(y_1|\mathbf{z},y_2)$ impose additional assumptions compared with standard IV methods. When

$y_2$ has special features (such as being binary, or even a corner solution), models for $E(y_2|\mathbf{z})$ can be used to generate instruments (not regressors) for $y_2$. The resulting IV estimates are robust to misspecification of the model for $E(y_2|\mathbf{z})$ and the first-step estimation can be ignored asymptotically.

# 2. Correlated Random Coefficient Models

Control function methods can be used for random coefficient models – that is, models where unobserved heterogeneity interacts with endogenous explanatory variables. In some cases, CF methods are indispensable; in other cases, standard IV methods are more robust. To illustrate, we modify equation (1.1) as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + u_1, \tag{2.1}$$

where $\mathbf{z}_1$ is $1 \times L_1$, $y_2$ is the endogenous explanatory variable, and $a_1$, the "coefficient" on $y_2$ – an unobserved random variable. [It is now convenient to set apart the intercept.] We could replace $\boldsymbol{\delta}_1$ with a random vector, say $\mathbf{d}_1$, and this would not affect our analysis of the IV estimator (but, as we will see, does change the control function estimator). Following Heckman and Vytlacil (1998), we refer to (2.1) as a **correlated random coefficient** (**CRC**) **model**.

It is convenient to write $a_1 = \alpha_1 + v_1$ where $\alpha_1 = \mathrm{E}(a_1)$ is the object of interest. (In the context of treatment effect estimation, $\alpha_1$ is the average treatment effect.) We can rewrite the equation as

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + v_1 y_2 + u_1 \equiv \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + e_1, \tag{2.2}$$

where $e_1 = v_1 y_2 + u_1$. Equation (2.2) shows explicitly a constant coefficient on $y_2$ (which we hope to estimate) but also an interaction between the observed heterogeneity, $v_1$, and $y_2$. Remember, (2.2) is a population model. For a random draw, we would write $y_{i1} = \eta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + v_{i1} y_{i2} + u_{i1}$, which makes it clear that $\boldsymbol{\delta}_1$ and $\alpha_1$ are parameters to estimate and $v_{i1}$ is specific to observation $i$.

As discussed in Wooldridge (1997, 2003), the potential problem with applying instrumental

variables (2SLS) to (2.2) is that the error term $v_1y_2 + u_1$ is not necessarily uncorrelated with the instruments **z**, even if we make the assumptions

$$\mathrm{E}(u_1|\mathbf{z}) = \mathrm{E}(v_1|\mathbf{z}) = 0, \tag{2.3}$$

which we maintain from here on. Generally, the term $v_1y_2$ can cause problems for IV estimation, but it is important to be clear about the nature of the problem. If we are allowing $y_2$ to be correlated with $u_1$ then we also want to allow $y_2$ and $v_1$ to be correlated. In other words, $\mathrm{E}(v_1y_2) = \mathrm{Cov}(v_1,y_2) \equiv \tau_1 \neq 0$. But a nonzero unconditional covariance is *not* a problem with applying IV to (2.2): it simply implies that the composite error term, $e_1$, has (unconditional) mean $\tau_1$ rather than a zero. As we know, a nonzero mean for $e_1$ means that the orginal intercept, $\eta_1$, would be inconsistenly estimated, but this is rarely a concern.

Therefore, we can allow $\mathrm{Cov}(v_1,y_2)$, the unconditional covariance, to be unrestricted. But the usual IV estimator is generally inconsistent if $\mathrm{E}(v_1y_2|\mathbf{z})$ depends on **z**. Note that, because $\mathrm{E}(v_1|\mathbf{z}) = 0$, $\mathrm{E}(v_1y_2|\mathbf{z}) = \mathrm{Cov}(v_1,y_2|\mathbf{z})$. Therefore, as shown in Wooldridge (2003), a sufficient condition for the IV estimator applied to (2.2) to be consistent for $\boldsymbol{\delta}_1$ and $\alpha_1$ is

$$\mathrm{Cov}(v_1,y_2|\mathbf{z}) = \mathrm{Cov}(v_1,y_2). \tag{2.4}$$

The 2SLS intercept estimator is consistent for $\eta_1 + \tau_1$. Condition (2.4) means that the conditional covariance between $v_1$ and $y_2$ is not a function of **z**, but the unconditional covariance is unrestricted.

Because $v_1$ is unobserved, we cannot generally verify (2.4). But it is easy to find situations where it holds. For example, if we write

$$y_2 = m_2(\mathbf{z}) + v_2 \tag{2.5}$$

and assume $(v_1, v_2)$ is independent of **z** (with zero mean), then (2.4) is easily seen to hold

because $\text{Cov}(v_1, y_2|\mathbf{z}) = \text{Cov}(v_1, v_2|\mathbf{z})$, and the latter cannot be a function of $\mathbf{z}$ under independence. Of course, assuming $v_2$ in (2.5) is independent of $\mathbf{z}$ is a strong assumption even if we do not need to specify the mean function, $m_2(\mathbf{z})$. It is much stronger than just writing down a linear projection of $y_2$ on $\mathbf{z}$ (which is no real assumption at all). As we will see in various models in Part IV, the representation (2.5) with $v_2$ independent of $\mathbf{z}$ is not suitable for discrete $y_2$, and generally (2.4) is not a good assumption when $y_2$ has discrete characteristics. Further, as discussed in Card (2001), (2.4) can be violated even if $y_2$ is (roughly) continuous. Wooldridge (2005) makes some headway in relaxing (2.44) by allowing for parametric heteroskedasticity in $u_1$ and $v_2$.

A useful extension of (1.1) is to allow observed exogenous variables to interact with $y_2$. The most convenient formulation is

$$y_1 = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + (\mathbf{z}_1 - \boldsymbol{\psi}_1)y_2\boldsymbol{\gamma}_1 + v_1 y_2 + u_1 \tag{2.6}$$

where $\boldsymbol{\psi}_1 \equiv \text{E}(\mathbf{z}_1)$ is the $1 \times L_1$ vector of population means of the exogenous variables and $\boldsymbol{\gamma}_1$ is an $L_1 \times 1$ parameter vector. As we saw in Chapter 4, subtracting the mean from $\mathbf{z}_1$ before forming the interaction with $y_2$ ensures that $\alpha_1$ is the average partial effect.

Estimation of (2.6) is simple if we maintain (2.4) [along with (2.3) and the appropriate rank condition]. Typically, we would replace the unknown $\boldsymbol{\psi}_1$ with the sample averages, $\bar{\mathbf{z}}_1$, and then estimate

$$y_{i1} = \theta_1 + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}\boldsymbol{\gamma}_1 + error_i \tag{2.7}$$

by instrumental variables, ignoring the estimation error in the population mean. The only issue is choice of instruments, which is complicated by the interaction term. One possibility is to use interactions between $\mathbf{z}_{i1}$ and all elements of $\mathbf{z}_i$ (including $\mathbf{z}_{i1}$). This results in many

overidentifying restrictions, even if we just have one instrument $z_{i2}$ for $y_{i2}$. Alternatively, we could obtain fitted values from a first stage linear regression $y_{i2}$ on $\mathbf{z}_i$, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$, and then use IVs $[1, \mathbf{z}_i, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}]$, which results in as many overidentifying restrictions as for the model without the interaction. Importantly, the use of $(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}$ as IVs for $(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2}$ is asymptotically the same as using instruments $(\mathbf{z}_{i1} - \boldsymbol{\psi}_1) \cdot (\mathbf{z}_i\boldsymbol{\pi}_2)$, where $L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2$ is the linear projection. In other words, consistency of this IV procedure does not in any way restrict the nature of the distribution of $y_2$ given $\mathbf{z}$. Plus, although we have generated instruments, the assumptions sufficient for ignoring estimation of the instruments hold, and so inference is standard (perhaps made robust to heteroskedasticity, as usual).

We can just identify the parameters in (2.6) by using a further restricted set of instruments, $[1, \mathbf{z}_{i1}, \hat{y}_{i2}, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)\hat{y}_{i2}]$. If so, it is important to use these as instruments and not as regressors. If we add the assumption. The latter procedure essentially requires a new assumption:

$$\mathrm{E}(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\pi}_2 \tag{2.8}$$

(where $\mathbf{z}$ includes a constant). Under (2.3), (2.4), and (2.8), it is easy to show

$$\mathrm{E}(y_1|\mathbf{z}) = (\eta_1 + \tau_1) + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\pi}_2) + (\mathbf{z}_1 - \boldsymbol{\psi}_1) \cdot (\mathbf{z}\boldsymbol{\pi}_2)\boldsymbol{\gamma}_1, \tag{2.9}$$

which is the basis for the Heckman and Vytlacil (1998) plug-in estimator. The usual IV approach applied to (2.7) simply relaxes (2.8) and does not require adjustments to the standard errors (because it uses generated instruments, not generated regressors).

We can also use a control function approach if we assume

$$\mathrm{E}(u_1|\mathbf{z}, v_2) = \rho_1 v_2, \mathrm{E}(v_1|\mathbf{z}, v_2) = \xi_1 v_2. \tag{2.10}$$

Then

$$\mathrm{E}(y_1|\mathbf{z}, y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \xi_1 v_2 y_2 + \rho_1 v_2, \tag{2.11}$$

and this equation is estimable once we estimate $\boldsymbol{\pi}_2$. Garen's (1984) control function procedure is to first regress $y_2$ on $\mathbf{z}$ and obtain the reduced form residuals, $\hat{v}_2$, and then to run the OLS regression $y_1$ on $1, \mathbf{z}_1, y_2, \hat{v}_2 y_2, \hat{v}_2$. Under the maintained assumptions, Garen's method consistently estimates $\boldsymbol{\delta}_1$ and $\alpha_1$. Because the second step uses generated regressors, the standard errors should be adjusted for the estimation of $\boldsymbol{\pi}_2$ in the first stage. Nevertheless, a test that $y_2$ is exogenous is easily obtained from the usual $F$ test of $\mathrm{H}_0 : \xi_1 = 0, \rho_1 = 0$ (or a heteroskedasticity-robust version). Under the null, no adjustment is needed for the generated standard errors.

Garen's assumptions are more restrictive than those needed for the standard IV estimator to be consistent. For one, it would be a fluke if (2.10) held without the conditional covariance $\mathrm{Cov}(v_1, y_2|\mathbf{z})$ being independent of $\mathbf{z}$. Plus, like HV (1998), Garen relies on a linear model for $\mathrm{E}(y_2|\mathbf{z})$. Further, Garen adds the assumptions that $\mathrm{E}(u_1|v_2)$ and $\mathrm{E}(v_1|v_2)$ are linear functions, something not needed by the IV approach.

Of course, one can make Garen's approach less parametric by replacing the linear functions in (2.10) with unknown functions. But independence of $(u_1, v_1, v_2)$ and $\mathbf{z}$ – or something very close to independence – is needed. And this assumption is not needed for the usual IV estimator,

If the assumptions needed for Garen's CF estimator to be consistent hold, it is likely more efficient than the IV estimator, although a comparison of the correct asymptotic variances is complicated. Again, there is a tradeoff between efficiency and robustness.

In the case of binary $y_2$, we have what is often called the "switching regression" model. Now, the right hand side of equation (2.11) represents $E(y_1|\mathbf{z}, v_2)$ where $y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0]$. If we assume (2.10) and that $v_2|\mathbf{z}$ is $\mathrm{Normal}(0, 1)$, then

$$E(y_1|\mathbf{z},y_2) = \eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)y_2, \tag{2.12}$$

where

$$h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) = y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2) \tag{2.13}$$

is the generalized residual function. The two-step estimation method is the one due to Heckman (1976).

There are two ways to embellish the model. The first is common: interact $(\mathbf{z}_1 - \boldsymbol{\mu}_1)$ with $y_2$ to allow different slopes for the "treated" and non-treated groups (keeping $\alpha_1$ as the average treatment effect). With this extension, the CF regression is

$$y_{i1} \text{ on } 1,\ \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2},\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}, \tag{2.14}$$

which is the same as the more familiar regression,

$$y_{i1} \text{ on } 1,\ \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2},\ (1-y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$$

Of course, this latter regression is also identical to running two separate regressions, including the IMRs for $y_2 = 0$ and $y_2 = 1$. The estimate of $\alpha_1$ is the difference in the two intercepts. The combined regression is convenient for putting proper emphasize on $\alpha_1$, which is the population average effect of the binary variable, $y_2$. The combined regression also makes it straightforward to program a bootstrap procedure for inference.

An extension that is not so common – in fact, it seems not to appear in the literature – comes from allowing $\mathbf{z}_1$ to also interact with heterogeneity, as in

$$y_1 = \mathbf{z}_1\mathbf{d}_1 + a_1 y_2 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\mathbf{g}_1 + u_1. \tag{2.15}$$

Now all coefficients are heterogeneous. If we assume that $E(a_1|v_2)$, $E(\mathbf{d}_1|v_2)$, and $E(\mathbf{g}_1|v_2)$ are linear in $v_2$, then

$$\begin{aligned} E(y_1|\mathbf{z},y_2) &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\xi}_1 + \rho_1 E(v_2|\mathbf{z},y_2) + \xi_1 E(v_2|\mathbf{z},y_2)y_2 \\ &\quad + \mathbf{z}_1 E(v_2|\mathbf{z},y_2)\boldsymbol{\psi}_1 + y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)E(v_2|\mathbf{z},y_2)\boldsymbol{\omega}_1 \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) + \xi_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)y_2 \\ &\quad + h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)\mathbf{z}_1\boldsymbol{\psi}_1 + h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2)y_2(\mathbf{z}_1 - \boldsymbol{\mu}_1)\boldsymbol{\omega}_1. \end{aligned} \tag{2.16}$$

After the first-stage probit, the second-stage regression can be obtained as

$$\begin{aligned} y_{i1} \text{ on } 1,\ \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \alpha_1 y_{i2} + (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)y_{i2},\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2),\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}, \\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1},\ h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1). \end{aligned} \tag{2.17}$$

across all observations *i*. Bootstrapping can be used to obtain valid standard errors because the first-stage estimation is just probit and the second stage is just linear regression.

If not for the term $v_1y_2$ in (2.6), we could, in a much more robust manner, apply IV directly to (2.7) (and the standard errors are easier to obtain, too). The IVs would be, say, $[1, \mathbf{z}_{i1}, \hat{\Phi}_{i2}, (\mathbf{z}_{i1} - \bar{\mathbf{z}}_1) \cdot \hat{\Phi}_{i2}]$, and the same procedure consistently estimates the average effects whether or not there are random coefficients on $\mathbf{z}_{i1}$.

Interestingly, the addition of the terms $h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)\mathbf{z}_{i1}$ and $h_2(y_{i2},\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)y_{i2}(\mathbf{z}_{i1} - \bar{\mathbf{z}}_1)$ has similarities to methods that allow $E(u_1|v_2)$ and $E(v_1|v_2)$ to be more flexible. For example, as shown in Lee (1982) and Heckman and MaCurdy (1986), if $E(u_1|v_2) = \rho_1 v_2 + \kappa_1(v_2^2 - 1)$, then

$$E(u_1|\mathbf{z},y_2) = \rho_1 h_2(y_2,\mathbf{z}\boldsymbol{\delta}_2) + \kappa_1[(1 - y_2)\mathbf{z}\boldsymbol{\delta}_2\lambda(-\mathbf{z}\boldsymbol{\delta}_2) - y_2\mathbf{z}\boldsymbol{\delta}_2\lambda(\mathbf{z}\boldsymbol{\delta}_2)]$$

and a similar expression holds for $E(v_1|v_2)$, where $v_1$ is the unobservable interacting with $y_2$ in (2.2). The generalized residual is now more complicated by easy to estimate. In addition to $y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ and $(1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$, we would also include $(1 - y_{i2})\mathbf{z}_i\hat{\boldsymbol{\delta}}_2\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ and $y_{i2}\mathbf{z}_i\hat{\boldsymbol{\delta}}_2\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ as regressors (in the model with constant slopes on the exogenous variables). In the general model (2.15), these two terms would also be interacted with $\mathbf{z}_{i1}$.

Newey (1988), in the standard switching regression framework, proposed a flexible two-step procedure that estimates $\boldsymbol{\delta}_2$ semiparametrically in the first stage – see Powell (1994) for a survey of such methods – and then uses series in $\mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ in place of the usual IMR terms. He obtains valid standard errors and, in most cases, bootstrapping is valid, too. In the sample selection case – so they estimate an equation for $y_2 = 1$ – Powell, Newey, and Walker (1990) implement both kernel and series estimators using the Mroz (1987) data. Because they do not reject the standard probit model for the selection equation, they estimate $\boldsymbol{\delta}_2$ by probit, and then include the terms $\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ and $\mathbf{z}_i\hat{\boldsymbol{\delta}}_2\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ (which are obtained by cross validation), and so their procedure in this case is identical to the proposal by Heckman and MaCurdy (1986).

Of course, the index restriction in the selection equation, namely, $P(y_2 = 1|\mathbf{z}) = P(y_2 = 1|\mathbf{z}\boldsymbol{\delta}_2)$ may be too restrictive. In principle, on may estimate the selection probability using a fully nonparametric estimator, and then use either a partial linear model approach (Robinson (1987)) based on kernel or series estimation. Powell (1994) describes this possibility. Alternatively, a combination of, say, a heteroskedastic probit model for selection, and a quadratic expectation might provide useful evidence for sensitity of the estimated $\alpha_1$. More precisly, if we specify $v_2|\mathbf{z}\sim Normal(0, \exp(2\mathbf{z}_2\boldsymbol{\eta}_2))$, where $\mathbf{z}_2$ is a subset of $\mathbf{z}$ that excludes a constant, then we can define a standard normal variable that is independnet of $\mathbf{z}$ as $e_2 = \exp(-\mathbf{z}_2\boldsymbol{\eta}_2)v_2$. Then, for example, we can assume $E(u_1|e_2) = \rho_1 v_2 + \kappa_1(e_2^2 - 1)$. With $\kappa_1 = 0$, we just replace the generalized residual $h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)$ with

$$h_2(y_{i2}, \mathbf{z}_i\hat{\boldsymbol{\delta}}_2, \mathbf{z}_{i2}\hat{\boldsymbol{\eta}}_2) = y_2\lambda(\exp(-\mathbf{z}_{i2}\hat{\boldsymbol{\eta}}_2)\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_2)\lambda(\exp(-\mathbf{z}_{i2}\hat{\boldsymbol{\eta}}_2)\mathbf{z}_i\hat{\boldsymbol{\delta}}_2),$$

where $\hat{\boldsymbol{\delta}}_2$ and $\hat{\boldsymbol{\eta}}_2$ are obtained from "heteroskedastic probit" estimation. An open question is whether these kinds of computationally simple yet flexible approaches do notably less well

than full semiparametric or nonparametric procedures.

Finally, we should not forget that maximum likelihood estimation is an alternative to two-step estimation. If $D(y_2|\mathbf{z})$ is specified as a probit and all unobservables are assumed to be jointly normal and independent of $\mathbf{z}$, $D(y_1|y_2,\mathbf{z})$ can be obtained and all parameters can be estimated jointly. A joint MLE is computationally much more demanding for the embellishments just mentioned, such as replacing $\boldsymbol{\delta}_1$ and $\boldsymbol{\xi}_1$ with random vectors that can be correlated with $y_2$.

# 3. Some Common Nonlinear Models and Limitations of the CF Approach

Like standard IV methods, control function approaches are more difficult to apply to nonlinear models, even relatively simple ones. Methods are available when the endogenous explanatory variables are continuous, but few if any results apply to cases with discrete $y_2$. Therefore, maximum likelihood approaches continue to be popular for nonlinear models.

## 3.1. Binary and Fractional Responses

The probit model provides a good illustration of the general approach. With a single endogenous explanatory variable, the simplest specification is

$$y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \geq 0], \tag{3.1}$$

where $u_1|z \sim$Normal$(0,1)$. But the analysis goes through if we replace $(z_1, y_2)$ with any known function $g_1(z_1, y_2)$, provided we have sufficient identifying assumptions. An example is $y_1 = [\mathbf{z}_1\boldsymbol{\delta}_1 + y_2\mathbf{z}_1\boldsymbol{\alpha}_1 + \gamma_1 y_2^2 + u_1 > 0]$. The nonlinearity in $y_2$ is not itself a problem (unless we inappropriately try to mimic 2SLS – more on this later).

The Smith-Blundell (1986) and Rivers-Vuong (1988) approach is to make a homoskedastic-normal assumption on the reduced form for $y_2$,

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2, \;\; v_2|\mathbf{z} \sim \text{Normal}(0, \tau_2^2). \tag{3.2}$$

A key point is that the RV approach essentially requires

$$(u_1, v_2) \text{ independent of } \mathbf{z}; \tag{3.3}$$

as we will see in the next section, semiparametric and nonparametric CF methods also rely on (3.3), or at least something close to it..

If we assume

$$(u_1, v_2) \sim \text{Bivariate Normal} \tag{3.4}$$

with $\rho_1 = Corr(u_1, v_2)$, then we can proceed with MLE based on $f(y_1, y_2|\mathbf{z})$. A simpler two-step approach, which is convenient for testing $H_0 : \rho_1 = 0$ ($y_2$ is exogenous), is also available, and it works if we replace the normality assumption in (3.2), the independence assumption in (3.3), and joint normality in (3.4) with

$$D(u_1|v_2, \mathbf{z}) = \text{Normal}(\theta_1 v_2, 1 - \rho_1^2), \tag{3.5}$$

where $\theta_1 = \rho_1/\tau_2$ is the regression coefficient. That we can relax the assumptions to some degree using a two-step CF approach has implications for less parametric approaches. Certainly we can relax the homoskedasticity and linear expectation in (3.3) without much additional work, as discussed in Wooldridge (2005).

Under the weaker assumption (3.5) we can write

$$P(y_1 = 1|\mathbf{z}, y_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_{\rho 1} + \alpha_{\rho 1} y_2 + \theta_{\rho 1} v_2) \tag{3.6}$$

where each coefficient is multiplied by $(1 - \rho_1^2)^{-1/2}$.

The RV two-step approach is

(1) OLS of $y_2$ on $\mathbf{z}$, to obtain the residuals, $\hat{v}_2$.

(2) Probit of $y_1$ on $\mathbf{z}_1, y_2, \hat{v}_2$ to estimate the scaled coefficients.

The original coefficients, which appear in the partial effects, are easily obtained from the set of two-step estimates:

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{\rho 1}/(1 + \hat{\theta}_{\rho 1}^2 \hat{\tau}_2^2)^{1/2}, \tag{3.7}$$

where $\hat{\theta}_{\rho 1}$ is the coeffcient on $\hat{v}_2$ and $\hat{\tau}_2^2$ is the usual error variance estimator from the first step

OLS, and $\hat{\boldsymbol{\beta}}_{\rho 1}$ includes $\hat{\boldsymbol{\delta}}_{\rho 1}$ and $\hat{\alpha}_{\rho 1}$. Standard errors can be obtained from the delta method of bootstrapping. Of course, they are computed directly from MLE. Partial effects are based on $\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1)$ where $\mathbf{x}_1 = (\mathbf{z}_1, y_2)$. It should be clear that nothing changes for estimation if $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, y_2)$; of course, we would change how partial effects are computed to account for the specific function $\mathbf{g}_1(\cdot, \cdot)$.

Testing the null hypothesis that $y_2$ is exogenous is simple using the two-step control function approache. Asymptotically, a simple $t$ test on $\hat{v}_2$ is valid to test $\mathrm{H}_0 : \rho_1 = 0$.

Under (3.3), we can also apply maximum likelihood by combining (3.2) and (3.6), recognizing that $v_2 = y_2 - \mathbf{z}\boldsymbol{\pi}_2$ and estimating all parameters jointly. For details, see Wooldridge (2002, Section 15.7.2).

A different way to obtain partial effects is to use the average structural function approach, which leads to estimation of $\mathrm{E}_{v_2}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_{\rho 1} + \theta_{\rho 1} v_2)]$. Whether or not $v_2$ is normally distributed, a consistent, $\sqrt{N}$-asymptotically normal estimator of the average structural function (evaluated at a given vector $\mathbf{x}_1$) is

$$\widehat{\mathrm{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{\rho 1} + \hat{\theta}_{\rho 1}\hat{v}_{i2}); \tag{3.8}$$

that is, we average out the reduced form residuals, $\hat{v}_{i2}$. This formulation is also useful for more complicated models.

Given that the probit structural model is essentially arbitrary, one might be so bold as to specify models for $P(y_1 = 1|\mathbf{z}_1, y_2, v_2)$ directly. For example, we can add polynomials in $v_2$ or even interact $v_2$ with elements of $\mathbf{x}_1$ side a probit or logit function. We return to such possibilities in the next section.

The two-step CF approach easily extends to fractional responses. Now, we start with an omitted variables formulation in the conditional mean:

$$E(y_1|\mathbf{z},y_2,q_1) = E(y_1|\mathbf{z}_1,y_2,q_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + q_1), \tag{3.9}$$

where $\mathbf{x}_1$ is a function of $(\mathbf{z}_1,y_2)$ and $q_1$ contains unobservables. As usual, we need some exclusion restrictions, embodied by omitting $\mathbf{z}_2$ from $\mathbf{x}_1$. The specification in equation (3.9) allows for responses at the corners, zero and one, and $y_1$ may take on any values in between. Under the assumption that

$$D(q_1|v_2,\mathbf{z}) \sim \text{Normal}(\theta_1 v_2, \eta_1^2) \tag{3.10}$$

Given (3.9) and (3.10), it can be shown, using the mixing property of the normal distribution, that

$$E(y_1|\mathbf{z},y_2,v_2) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{\eta 1} + \theta_{\eta 1} v_2), \tag{3.11}$$

where the index "$\eta$" denotes coefficients multiplied by $(1 + \eta_1^2)^{-1/2}$. Because the Bernoulli log likelihood is in the linear exponential family, maximizing it consistently estimates the parameters of a correctly specified mean; naturally, the same is true for two-step estimation. That is, the *same* two-step method can be used in the binary and fractional cases. Of course, the variance associated with the Bernoulli distribution is generally incorrect. In addition to correcting for the first-stage estimates, a robust sandwich estimator should be computed to account for the fact that $D(y_1|\mathbf{z},y_2)$ is not Bernoulli. The best way to compute partial effects is to use (3.8), with the slight notational change that the implicit scaling in the coefficients is different. By using (3.8), we can directly use the scaled coefficients estimated in the second stage – a feature common across CF methods for nonlinear models. The bootstrap that reestimates the first and second stages for each iteration is an easy way to obtain standard

errors. Of course, having estimates of the parameters up to a common scale allows us to determine signs of the partial effects in (3.9) as well as relative partial effects on the continuous explanatory variables.

Wooldridge (2005) describes some simple ways to make the analysis starting from (3.9) more flexible, including allowing $Var(q_1|v_2)$ to be heteroskedastic. We can also use strictly monotonic transformations of $y_2$ in the reduced form, say $h_2(y_2)$, regardless of how $y_2$ appears in the structural model: the key is that $y_2$ can be written as a function of $(\mathbf{z}, v_2)$. The extension to multivariate $\mathbf{y}_2$ is straightforward with sufficient instruments provide the elements of $\mathbf{y}_2$, or strictly monotonic functions of them, have reduced forms with additive errors that are effectively indendent of $\mathbf{z}$. (This assumption rules out applications to $y_2$ that are discrete (binary, multinomial, or count)or have a discrete component (corner solution).

The control function approach has some decided advantages over another two-step approach – one that appears to mimic the 2SLS estimation of the linear model. Rather than conditioning on $v_2$ along with $\mathbf{z}$ (and therefore $y_2$) to obtain $\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{P}(y_1 = 1|\mathbf{z}, y_2, v_2)$, we can obtain $\mathrm{P}(y_1 = 1|\mathbf{z})$. To find the latter probability, we plug in the reduced form for $y_2$ to get $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2) + \alpha_1 v_2 + u_1 > 0]$. Because $\alpha_1 v_2 + u_1$ is independent of $\mathbf{z}$ and $(u_1, v_2)$ has a bivariate normal distribution, $\mathrm{P}(y_1 = 1|\mathbf{z}) = \Phi\{[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1(\mathbf{z}\boldsymbol{\delta}_2)]/\omega_1\}$ where $\omega_1^2 \equiv \mathrm{Var}(\alpha_1 v_2 + u_1) = \alpha_1^2\tau_2^2 + 1 + 2\alpha_1\mathrm{Cov}(v_2, u_1)$. (A two-step procedure now proceeds by using the same first-step OLS regression – in this case, to get the fitted values, $\hat{y}_{i2} = \mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ – now followed by a probit of $y_{i1}$ on $\mathbf{z}_{i1}, \hat{y}_{i2}$. It is easily seen that this method estimates the coefficients up to the common scale factor $1/\omega_1$, which can be any positive value (unlike in the CF case, where we know the scale factor is greater than unity).

One danger with plugging in fitted values for $y_2$ is that one might be tempted to plug $\hat{y}_2$ into nonlinear functions, say $y_2^2$ or $y_2\mathbf{z}_1$. This does not result in consistent estimation of the scaled parameters or the partial effects. If we believe $y_2$ has a linear RF with additive normal error independent of $\mathbf{z}$, the addition of $\hat{v}_2$ solves the endogeneity problem regardless of how $y_2$ appears. Plugging in fitted values for $y_2$ only works in the case where the model is linear in $y_2$. Plus, the CF approach makes it much easier to test the null that for endogeneity of $y_2$ as well as compute APEs.

In standard index models such as (3.9), or, if you prefer, (3.1), the use of control functions to estimate the (scaled) parameters and the APEs produces no surprises. However, one must take care when, say, we allow for random slopes in nonlinear models. For example, suppose we propose a random coefficient model

$$E(y_1|\mathbf{z},y_2,\mathbf{c}_1) = E(y_1|\mathbf{z}_1,y_2,\mathbf{c}_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + a_1y_2 + q_1), \tag{3.12}$$

where $a_1$ is random with mean $\alpha_1$ and $q_1$ again has mean of zero. If we want the partial effect of $y_2$, evaluated at the mean of heterogeneity, we have

$$\alpha_1\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2), \tag{3.13}$$

where $\phi(\cdot)$ is the standard normal pdf, and this equation is obtained by differentiating (3.12) with respect to $y_2$ and then plugging in $a_1 = \alpha_1$ and $q_1 = 0$. Suppose we write $a_1 = \alpha_1 + d_1$ and assume that $(d_1,q_1)$ is bivariate normal with mean zero. Then, for given $(\mathbf{z}_1,y_2)$, the average structural function can be shown to be

$$E_{(d_1,q_1)}[\Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2 + d_1y_2 + q_1)] = \Phi[(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1y_2)/(1 + \sigma_q^2 + 2\sigma_{dq}y_2 + \sigma_d^2y_2^2)^{1/2}], \tag{3.14}$$

where $\sigma_q^2 = Var(q_1)$, $\sigma_d^2 = Var(d_1)$, and $\sigma_{dq} = Cov(d_1,q_1)$. The average partial effect with respect to, say, $y_2$, is the derivative of this function with respect to $y_2$. While this partial effect

depends on $\alpha_1$, it is messier than (3.13) and need not even have the same sign as $\alpha_1$. Wooldridge (2005) discusses related issues in the context of probit models with exogenous variables and heteroskedasticity. In one example, he shows that, depending on whether heteroskedasticity in the probit is due to heteroskedasticity in $Var(u_1|\mathbf{x}_1)$, where $u_1$ is the latent error, or due to random slopes, the APEs are completely different in general. The same is true here: the APE when the coefficient on $y_2$ is random is generally very different from the APE obtained if we maintain $a_1 = \alpha_1$ but allow $Var(q_1|v_2)$ to be heteroskedastic. In the latter case, the APE is a positive multiple of $\alpha_1$.

Incidentally, we can estimate the APE in (3.14) fairly generally. A parametric approach is to assume joint normality of $(d_1, q_1, v_2)$ (and independence with $\mathbf{z}$). Then, with a normalization restriction, it can be shown that

$$E(y_1|\mathbf{z}, v_2) = \Phi[(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2)/(1 + \eta_1 y_2 + \lambda_1 y_2^2)^{1/2}], \tag{3.15}$$

which can be estimated by inserting $\hat{v}_2$ for $v_2$ and using nonlinear least squares or Bernoulli QMLE. (The latter is often called "heteroskedastic probit" when $y_1$ is binary.) This procedure can be viewed as an extension to Garen's method for linear models with correlated random coefficients.

Estimation, inference, and interpretation would be especially straightforward (the latter possibly using the bootstrap) if we squint and pretend the term $(1 + \eta_1 y_2 + \lambda_1 y_2^2)^{1/2}$ is not present. Then, estimation would simply be Bernoulli QMLE of $y_{i1}$ on $\mathbf{z}_{i1}$, $y_{i2}$, $\hat{v}_{i2}$, and $y_{i2}\hat{v}_{i2}$, which means that we just add the interaction to the usual Rivers-Vuong procedure. The APE for $y_2$ would be estimated by taking the derivative with respect to $y_2$ and averaging out $\hat{v}_{i2}$, as usual:

$$N^{-1}\sum_{i=1}^{N}(\hat{\alpha}_1 + \hat{\psi}_1\hat{v}_{i2}) \cdot \phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2 + \hat{\theta}_1\hat{v}_{i2} + \hat{\psi}_1 y_2\hat{v}_{i2}), \tag{3.16}$$

and evaluating this at chosen values for $(\mathbf{z}_1, y_2)$ (or using further averaging across the sample values). This simplification cannot be reconciled with (3.9), but it is in the spirit of adding flexibility to a standard approach and treating functional forms as approximations. As a practical matter, we can compare this with the APEs obtained from the standard Rivers-Vuong approach, and a simple test of the null hypothesis that the coefficient on $y_2$ is constant is $H_0 : \psi_1 = 0$ (which should account for the first step estimation of $\hat{\boldsymbol{\pi}}_2$). The null hypothesis that $y_2$ is exogenous is the joint test $H_0 : \theta_1 = 0, \psi_1 = 0$, and in this case no adjustment is needed for the first-stage estimation. And why stop here? If we, add, say, $y_2^2$ to the structural model, we might add $\hat{v}_2^2$ to the estimating equation as well. It would be very difficult to relate parameters estimated from the CF method to parameters in an underlying structural model; indeed, it would be difficult to find a structural model given rise to this particular CF approach. But if the object of interest are the average partial effects, the focus on flexible models for $E(y_1|\mathbf{z}_1, y_2, v_2)$ can be liberating (or disturbing, depending on one's point of view about "structural" parameters).

Lewbel (2000) has made some progress in estimating parameters up to scale in the model $y_1 = 1[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 > 0],$ where $y_2$ might be correlated with $u_1$ and $\mathbf{z}_1$ is a $1 \times L_1$ vector of exogenous variables. Lewbel's (2000) general approach applies to this situation as well. Let $\mathbf{z}$ be the vector of all exogenous variables uncorrelated with $u_1$. Then Lewbel requires a continuous element of $\mathbf{z}_1$ with nonzero coefficient – say the last element, $z_{L_1}$– that does not appear in $D(u_1|y_2, \mathbf{z})$. (Clearly, $y_2$ cannot play the role of the variable excluded from $D(u_1|y_2, \mathbf{z})$ if $y_2$ is thought to be endogenous.) When might Lewbel's exclusion restriction

hold? Sufficient is $y_2 = g_2(\mathbf{z}_2) + v_2,$ where $(u_1, v_2)$ is independent of $\mathbf{z}$ and $\mathbf{z}_2$ does not contain $z_{L_1}$. But this means that we have imposed an exclusion restriction on the reduced form of $y_2$, something usually discouraged in parametric contexts. Randomization of $z_{L_1}$ does *not* make its exclusion from the reduced form of $y_2$ legitimate; in fact, one often hopes that an instrument for $y_2$ is effectively randomized, which means that $z_{L_1}$ does *not* appear in the structural equation but does appear in the reduced form of $y_2$ – the opposite of Lewbel's assumption. Lewbel's assumption on the "special" regressor is suited to cases where a quantity that only affects the response, $y_1$, is randomized. A randomly generated project cost presented to subjects in a willingness-to-pay study is one possibility. Even in such scenarios, one cannot identify the effects of covariates on willingness to pay because coefficients are identified only up to scale.

Returning to the probit response function in (3.9), we can understand the limits of the CF approach for estimating nonlinear models with discrete EEVs. The Rivers-Vuong approach, and its extension to fractional responses, cannot be expected to produce consistent estimates of the parameters or APEs for discrete $y_2$. The problem is that we cannot write

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{3.17}$$

$$D(v_2|\mathbf{z}) = D(v_2) = \text{Normal}(0, \tau_2^2). \tag{3.18}$$

In other words, unlike when we estimate a linear structural equation, the reduced form in the RV approach is not just a linear projection – far from it. In the extreme we have completely specified $D(y_2|\mathbf{z})$ as homoskedastic normal, which is clearly violated if $y_2$ is a binary variable, a count variable, or a corner solution (commonly called a "censored" variable). Unfortunately, even just assuming independence between $v_2$ and $\mathbf{z}$ rules out discrete $y_2$, an assumption that plays an important role even in fully nonparametric approaches. The bottom line is that there

are no known two-step estimation methods that allow one to estimate a probit model or fractional probit model with discrete $y_2$, even if we make strong distributional assumptions.

Possibly because of the absense of valid two-step methods with discrete EEVs, some poor strategies still linger. For example, suppose $y_1$ and $y_2$ are both binary, (3.1) holds, $y_2$ follows the index model

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0], \tag{3.19}$$

and we maintain joint normality of $(u_1, v_2)$ – now both with unit variances – and, of course, independence between the errors and $\mathbf{z}$. Because $D(y_2|\mathbf{z})$ follows a standard probit, it is tempting to try to mimic 2SLS as follows: (i) Run probit of $y_2$ on $\mathbf{z}$ and get the fitted probabilities, $\hat{\Phi}_2 = \Phi(\mathbf{z}\hat{\boldsymbol{\delta}}_2)$. (ii) Run probit of $y_1$ on $\mathbf{z}_1, \hat{\Phi}_2$; that is, just replace each $y_{i2}$ with its fitted probability, $\hat{\Phi}_{i2}$. This does not work, as it would require passing the expected value passes through a nonlinear function. Some have called prodedures like this a "forbidden regression." We could find $E(y_1|\mathbf{z},y_2)$ as a function of the structural and reduced form parameters, insert the first-stage estimates of the RF parameters, and then use binary response estimation in the second stage. But the estimator is not probit with the fitted probabilities plugged in for $y_2$. Currently, the only strategy we have is maximum likelihood estimation based on $f(y_1|y_2,\mathbf{z})f(y_2|\mathbf{z})$, which is not difficult. Wooldridge (2002, Section 15.7.3) contains the likelihood function. [The dearth of options that allow some robustness to distributional assumptions on $y_2$ helps explain why some authors, notably Angrist (2001), have promoted the idea of just using linear probability models estimated by 2SLS. This strategy seems to provide good estimates of the average treatment effect in many applications. But it also seems true that MLE based on joint normality might yield useful approximations to the APEs, too, even if the distributional functions are not entirely correct. Such a view argues for fully robust inference

in the context of misspecified maximum likelihood, as in White (1982).]

An issue that comes up occasionally is whether "bivariate probit" software can be used to estimate the probit model with a binary endogenous variable. In fact, the answer is yes, and the endogenous variables can appear in any way in the model, particularly interacted with exogenous variables. The key is that the likelihood function is constructed from $f(y_1|y_2,\mathbf{x}_1)f_2(y_2|\mathbf{x}_2)$, and so its form does not change if $\mathbf{x}_1$ includes $y_2$. (Of course, one should have at least one exclusion restriction in the case $\mathbf{x}_1$ does depend on $y_2$.) MLE, of course, has all of its desirable properties, and the parameter estimates needed to compute APEs are provided directly.

If $y_1$ is a fractional response satisfying (3.9), $y_2$ follows (3.19), and $(q_1, v_2)$ are jointly normal and independent of $\mathbf{z}$, a two-step method based on $E(y_1|\mathbf{z},y_2)$ is possible; the expectation is not in closed form, and estimation cannot proceed by simply adding a control function to a Bernoulli QMLE. But it should not be difficult to implement. Full MLE for a fractional response is more difficult than for a binary response, particularly if $y_1$ takes on values at the endpoints with positive probability.

An essentially parallel discussion holds for ordered probit response models, where $y_1$ takes on the ordered values $\{0,1,\ldots,J\}$. The RV procedure, and its extensions, applies immediately. In computing partial effects on the response probabilities, we simply average out the reduced for residuals, as in equation (3.8). The comments about the forbidden regression are immediately applicable, too: one cannot simply insert, say, fitted probabilities for the binary EEV $y_2$ into an ordered probit model for $y_1$ and hope for consistent estimates of anything of interest.

Likewise, methods for Tobit models when $y_1$ is a corner solution, such as labor supply or

charitable contributions, are analyzed in a similar fashion. If $y_2$ is a continuous variable, CF methods for consistent estimation can be obtained, at least under the assumptions used in the RV setup. Smith and Blundell (1986) and Wooldridge (2002, Chapter 16) contain treatments. The embellishments described above, such as letting $D(u_1|v_2)$ be a flexible normal distribution, carry over immediately to Tobit case, as do the cautions in looking for simple two-step methods when $D(y_2|\mathbf{z})$ is discrete. Maximum likleihood estimation of all parameters jointly is also quite feasible.

## 3.2. Multinomial and Ordered Responses

Allowing endogenous explanatory variables (EEVs) in multinomial response models is notoriously difficult, even for continuous endogenous variables. There are two basic reasons. First, multinomial probit (MNP), which mixes well well a reduced form normality assumption for $D(y_2|\mathbf{z})$, is still computationally difficult for even a moderate number of choices. Apparently, no one has undertaken a systematic treatment of MNP with EEVs, including how to obtain partial effects.

The multinomial logit (MNL) model and its extensions, such as nested logit and random coefficient versions, are much simpler computationally with lots of alternatives. Unfortunately, the normal distribution does not mix well with the extreme value distribution, and so, if we begin with a structural MNL model (or conditional logit), the estimating equations obtained from a CF approach are difficult to obtain, and MLE is very difficult, too, even if we assume a normal distribution in the reduced form(s).

Recently, some authors have suggested taking a practical approach to allowing continuous EEVs in multinomial response. The suggestions for binary and fractional responses in the

previous subsection – namely, use probit, or even logit, with flexible functions of both the observed variables and the reduced form residuals – is in this spirit.

Again it is convenient to model the source of endogeneity as an omitted variable. Let $y_1$ be the (unordered) multinomial response taking values $\{0, 1, \ldots, J\}$, let $\mathbf{z}$ be the vector of endogenous variables, and let $\mathbf{y}_2$ be a vector of endogenous variables. If $r_1$ represents omitted factors that the researcher would like to control for, then the structural model consists of specifications for the response probabilities

$$\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1), j = 0, 1, \ldots, J. \tag{3.20}$$

The average partial effects, as usual, are obtained by averaging out the unobserved heterogeneity, $r_1$. Assume that $\mathbf{y}_2$ follows the linear reduced form

$$\mathbf{y}_2 = \mathbf{z}\boldsymbol{\Pi}_2 + \mathbf{v}_2. \tag{3.21}$$

Typically, at least as a first attempt, we would assume a convenient joint distribution for $(r_1, \mathbf{v}_2)$, such as multivariate normal and independent of $\mathbf{z}$. This approach has been applied when the response probabilities, conditional on $r_1$, have the conditional logit form. For example, Villas-Boas and Winer (1999) apply this approach to modeling brand choice, where prices are allowed to correlated with unobserved tastes that affect brand choice. In implementing the CF approach, the problem in starting with a multinomial or conditional logit model for (3.20) is computational. Nevertheless, estimation is possible, particular if one uses simulation methods of estimation briefly mentioned in the previous subsection.

A much simpler control function approach is obtained if we skip the step of modeling $\mathrm{P}(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ and jump directly to convenient models for $\mathrm{P}(y_1 = j|\mathbf{z}_{i1}, \mathbf{y}_2, \mathbf{v}_2) = \mathrm{P}(y_1 = j|\mathbf{z}, \mathbf{y}_2)$. Petrin and Train (2006) are proponents of this solution.

The idea is that any parametric model for $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ is essentially arbitrary, so, if we can recover quantities of interest directly from $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, why not specify these probabilities directly? If we assume that $D(r_1|\mathbf{z}, \mathbf{y}_2) = D(r_1|\mathbf{v}_2)$, and that $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ can be obtained from $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, r_1)$ by integrating the latter with respect to $D(r_1|\mathbf{v}_2)$ then we can estimate the APEs directly from $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ by averaging out across the reduced form residuals, as in previous cases.

Once we have selected a model for $P(y_1 = j|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, which could be multinomial logit, conditional logit, or nested logit, we can apply a simple two-step procedure. First, estimate the reduced form for $\mathbf{y}_{i2}$ and obtain the residuals, $\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \mathbf{z}_i\hat{\mathbf{\Pi}}_2$. (Alternatively, we can use strictly monotonic transformations of the elements of $\mathbf{y}_{i2}$.) Then, we estimate a multinomial response model with explanatory variables $\mathbf{z}_{i1}, \mathbf{y}_{i2},$ and $\hat{\mathbf{v}}_{i2}$. As always with control function approaches, we need enough exclusion restrictions in $\mathbf{z}_{i1}$ to identify the parameters and APEs. We can include nonlinear functions of $(\mathbf{z}_{i1}, \mathbf{y}_{i2}, \hat{\mathbf{v}}_{i2})$, including quadratics and interactions for more flexibility.

Given estimates of the probabilities $p_j(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$, we can estimate the average partial effects on the structural probabilities by estimating the average structural function:

$$\widehat{\text{ASF}}(\mathbf{z}_1, \mathbf{y}_2) = N^{-1} \sum_{i=1}^{N} p_j(\mathbf{z}_1, \mathbf{y}_2, \hat{\mathbf{v}}_{i2}). \tag{3.22}$$

Then, we can take derivatives or changes of $\widehat{\text{ASF}}(\mathbf{z}_1, \mathbf{y}_2)$ with respect to elements of $(\mathbf{z}_1, \mathbf{y}_2)$, as usual. While the delta method can be used to obtain analytical standard errors, the bootstrap is simpler and feasible if one uses, say, conditional logit.

In an application to choice of television service, Petrin and Train (2006) find the CF

approach gives remarkably similar parameter estimates to the approach proposed by Berry, Pakes, and Levinsohn (1995), which we touch on in the cluster sample notes.

When the EEVs are discrete, the CF arguments above do not apply. One can often implement maximum likelihood without too much difficulty. For example, Adams, Chiang, and Jensen (2003) use MLE when the scalar $y_2$ follows an ordered probit.

## 3.3. Exponential Models

Exponential models represent a middle ground between linear models and discrete response models: to allow for EEVs in an exponential model, we need to impose more assumptions than needed for standard linear models but fewer assumptions than discrete response models. Both IV approaches and CF approaches are available for exponential models, the latter having been worked out for continuous and binary EEVs. With a single EEV, write

$$E(y_1|\mathbf{z},y_2,r_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + r_1), \tag{3.23}$$

where $r_1$ is the omitted variable. (Extensions to general nonlinear functions of $(\mathbf{z}_1, y_2)$ are immediate; we just add those functions with linear coefficients to (3.23). Leading cases are polynomials and interactions.) Suppose first that $y_2$ has a standard linear reduced form with an additive, independent error:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 \tag{3.24}$$

$$D(r_1, v_2|\mathbf{z}) = D(r_1, v_2), \tag{3.25}$$

so that $(r_1, v_2)$ is independent of $\mathbf{z}$. Then

$$E(y_1|\mathbf{z},y_2) = E(y_1|\mathbf{z}, v_2) = E[\exp(r_1)|v_2]\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2). \tag{3.26}$$

If $(r_1, v_2)$ are jointly normal, then $E[\exp(r_1)|v_2] = \exp(\theta_1 v_2)$, where we set the intercept to

zero, assuming $\mathbf{z}_1$ includes an intercept. This assumption can hold more generally, too. Then

$$E(y_1|\mathbf{z},y_2) = E(y_1|\mathbf{z},v_2) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2), \tag{3.27}$$

and this expectation immediately suggest a two-step estimation procedure. The first step, as before, is to estimate the reduced form for $y_2$ and obtain the residuals. Then, include $\hat{v}_2$, along with $\mathbf{z}_1$ and $y_2$, in nonlinear regression or, especially if $y_1$ is a count variable, in a Poisson QMLE analysis. Like NLS, it requires only (3.27) to hold. A $t$ test of $H_0 : \theta_1 = 0$ is valid as a test that $y_2$ is exogenous. Average partial effects on the mean are obtained from

$$\left[ N^{-1} \sum_{i=1}^{N} \exp(\hat{\theta}_1 \hat{v}_{i2}) \right] \exp(\mathbf{z}_1 \hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2).$$

Proportionate effects on the expected value, that is elasticities and semi-elasticities, do not depend on the scale factor out front.

Like in the binary case, we can use a random coefficient model to suggest more flexible CF methods. For example, if we start with

$$\begin{aligned} E(y_1|\mathbf{z},y_2,a_1,r_1) &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + a_1 y_2 + r_1) \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + d_1 y_2 + r_1) \end{aligned} \tag{3.28}$$

and assume trivariate normality of $(d_1, r_1, v_2)$ (and independence from $\mathbf{z}$), then it can be shown that

$$\begin{aligned} E(y_1|\mathbf{z},v_2) = \exp(&\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2 \\ &+ (\sigma_r^2 + 2\sigma_{dr} y_2 + \sigma_d^2 y_2^2)/2). \end{aligned} \tag{3.29}$$

Therefore, the estimating equation involves a quadratic in $y_2$ and an interaction between $y_2$ and $v_2$. Notice that the term $(\sigma_r^2 + 2\sigma_{dr} y_2 + \sigma_d^2 y_2^2)/2$ is present even if $y_2$ is exogenous, that is, $\theta_1 = \psi_1 = 0$. If $\sigma_{dr} = Cov(d_1, r_1) \neq 0$ then (3.29) does not even identify $\alpha_1 = E(a_1)$ (we

would have to use higher-order moments, such as a variance assumption). But (3.29) *does* identify the average structural function (and, therefore, APEs). We just absorb $\sigma_r^2$ into the intercept, combine the linear terms in $y_2$, and add the quadratic in $y_2$. So, we would estimate

$$E(y_1|\mathbf{z}, v_2) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \rho_1 y_2 + \theta_1 v_2 + \psi_1 y_2 v_2 + \eta_1 y_2^2) \tag{3.30}$$

using a two-step QMLE. The ASF is more complicated, and estimated as

$$\widehat{ASF}(\mathbf{z}_1, y_2) = \left[ N^{-1} \sum_{i=1}^{N} \exp(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\rho}_1 y_2 + \hat{\theta}_1 \hat{v}_{i2} + \hat{\psi}_1 y_2 \hat{v}_{i2} + \hat{\eta}_1 y_2^2) \right], \tag{3.31}$$

which, as in the probit example, implies that the APE with respect to $y_2$ need not have the same sign as $\alpha_1$.

Our inability to estimate $\alpha_1$ even in this very parametric setting is just one example of how delicate identification of parameters in standard index models can be. Natural extensions to models with random slopes generally cause even the mean heterogeneity ($\alpha_1$ above) to be unidentified. Again, it must be emphasized that the loss of identification holds even if $y_2$ is assumed exogenous.

If $y_2$ is a binary model following a probit, then a CF approach due to Terza (1998) can be used. We return to the model in (3.23) where, for simplicity, we assume $y_2$ is not interacted with elements of $\mathbf{z}_1$; the extension is immediate.We can no longer assume (3.24) and (3.25). Instead, replace (3.24)

$$y_2 = 1[\mathbf{z}\boldsymbol{\pi}_2 + v_2 > 0] \tag{3.32}$$

and still adopt (3.25). In fact, we assume $(r_1, v_2)$ is jointly normal. To implement a CF approach, we need to find

$$\begin{aligned} E(y_1|\mathbf{z},y_2) &= E[E(y_1|\mathbf{z},v_2)|\mathbf{z},y_2] \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)E[\exp(\eta_1 + \theta_1 v_2)|\mathbf{z},y_2] \\ &= \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2,\mathbf{z}\boldsymbol{\pi}_2,\theta_1), \end{aligned} \tag{3.34}$$

where we absorb $\eta_1$ into the intercept in $\mathbf{z}_1$ without changing notation and

$$\begin{aligned} h(y_2,\mathbf{z}\boldsymbol{\pi}_2,\theta_1) = \exp(\theta_1^2/2)\{&y_2\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2) \\ &+ (1-y_2)[1-\Phi(\theta_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1-\Phi(\mathbf{z}\boldsymbol{\pi}_2)]\}, \end{aligned} \tag{3.35}$$

as shown by Terza (1998). Now, $\boldsymbol{\pi}_2$ is estimated by a first-stage probit, and then NLS or, say, Poisson QMLE can be applied to the mean function

$$\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)h(y_2,\mathbf{z}\hat{\boldsymbol{\pi}}_2,\theta_1). \tag{3.36}$$

As usual, unless $\theta_1 = 0$, one must account for the estimation error in the first step when obtaining inference in the second. Terza (1998) contains analytical formulas, or one may use the bootstrap.

In the exponential case, an alternative to either of the control function approaches just presented is available – and, it produces consistent estimators regardless of the nature of $y_2$. Write $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1,\mathbf{y}_2)$ as any function of exogenous and endogenous variables. If we start with

$$E(y_1|\mathbf{z},\mathbf{y}_2,r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1) \tag{3.37}$$

then we can use a transformation due to Mullahy (1997) to consistently estimate $\boldsymbol{\beta}_1$ by method of moments. By definition, and assuming only that $y_1 \geq 0$, we can write

$$\begin{aligned} y_1 &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1)a_1 \\ &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1)\exp(r_1)a_1,\ E(a_1|\mathbf{z},\mathbf{y}_2,r_1) = 1. \end{aligned}$$

If $r_1$ is independent of $\mathbf{z}$ then

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1|\mathbf{z}] = E[\exp(r_1)|\mathbf{z}] = E[\exp(r_1)] = 1, \tag{3.38}$$

where the last equality is just a normalization that defines the intercept in $\boldsymbol{\beta}_1$. Therefore, we have conditional moment conditions

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 - 1|\mathbf{z}] = 0, \tag{3.39}$$

which depends on the unknown parameters $\boldsymbol{\beta}_1$ and observable data. Any function of $\mathbf{z}$ can be used as instruments in a nonlinear GMM procedure. An important issue in implementing the procedure is choosing instruments. See Mullahy (1997) for further discussion.

# 4. Semiparametric and Nonparametric Approaches

Blundell and Powell (2004) show how to relax distributional assumptions on $(u_1, v_2)$ in the model model $y_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + u_1 > 0]$, where $\mathbf{x}_1$ can be any function of $(\mathbf{z}_1, y_2)$. The key assumption is that $y_2$ can be written as $y_2 = g_2(\mathbf{z}) + v_2,$ where $(u_1, v_2)$ is independent of $\mathbf{z}$. The independence of the additive error $v_2$ and $\mathbf{z}$ pretty much rules out discreteness in $y_2$, even though $g_2(\cdot)$ can be left unspecified. Under the independence assumption,

$$\mathrm{P}(y_1 = 1|\mathbf{z}, v_2) = \mathrm{E}(y_1|\mathbf{z}, v_2) = H(\mathbf{x}_1\boldsymbol{\beta}_1, v_2) \tag{4.1}$$

for some (generally unknown) function $H(\cdot, \cdot)$. The average structural function is just $\mathrm{ASF}(\mathbf{z}_1, y_2) = \mathrm{E}_{v_{i2}}[H(\mathbf{x}_1\boldsymbol{\beta}_1, v_{i2})]$. We can estimate $H$ and $\boldsymbol{\beta}_1$ quite generally by first estimating the function $g_2(\cdot)$ and then obtaining residuals $\hat{v}_{i2} = y_{i2} - \hat{g}_2(\mathbf{z}_i)$. Blundell and Powell (2004) show how to estimate $H$ and $\boldsymbol{\beta}_1$ (up to scale) and $G(\cdot)$, the distribution of $u_1$. The ASF is obtained from $G(\mathbf{x}_1\boldsymbol{\beta}_1)$. We can also estimate the ASF by averaging out the reduced form residuals,

$$\widehat{\text{ASF}}(\mathbf{z}_1, y_2) = N^{-1} \sum_{i=1}^{N} \hat{H}(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_1, \hat{v}_{i2}); \tag{4.2}$$

derivatives and changes can be computed with respect to elements of $(\mathbf{z}_1, y_2)$.

Blundell and Powell (2003) allow $P(y_1 = 1|\mathbf{z}, y_2)$ to have the general form $H(\mathbf{z}_1, y_2, v_2)$, and then the second-step estimation is entirely nonparametric. They also allow $\hat{g}_2(\cdot)$ to be fully nonparametric. But parametric approximations in each stage might produce good estimates of the APEs. For example, $y_{i2}$ can be regressed on flexible functions of $\mathbf{z}_i$ to obtain $\hat{v}_{i2}$. Then, one can estimate probit or logit models in the second stage that include functions of $\mathbf{z}_1$, $y_2$, and $\hat{v}_2$ in a flexible way – for example, with levels, quadratics, interactions, and maybe even higher-order polynomials of each. Then, one simply averages out $\hat{v}_{i2}$, as in equation (4.2). Valid standard errors and test statistics can be obtained by bootstrapping or by using the delta method.

In certain cases, an even more parametric approach suggests itself. Suppose we have the exponential regression

$$E(y_1|\mathbf{z}, y_2, r_1) = \exp(\mathbf{x}_1 \boldsymbol{\beta}_1 + r_1), \tag{4.3}$$

where $r_1$ is the unobservable. If $y_2 = \mathbf{g}_2(\mathbf{z})\boldsymbol{\pi}_2 + v_2$ and $(r_1, v_2)$ is independent of $\mathbf{z}$, then

$$E(y_1|\mathbf{z}_1, y_2, v_2) = h_2(v_2) \exp(\mathbf{x}_1 \boldsymbol{\beta}_1), \tag{4.4}$$

where now $h_2(\cdot)$ is an unknown function. It can be approximated using series, say, and, of course, first-stage residuals $\hat{v}_2$ replace $v_2$.

Blundell and Powell (2003) consider a very general setup, which starts with $y_1 = g_1(\mathbf{z}_1, \mathbf{y}_2, u_1)$, and then discuss estimation of the ASF, given by

$$ASF_1(\mathbf{z}_1, \mathbf{y}_2) = \int g_1(\mathbf{z}_1, \mathbf{y}_2, u_1) dF_1(u_1), \tag{4.5}$$

where $F_1$ is the distribution of $u_1$. The key restrictions are that $\mathbf{y}_2$ can be written as

$$\mathbf{y}_2 = \mathbf{g}_2(\mathbf{z}) + \mathbf{v}_2, \tag{4.6}$$

where $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$. The additive, independent reduced form errors in (4.6) effectively rule out applications to discrete $\mathbf{y}_2$. Conceptually, Blundell and Powell's method is straightforward, as it is a nonparametric extenstion of parametric approaches. First, estimate $\mathbf{g}_2$ nonparametrically (which, in fact, may be done via a flexible parametric model, or kernel estimators). Obtain the residuals $\hat{\mathbf{v}}_{i2} = \mathbf{y}_{i2} - \hat{\mathbf{g}}_2(\mathbf{z}_i)$. Next, estimate $E(y_1|\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2) = h_1(\mathbf{z}_1, \mathbf{y}_2, \mathbf{v}_2)$ using nonparametrics, where $\hat{\mathbf{v}}_{i2}$ replaces $\mathbf{v}_2$. Identification of $h_1$ holds quite generally, provided we have sufficient exclusion restrictions (elements in $\mathbf{z}$ not in $\mathbf{z}_1$). BP discuss some potential pitfalls. Once we have $\hat{h}_1$, we can consistently estimate the ASF. For given $\mathbf{x}_1^o = (\mathbf{z}_1^o, \mathbf{y}_2^o)$, the ASF can always be written, using iterated expectations, as

$$E_{\mathbf{v}_2}\{E[g_1(\mathbf{x}_1^o, u_1)|\mathbf{v}_2]\}.$$

Under the assumption that $(u_1, \mathbf{v}_2)$ is independent of $\mathbf{z}$, $E[g_1(\mathbf{x}_1^o, u_1)|\mathbf{v}_2] = h_1(\mathbf{x}_1^o, \mathbf{v}_2)$ – that is, the regression function of $y_1$ on $(\mathbf{x}_1, \mathbf{v}_2)$. Therefore, a consistent estimate of the ASF is

$$N^{-1}\sum_{i=1}^{N} \hat{h}_1(\mathbf{x}_1, \hat{\mathbf{v}}_{i2}). \tag{4.7}$$

While semiparametric and parametric methods when $y_2$ (or, more generally, a vector $\mathbf{y}_2$) are continuous – actually, have a reduced form with an additive, independent error – they do not currently help us with discrete EEVs.

With univariate $y_2$, it possible to relax the additivity of $v_2$ in the reduced form equation under monotonicity assumptions. Like Blundell and Powell (2003), Imbens and Newey (2006) consider the triangular system, but without additivity in the reduced form of $y_2$. The structureal

equation is

$$y_1 = g_1(\mathbf{z}_1, y_2, \mathbf{u}_1), \tag{4.8}$$

where $\mathbf{u}_1$ is a vector heterogeneity (whose dimension may not even be known), and the reduced form for $y_2$ is

$$y_2 = g_2(\mathbf{z}, e_2), \tag{4.9}$$

where $g_2(\mathbf{z}, \cdot)$ is strictly monotonic. This assumption rules out discrete $y_2$ but allows some interaction between the unobserved heterogeneity in $y_2$ and the exogenous variables. As one special case, Imbens and Newey show that, if $(\mathbf{u}_1, e_2)$ is assumed to be independent of $\mathbf{z}$, then a valid control function that can be used in a second stage is $v_2 \equiv F_{y_2|\mathbf{z}}(y_2, \mathbf{z})$, where $F_{y_2|\mathbf{z}}$ is the conditional distribution of $y_2$ given $\mathbf{z}$. Imbens and Newey describe identification of various quantities of interest, including the quantile structural function (QSF). When $u_1$ is a scalar and monotonically increasing in $u_1$, the QSF is

$$QSF_\tau(\mathbf{x}_1) = g_1(\mathbf{x}_1, \text{Quant}_\tau(u_1)), \tag{4.10}$$

where $\text{Quant}_\tau(u_1)$ is the $\tau^{th}$ of $u_1$. We consider quantile methods in more detail in the quantile methods notes.

# 5. Methods for Panel Data

We can combine methods for handling correlated random effects models with control function methods to estimate certain nonlinear panel data models with unobserved heterogeneity and EEVs. Here we provide as an illustration a parametric approach used by Papke and Wooldridge (2008), which applies to binary and fractional responses. The

manipulations are routine but point to more flexible ways of estimating the average marginal effects. It is important to remember that we currently have no way of estimating, say, unobserved effects models for fractional response variables, either with or without endogenous explanatory variables, without imposing some restrictions on the distribution of heterogeneity given the exogenous variables. Even the approaches that treat the unobserved effects as parameters – and use large $T$ approximations – to not allow endogenous regressors. Plus, recall from the nonlinear panel data notes that most results are for the case where the data are assumed independent across time. Jackknife approaches further assume homogeneity across time.

We write the model with time-constant unobserved heterogeneity, $c_{i1}$, and time-varying unobservables, $v_{it1}$, as

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, c_{i1}, v_{it1}) = E(y_{it1}|y_{it2}, \mathbf{z}_{it1}, c_{i1}, v_{it1}) = \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1}). \tag{5.1}$$

Thus, there are two kinds of potential omitted variables. We allow the heterogeneity, $c_{i1}$, to be correlated with $y_{it2}$ and $\mathbf{z}_i$, where $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT})$ is the vector of strictly exogenous variables (conditional on $c_{i1}$). The time-varying omitted variable is uncorrelated with $\mathbf{z}_i$ – strict exogeneity – but may be correlated with $y_{it2}$. As an example, $y_{it1}$ is a female labor force participation indicator and $y_{it2}$ is other sources of income. Or, $y_{it1}$ is a test pass rate, and the school leve, and $y_{it2}$ is a measure of spending per student.

When we write $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$, we are assuming $\mathbf{z}_{it2}$ can be excluded from the "structural" equation (4.1). This is the same as the requirement for fixed effects two stage least squares estimation of a linear model.

To proceed, we first model the heterogeneity using a Chamberlain-Mundlak approach:

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}, a_{i1}|\mathbf{z}_i \sim \text{Normal}(0, \sigma^2_{a_1}). \tag{5.2}$$

We could allow the elements of $\mathbf{z}_i$ to appear with separate coefficients, too. Note that only exogenous variables are included in $\bar{\mathbf{z}}_i$. Plugging into (5.1) we have

$$\begin{aligned} E(y_{it1}|y_{it2}, \mathbf{z}_i, a_{i1}, v_{it1}) &= \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1} + v_{it1}) \\ &\equiv \Phi(\alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + r_{it1}). \end{aligned} \tag{5.3}$$

Next, we assume a linear reduced form for $y_{it2}$:

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\delta_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2}, t = 1, \ldots, T, \tag{5.4}$$

where, if necessary, we can allow the coefficients in (5.4) to depend on $t$. The addition of the time average of the strictly exogenous variables in (5.4) follows from the Mundlak (1978) device. The nature of endogeneity of $y_{it2}$ is through correlation between $r_{it1} = a_{i1} + v_{it1}$ and the reduced form error, $v_{it2}$. Thus, $y_{it2}$ is allowed to be correlated with unobserved heterogeneity and the time-varying omitted factor. We also assume that $r_{it1}$ given $v_{it2}$ is conditionally normal, which we write as

$$r_{it1} = \eta_1 v_{it2} + e_{it1}, \tag{5.5}$$

$$e_{it1}|(\mathbf{z}_i, v_{it2}) \sim \text{Normal}(0, \sigma^2_{e_1}), t = 1, \ldots, T. \tag{5.6}$$

Because $e_{it1}$ is independent of $(\mathbf{z}_i, v_{it2})$, it is also independent of $y_{it2}$. Using a standard mixing property of the normal distribution,

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2}) = \Phi(\alpha_{e1} y_{it2} + \mathbf{z}_{it1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2}) \tag{5.7}$$

where the "$e$" subscript denotes division by $(1 + \sigma^2_{e_1})^{1/2}$. This equation is the basis for CF estimation.

The assumptions used to obtain (5.7) would not hold for $y_{it2}$ having discreteness or

substantively limited range in its distribution. It is straightfoward to include powers of $v_{it2}$ in (5.7) to allow greater flexibility. Following Wooldridge (2005) for the cross-sectional case, we could even model $r_{it1}$ given $v_{it2}$ as a heteroskedastic normal.

In deciding on estimators of the parameters in (5.7), we must note that the explanatory variables, while contemporaneous exogenous by construction, are not usually strictly exogenous. In particular, we allow $y_{is2}$ to be correlated with $v_{it1}$ for $t \neq s$. Therefore, generalized estimation equations, that assume strict exogeneity – see the notes on nonlinear panel data models – will not be consistent in general. We could apply method of moments procedures. A simple approach is to use use pooled nonlinear least squares or pooled quasi-MLE, using the Bernoulli log likelihood. (The latter fall under the rubric of generalized linear models.) Of course, we want to allow arbitrary serial dependence and $Var(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2})$ in obtaining inference, which means using a robust sandwich estimator.

The two step procedure is (i) Estimate the reduced form for $y_{it2}$ (pooled across $t$, or maybe for each $t$ separately; at a minimum, different time period intercepts should be allowed). Obtain the residuals, $\hat{v}_{it2}$ for all $(i,t)$ pairs. The estimate of $\boldsymbol{\delta}_2$ is the fixed effects estimate. (ii) Use the pooled probit QMLE of $y_{it1}$ on $y_{it2}, \mathbf{z}_{it1}, \bar{\mathbf{z}}_i, \hat{v}_{it2}$ to estimate $\alpha_{e1}, \boldsymbol{\delta}_{e1}, \psi_{e1}, \boldsymbol{\xi}_{e1}$ and $\eta_{e1}$.

Because of the two-step procedure, the standard errors in the second stage should be adjusted for the first stage estimation. Alternatively, bootstrapping can be used by resampling the cross-sectional units. Conveniently, if $\eta_{e1} = 0$, the first stage estimation can be ignored, at least using first-order asymptotics. Consequently, a test for endogeneity of $y_{it2}$ is easily obtained as an asymptotic $t$ statistic on $\hat{v}_{it2}$; it should be make robust to arbitrary serial correlation and misspecified variance. Adding first-stage residuals to test for endogeneity of an explanatory variables dates back to Hausman (1978). In a cross-sectional contexts, Rivers and

Vuong (1988) suggested it for the probit model.

Estimates of average partial effects are based on the average structural function

$$E_{(c_{i1},v_{it1})}[\Phi(\alpha_1 y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_1 + c_{i1} + v_{it1})] \tag{5.8}$$

with respect to the elements of $(y_{t2}, \mathbf{z}_{t1})$. It can be shown that

$$E_{(\bar{\mathbf{z}}_i,v_{it2})}[\Phi(\alpha_{e1} y_{t2} + \mathbf{z}_{t1}\boldsymbol{\delta}_{e1} + \psi_{e1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{e1} + \eta_{e1} v_{it2})]; \tag{5.9}$$

that is, we "integrate out" $(\bar{\mathbf{z}}_i, v_{it2})$ and then take derivatives or changes with respect to the elements of $(\mathbf{z}_{t1} y_{t2})$. Because we are not making a distributional assumption about $(\bar{\mathbf{z}}_i, v_{it2})$, we instead estimate the APEs by averaging out $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$ across the sample, for a chosen *t*:

$$N^{-1}\sum_{i=1}^{N} \Phi(\hat{\alpha}_{e1} y_{t2} + \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_{e1} + \hat{\psi}_{e1} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{e1} + \hat{\eta}_{e1}\hat{v}_{it2}). \tag{5.10}$$

APEs computed from (5.10) – typically with further averaging out across *t* and the values of $y_{t2}$ and $\mathbf{z}_{t1}$ – can be compared directly with linear model estimates, particular fixed effects IV estimates.

We can use the approaches of Altonji and Matzkin (2005) and Blundell and Powell (2003) to make the analysis less parametric. For example, we might replace (5.4) with $y_{it2} = g_2(\mathbf{z}_{it}, \bar{\mathbf{z}}_i) + v_{it2}$ (or use functions in addition to , $\bar{\mathbf{z}}_i$, as in AM). Then, we could maintain

$$D(c_{i1} + v_{it1}|\mathbf{z}_i, y_{it2}) = D(c_{i1} + v_{it1}|\bar{\mathbf{z}}_i, v_{it2}).$$

In the first estimation step, $\hat{v}_{it2}$ is obtained from a nonparametric or semiparametric pooled estimation. Then the function

$$E(y_{it1}|y_{it2}, \mathbf{z}_i, v_{it2}) = h_1(\mathbf{x}_{it1}\boldsymbol{\beta}_1, \bar{\mathbf{z}}_i, v_{it2})$$

can be estimated in a second stage, with the first-stage residuals, $\hat{v}_{it2}$, inserted. Generally,

identification holds because the $v_{it2}$ varying over time separately from $\mathbf{x}_{it1}$ due to time-varying exogenous instruments $\mathbf{z}_{it2}$. The inclusion of $\bar{\mathbf{z}}_i$ requires that we have at least one time-varying, strictly exogenous instrument for $y_{it2}$.

# References

Adams, J.D., E.P. Chiang, and J.L. Jensen (2003), "The Influence of Federal Laboratory R&D on Industrial Research," *Review of Economics and Statistics* 85, 1003-1020.

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

Angrist, J.D. (1991), "Estimations of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics* 19, 2-16.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

Blundell, R. and J.L. Powell (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econonometrics: Theory and Applications*, Eighth World Congress, Volume 2, M. Dewatripont, L.P. Hansen and S.J. Turnovsky, eds. Cambridge: Cambridge University Press, 312-357.

Blundell, R. and J.L. Powell (2004), "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies* 71, 655-679.

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

Garen, J. (1984), "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica* 52, 1199-1218.

Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251-1271.

Heckman, J.J. (1976), "The Common Structure of Statistical Models of Truncation, Sample

Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475-492.

Heckman, J.J. and T.E. MaCurdy (1986), "Labor Econometrics," in *Handbook of Econometrics*, Volume 3. Z. Griliches and M.D. Intriligator (eds.), 1918-1977. Amsterdam: Elsevier.

Heckman, J.J. and E. Vytlacil (1998), "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling," *Journal of Human Resources* 33, 974-987.

Imbens, G.W. and W.K. Newey (2006), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," mimeo, MIT Department of Economics.

Lee, L.-F. (1982), "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies* 49, 355-372.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics* 97, 145-177.

Mullahy, J. (1997), "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics* 79, 586-593.

Mundlak, Y. (1978), "On the Pooling of Time Series and Cross Section Data, *Econometrica* 46, 69-85.

Newey, W.K. (1988), "Adaptive Estimation of Regression Models via Moment Restrictions," *Journal of Econometrics* 38, 301-339.

Papke, L.E. and J.M. Wooldridge (2008), "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates," forthcoming, *Journal of Econometrics*.

Petrin, A. and K. Train (2006), "Control Function Corrections for Unobserved Factors in Differentiated Product Models," mimeo, University of Minnesota Department of Economics.

Powell, J.L. (1994), "Estimation of Semiparametric Models," in *Handbook of Econometrics*, Volume 4. R.F. Engle and D.L. McFadden (eds.), 2443-2521. Amsterdam: Elsevier.

Rivers, D. and Q.H. Vuong (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics* 39, 347-366.

Smith, R.J., and R.W. Blundell (1986), "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica* 54, 679-685.

Terza, J.V. (1998), "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects," *Journal of Econometrics* 84, 129-154.

Villas-Boas, J.M. and R.S. Winer (1999), "Endogeneity in Brand Choice Models," *Management Science* 45, 1324-1338.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50, 1-25.

Wooldridge, J.M. (1997), "On Two Stage Least Squares Estimation of the Average Treatment Effect in Random Coefficient Models," *Economics Letters* 56, 129-133.

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

Wooldridge, J.M. (2003), "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters* 79, 185-191.

Wooldridge, J.M. (2005), "Unobserved Heterogeneity and Estimation of Average Partial

Effects," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. D.W.K. Andrews and J.H. Stock (eds.), 27-55. Cambridge: CUP.