Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory Variables

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
East Lansing, MI 48824-1038
wooldri1@msu.edu

This version: June 2012

Abstract: This paper proposes a quasi-maximum likelihood framework for estimating nonlinear models with continuous or discrete endogenous explanatory variables. Both joint and two-step estimation procedures are considered. The joint estimation procedure can be viewed as quasi-limited information maximum likelihood, as one or both of the log likelihoods used may be misspecified. The two-step control function approach is computationally simple and leads to straightforward tests of endogeneity. In the case of discrete endogenous explanatory variables, I argue that the control function approach can be applied to generalized residuals to obtain average partial effects. The general results are applied to nonlinear models for fractional and nonnegative responses.

Keywords: Quasi-Maximum Likelihood, Control Function, Linear Exponential Family, Average Structural Function

1. Introduction

The most common method of estimating a linear model with one or more endogenous explanatory variables is two stage least squares (2SLS). Nevertheless, several authors have argued that the limited information maximum likelihood estimator – obtained under the nominal assumption of jointly normally distributed unobservables – can have better small-sample properties, particularly when there are many overidentifying restrictions (possibly in combination with "weak" instruments). See, for example, Bekker (1994) and Staiger and Stock (1997). Imbens and Wooldridge (2009) contains a summary and several references.

It is well known that endogeneity among explanatory variables is generally more difficult to handle in nonlinear models, although several special cases have been worked out. Unlike with a linear model with constant coefficients, where 2SLS can always be applied regardless of the nature of the endogenous explanatory variables (EEVs), with nonlinear models the probabilistic nature the EEVs – whether they are continuous, discrete, or some combinaton – plays a critical role. Methods where fitted values obtained in a first stage are plugged in for the EEVs in a second stage are generally inconsistent for the both the "structural" parameters and other quantities of interest, such as average partial (or marginal) effects.

For the most part, two approaches are used to estimate nonlinear models with EEVs. Maximum likelihood (conditional on the exogenous variables) is, in principle, available when a distribution (conditional on exogenous varaibles) for the EEVs is fully specified and a distribution of the response variable conditional on the EEVs (and exogenous variables) is specified or derived from a set of equations with unobserved errors. The MLE approach has been widely applied, especially for binary responses, but it has some limitations. For one, it can be computationally difficult with multiple EEVs. Perhaps more importantly, it requires specification of a full set of conditional distributions, and it is generally not robust if those assumptions are wrong.

A second general approach is a control function approach, where residuals from a first stage estimation procedure involving the EEVs are inserted into a second stage estimation problem. Rivers and Vuong (1988) for probit models and Smith and Blundell (1986) for Tobit models are popular examples. Wooldridge (2010) uses the control function (CF) approach in a variety of settings, including nonlinear models with cross section data or panel data. Recently, Blundell and Powell (2003, 2004) have shown that the approach has broad applicability in semiparametric and even nonparametric settings. BP show that quantities of interest – partial effects of the average structural function – are identified very generally, without distributional or functional form restrictions. (Wooldridge, 2005, argues that the concepts of the average structural function and average partial effects are similar, but the APE approach is more flexible in that it can allow unobservables that are not assumed to be independent of exogenous covariates.) In some cases, the practice of inserting first-stage fitted values for EEVs can produce consistent estimators of parameters up to a common scale factor, but the assumptions under which this occurs are very restrictive, and average partial effects are not easy to recover. In addition, the “fitted-value method” does not allow simple tests of the null that the suspected EEVs are exogenous.

The main drawback of the CF approach – even in BP’s general setting – is that the nature of the EEVs is restricted. It must be assumed that the reduced forms of the EEVs have additive errors that are independent of the variables exogenous in the structural equation. The assumption of additive, independent errors rules out discrete EEVs. Thus, while the BP approach allows for general response functions, its scope is restricted because it does not allow general EEVs.

Since the work of White (1982), econometricians have known that the parameter estimators obtained from maximum likelihood estimation of misspecified models can be given a useful interpretation, and it is possible to perform inference. Further, we know there are special cases where the so-called quasi-MLE actually identifies population parameters that index some feature of the distribution. See Gourieroux, Monfort, and Trognon (1984) for the case of conditional means and conditional variances. The first contribution of this paper is to show that there are situations of practical interest where using joint MLEs has significant robustness properties. In other words, even in models with EEVs, certain quasi-MLEs identify interesting parameters. Two important examples are the MLEs obtained for fractional responses with either continuous or binary EEVs. As it turns out, the log likelihood function for a binary response can be applied to fractional response variables under a conditional mean assumption without further restricting the conditional distribution. A practically important implication is that a joint estimation procedure that is available for binary responses can be applied to fractional responses. Because the method is quasi-MLE, robust inference needs to be used because the information matrix equality is generally violated.

A second example is when the response variable is a count variable or any other variable with an exponential conditional mean function. The general results here show that one could maximize a joint quasi-log likelihood associated with a Poisson response and a binary EEV. The one-step nature of the estimation procedure might improve over available two-step estimators, such as the one proposed by Terza (1998), while being just as robust and possibly more efficient.

A second contribution of the paper is to derive a class of tests for endogeneity in nonlinear models. The score principle is convenient for obtaining robust variable addition tests (VATs). Generally, the variable added to a standard second-stage MLE or quasi-MLE is a generalized residual.

The third contribution is more controversial. I propose an extension of the BP approach by suggesting that we adopt independence assumptions of unobservables in the structural equation conditional on generalized residuals obtained from the reduced forms for the EEVs. I show that, if we take the conditional independence (CI) assumption seriously, the average structural function – and, therefore, the average partial effects – are identified. Even if we do not fully believe the CI assumption, adding those GRs in the second step may provide reasonable approximations to average partial effects. After all, the score test – where the coefficient on the EEV is zero – is obtained by adding the GRs. Special cases of this approach have been suggested by Petrin and Train (2010) (multinomial response, continuous EEVs), Terza, Basu, and Rathouz (2008, binary response, binary EEV), and Wooldridge (2010, ordered response, binary EEV). Here I provide a unified setting and discuss why the approach is more convincing for continuous EEVs than discrete EEVs, but where the latter might be acceptable – especially in complicated settings.

The paper is organized as follows. In Section 2 I use a standard linear model as motivation by illustrating the robustness of the Gaussian limited information maximum likelihood estimator (LIML). The arguments in the linear case can be extended to nonlinear cases, and Section 3 lays out the simple general approach. Section 4 shows how the general approach can be applied to fractional response variables and nonnegative responses with an exponential mean function, including count responses.

The simple variable addition tests for testing the null that the EEVs are exogenous are derived in Section 5. These tests are easily obtained using standard software, and they motivate the general control function approach in Section 6 for handling endogeneity of continuous and discrete EEVs. Section 7 contains concluding remarks.

2. Motivation: A Linear Model

Consider a population linear model for a response variable y_1 with a single endogenous explanatory variable (EEV), y_2:

$$y_1 = \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{o1} + u_1, \tag{2.1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of a vector $\mathbf{z}$. We assume the vector $\mathbf{z}$ is exogenous in the sense that

$$E(\mathbf{z}' u_1) = \mathbf{0}. \tag{2.2}$$

In practice, $\mathbf{z}_1$ would include a constant, and so we assume that u_1 has a zero mean. We use the convention of putting "o" on the parameters because it is helpful to distinguish the population values from generic values in the parameter space.

The reduced form of y_2 is a linear projection in the population:

$$y_2 = \mathbf{z} \boldsymbol{\delta}_{o2} + v_2 \tag{2.3}$$

$$E(\mathbf{z}' v_2) = \mathbf{0} \tag{2.4}$$

where $\boldsymbol{\delta}_{o2}$ is $L \times 1$. Notice that nothing about the linear projection defined by (2.3) and (2.4) restricts the nature of y_2; it could be a discrete variable, including a binary variable. Also, (2.1) can be viewed as just a linear approximation to a underlying linear model, where (2.2) effectively defines α_{o1} and $\boldsymbol{\delta}_{o1}$.

Provided $E(\mathbf{z}'\mathbf{z})$ is nonsingular and $\boldsymbol{\delta}_{o22} \neq \mathbf{0}$, where $\boldsymbol{\delta}_{o2} = (\boldsymbol{\delta}_{o21}', \boldsymbol{\delta}_{o22}')'$, two stage least squares (2SLS) estimation under random sampling is consistent; see, for example, Wooldridge (2010, Chapter 5). An alternative approach, and one that is convenient for testing the null that y_2 is exogenous, is a control function approach. Write the linear projection of u_1 on v_2, in error form, as

$$u_1 = \gamma_{o1} v_2 + e_1, \tag{2.5}$$

where $\gamma_{o1} = E(v_2 u_1)/E(v_2^2)$ is the population regression coefficient. By construction, $E(v_2 e_1) = 0$ and $E(\mathbf{z}' e_1) = \mathbf{0}$.

If we plug (2.5) into (2.1) we can write

$$y_1 = \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + \gamma_{o1} v_2 + e_1 \tag{2.6}$$

$$E(\mathbf{z}' e_1) = \mathbf{0}, E(v_2 e_1) = 0, E(y_2 e_1) = 0 \tag{2.7}$$

Adding the reduced form error, v_2, to the structural equation “controls” for the endogeneity of y_2. If we could observe data on v_2, we could simply add it as a regressor. Instead, given a random sample of size N, we can estimate $\boldsymbol{\delta}_{o2}$ in a first stage by OLS and obtain the residuals, $\hat{v}_{i2}$, $i = 1, \ldots, N$. In a second stage we run the regression

$$y_{i1} \text{ on } y_{i2}, \mathbf{z}_{i1}, \text{ and } \hat{v}_{i2}, i = 1, \ldots, N. \tag{2.8}$$

The OLS estimators from (2.8) are the control function (CF) estimators. It is well known – for example, Hausman (1978) – that the CF estimates $\hat{\alpha}_1$ and $\hat{\boldsymbol{\delta}}_1$ are *identical* to the 2SLS estimates. Further, the regression-based Hausman test of the null that y_2 is exogenous is a *t* test of $H_0 : \gamma_{o1} = 0$. One may wish to make the test robust to heteroskedasticity, but there is no need to adjust for the first-stage estimation of $\boldsymbol{\delta}_{o2}$ under the null hypothesis. For further discussion, see Wooldridge (2010, Chapter 5).

Rather than use a two-step method, an alternative is to obtain the LIML estimator assuming that (u_1, v_2) is independent of $\mathbf{z}$ and bivariate normal, which implies that (e_1, v_2) is bivariate normal and independent of $\mathbf{z}$. The log likelihood for random draw i (conditional on $\mathbf{z}_i$), multiplied by two, is

$$-\log(\eta_1^2) - [y_{i1} - \alpha_1 y_{i2} - \mathbf{z}_{i1}\boldsymbol{\delta}_1 - \gamma_1(y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)]^2/\eta_1^2 - \log(\tau_2^2) - (y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)^2/\tau_2^2,$$

and the LIML estimators solve

$$\min_{\alpha_1,\boldsymbol{\delta}_1,\gamma_1,\boldsymbol{\delta}_2,\eta_1^2,\tau_2^2} \sum_{i=1}^{N} \{[y_{i1} - \alpha_1 y_{i2} - \mathbf{z}_{i1}\boldsymbol{\delta}_1 - \gamma_1(y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)]^2/\eta_1^2 + (y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)^2/\tau_2^2\} + \log(\eta_1^2) + \log(\tau_2^2)\}.$$

This setup is dubbed “LIML” because $D(y_2|\mathbf{z})$ is an unrestricted reduced from.

For the purposes of this paper, an interesting feature of the LIML estimator is that it is fully robust in the sense that it consistently estimates the parameters in (2.1) and (2.3) under only the zero covariance conditions in (2.2) and (2.4). To see this, write

$$\begin{aligned} y_1 &= \alpha_{o1} y_2 + \mathbf{z}_1\boldsymbol{\delta}_{o1} + \gamma_{o1}(y_2 - \mathbf{z}\boldsymbol{\delta}_{o2}) + e_1 \\ E(\mathbf{z}'e_1) &= \mathbf{0},\ E(y_2 e_1) = 0, \end{aligned}$$

which means that specific nonlinear functions of $(\alpha_{o1}, \boldsymbol{\delta}_{o1}, \gamma_{o1}, \boldsymbol{\delta}_{o2})$ index the linear projection of y_1 on $(y_2, \mathbf{z})$:

$$L(y_1|y_2, \mathbf{z}) = \alpha_{o1} y_2 + \mathbf{z}_1\boldsymbol{\delta}_{o1} + \gamma_{o1}(y_2 - \mathbf{z}\boldsymbol{\delta}_{o2}). \tag{2.9}$$

In addition,

$$L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\delta}_{o2}. \tag{2.10}$$

Together, by the minimum mean square error property of the linear projection, (2.9) and (2.10) imply that the parameters $\alpha_{o1}, \boldsymbol{\delta}_{o1}, \gamma_{o1}$, and $\boldsymbol{\delta}_{o2}$ solve

$$\min_{\alpha_1,\boldsymbol{\delta}_1,\gamma_1,\boldsymbol{\delta}_2} E\{[y_1-\alpha_1 y_2-\mathbf{z}_1\boldsymbol{\delta}_1-\gamma_1(y_2-\mathbf{z}\boldsymbol{\delta}_2)]^2\}+E[(y_2-\mathbf{z}\boldsymbol{\delta}_2)^2]$$

Weighting the expected squared errors by positive constants does not change the solutions. In fact, the first order conditions (FOCs) with respect to $\alpha_1,\ \boldsymbol{\delta}_1,\ \gamma_1,$ and $\boldsymbol{\delta}_2$ are

$$\begin{aligned}
-E\{y_2[y_1-\alpha_1 y_2-\mathbf{z}_1\boldsymbol{\delta}_1-\gamma_1(y_2-\mathbf{z}\boldsymbol{\delta}_2)]/\eta_1^2\} &= 0\\
-E\{\mathbf{z}_1'[y_1-\alpha_1 y_2-\mathbf{z}_1\boldsymbol{\delta}_1-\gamma_1(y_2-\mathbf{z}\boldsymbol{\delta}_2)]/\eta_1^2\} &= \mathbf{0}\\
-E\{(y_2-\mathbf{z}\boldsymbol{\delta}_2)[y_1-\alpha_1 y_2-\mathbf{z}_1\boldsymbol{\delta}_1-\gamma_1(y_2-\mathbf{z}\boldsymbol{\delta}_2)]/\eta_1^2\} &= 0\\
\gamma_1 E\{\mathbf{z}'[y_1-\alpha_1 y_2-\mathbf{z}_1\boldsymbol{\delta}_1-\gamma_1(y_2-\mathbf{z}\boldsymbol{\delta}_2)]/\eta_1^2\}-E[\mathbf{z}'(y_2-\mathbf{z}\boldsymbol{\delta}_2)]/\tau_2^2 &= \mathbf{0},
\end{aligned}$$

and $(\alpha_{o1},\boldsymbol{\delta}_{o1},\gamma_{o1},\boldsymbol{\delta}_{o2})$ solves these (uniquely) by definition of the linear projection. If we define

$$\begin{aligned}
\eta_{o1}^2 &\equiv E\{[y_1-\alpha_{o1}y_2-\mathbf{z}_1\boldsymbol{\delta}_{o1}-\gamma_{o1}(y_2-\mathbf{z}\boldsymbol{\delta}_{o2})]^2\} = E(e_1^2)\\
\tau_{o1}^2 &\equiv E[(y_2-\mathbf{z}\boldsymbol{\delta}_{o2})^2]
\end{aligned}$$

then the FOCs for η_1^2 and τ_2^2 can be written as

$$\begin{aligned}
-\frac{\eta_{o1}^2}{(\eta_1^2)^2}+\frac{1}{\eta_1^2} &= 0\\
-\frac{\tau_{o2}^2}{(\tau_2^2)^2}+\frac{1}{\tau_2^2} &= 0.
\end{aligned}$$

It follows that the solutions are η_{o1}^2 and τ_{o2}^2.

When equation (2.1) is just identified, it is well-known that the IV and LIML estimators are alebraically equivalent – which means, of course, that LIML is just as robust as IV. The argument above – which I believe is original – shows that LIML is just as robust as 2SLS even in the overidentified case.

It is fairly straightforward to extend the previous analysis to a vector $\mathbf{y}_2$ of EEVs. The bottom line is that the Gaussian log likelihood identifies the parameters of a linear model under the same identification condition as 2SLS. We do not need Gaussianity, homoskedasticity, or even linear conditional expectations. Of course, in general we should use robust inference of the kind discussed in White (1982) because the information matrix equality does not hold (for either of the conditional quasi-log likelihoods).

In the next section we argue that the findings for linear models can be extended to certain nonlinear models.

3. A Framework for Quasi-LIML for Nonlinear Models

Suppose that y_1 is a binary response and y_2 is continuous, and consider the model

$$y_1 = 1[\alpha_{o1}y_2 + \mathbf{z}_1\boldsymbol{\delta}_{o1} + u_1 \geq 0] \tag{3.1}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_{o2} + v_2 \tag{3.2}$$

The standard parametric assumptions are that (u_1, v_2) is bivariate normal with mean zero and independent of $\mathbf{z}$. Under normality it can be shown [for example, Wooldridge (2010, Section 15.7.2)] that

$$P(y_1 = 1|y_2, \mathbf{z}) = \Phi\left[\frac{(\alpha_{o1}y_2 + \mathbf{z}_1\boldsymbol{\delta}_{o1} + (\rho_{o1}/\tau_{o2})(y_2 - \mathbf{z}\boldsymbol{\delta}_{o2})}{(1-\rho_{o1}^2)^{1/2}}\right], \tag{3.3}$$

where $\tau_{o2}^2 = Var(v_2)$ and $\rho_{o1} = Corr(v_2, u_1)$. This formula is the basis for the Rivers-Vuong (1988) two-step approach to estimating scaled coefficients in a probit model with a continuous EEV.

We can easily see that the MLE based on $D(y_1|y_2, \mathbf{z})$ and $D(y_2|\mathbf{z})$ has some robustness properties. Suppose we *define* $v_2 = y_2 - \mathbf{z}\boldsymbol{\delta}_{o2}$, where $\boldsymbol{\delta}_{o2}$ is the vector of linear projection parameters, and $\tau_{o2}^2 \equiv E(v_2^2)$. As in Section 2, we know the Gaussian quasi-log-likelihood function identifies these parameters without further assumptions. Then, if we *assume*

$D(u_1|y_2,\mathbf{z}) = D(u_1|v_2)$ – which means that $D(u_1|y_2,\mathbf{z})$ depends on $(y_2,\mathbf{z})$ only through the linear function $y_2 - \mathbf{z}\boldsymbol{\delta}_2$ – and that $D(u_1|v_2)$ has mean linear in v_2 and is homoskedastic normal, the quasi-MLE is consistent even though the full distributional assumptions do not hold. In Section 4 we will show that this finding carries through if y_1 is a fractional response with a conditional mean that has a probit form.

Now consider a more general setup. Let $\boldsymbol{\theta}_{o1},$ and $\boldsymbol{\theta}_{o2}$ be the parameters appearing in the model for some feature of $D(\mathbf{y}_1|\mathbf{y}_2,\mathbf{z})$, where only $\boldsymbol{\theta}_{o2}$ appears in some feature of $D(\mathbf{y}_2|\mathbf{z})$. Let $q_2(\mathbf{y}_2,\mathbf{z},\boldsymbol{\theta}_2)$ and $q_1(\mathbf{y}_1,\mathbf{y}_2,\mathbf{z},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ be objective functions such that $\boldsymbol{\theta}_{o2}$ maximizes $E[q_2(\mathbf{y}_2,\mathbf{z},\boldsymbol{\theta}_2)]$ and $(\boldsymbol{\theta}_{o1},\boldsymbol{\theta}_{o2})$ maximizes $E[q_1(\mathbf{y}_1,\mathbf{y}_2,\mathbf{z},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)]$. Then $(\boldsymbol{\theta}_{o1},\boldsymbol{\theta}_{o2})$ maximizes

$$\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\{E[q_1(\mathbf{y}_1,\mathbf{y}_2,\mathbf{z},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)] + E[q_2(\mathbf{y}_2,\mathbf{z},\boldsymbol{\theta}_2)]\}. \tag{3.4}$$

If we can assume or establish uniqueness of $(\boldsymbol{\theta}_{o1},\boldsymbol{\theta}_{o2})$ – which typically follows under standard identification conditions – it follows that, under standard regularity conditions, the solutions $(\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2)$ to

$$\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2}\sum_{i=1}^{N}[q_1(\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{z}_i,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) + q_2(\mathbf{y}_{i2},\mathbf{z}_i,\boldsymbol{\theta}_2)] \tag{3.5}$$

are generally consistent for $(\boldsymbol{\theta}_{o1},\boldsymbol{\theta}_{o2})$.

In this paper we consider the case where $q_1(\bullet)$ and $q_2(\bullet)$ are quasi-log likelihoods. The challenge is to find interesting cases where quasi-log likelihoods identify the parameters of interest. Of course, in general a two-step procedure – where $\boldsymbol{\theta}_{o2}$ is estimated by $\tilde{\boldsymbol{\theta}}_2$ and then its estimator is plugged into a second step to obtain $\tilde{\boldsymbol{\theta}}_1$ – will also be consistent. The point here is that the one-step estimator that solves (3.5) are generally as robust as a two-step estimator. The one-step estimator makes inference more straightforward and, in some cases, it is more efficient. As in the linear case, the one-step estimator may have better finite-sample properties. Plus, as we will see in Section 4, in some cases there are no convenient two-step estimators yet a joint quasi-LIML is consistent and asymptotically normal.

With smooth objective functions in (3.5), asymptotic analysis follows from standard results on M-estimation [for example, Wooldridge (2010, Chapter 12)]. In general, one needs to use the White (1982) sandwich variance estimator for misspecified maximum likelihood.

In some cases it will happen that, for all outcomes $(\mathbf{y}_2, \mathbf{z})$, $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} E[q_1(\mathbf{y}_1, \mathbf{y}_2, \mathbf{z}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) | \mathbf{y}_2, \mathbf{z}], \tag{3.6}$$

which implies that the scores for $q_1(\cdot)$ and $q_2(\cdot)$ (evaluated at $\boldsymbol{\theta}_{o1}$ and $\boldsymbol{\theta}_{o2}$) are uncorrelated. In that case,

$$Avar\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_{o1}^{-1}\mathbf{B}_{o1}\mathbf{A}_{o1} + \mathbf{A}_{o2}^{-1}\mathbf{B}_{o2}\mathbf{A}_{o2} \tag{3.7}$$

where, for objective functions $g = 1, 2$,

$$\mathbf{A}_{og} = -E[\nabla_{\boldsymbol{\theta}}^2 q_g(\boldsymbol{\theta}_o)] \tag{3.8}$$

$$\mathbf{B}_{og} = E[\nabla_{\boldsymbol{\theta}} q_g(\boldsymbol{\theta}_o)' \nabla_{\boldsymbol{\theta}} q_g(\boldsymbol{\theta}_o)] \tag{3.9}$$

Further simplifications of the componets of the sandwiches are sometimes available on a case-by-case basis.

4. Examples of Quasi-LIMLs

We now show how the setup in Section 3 can be applied to several interesting examples. For notationaly simplicity, we do not use "*o*" to index the true parameters.

4.1. Models for Binary and Fractional Responses

Suppose that y_1 is a variable taking values in the unit interval, $[0,1]$. This includes the case where y_1 is binary but also allows y_1 to be a continuous proportion. y_1 can have both discrete and continuous characteristics (so, for example, y_1 can be a proportion that takes on zero or one with positive probability).

We set up the endogeneity of a covariate as an omitted variable problem, and start by assuming y_2 has a linear reduced form with substantive restrictions:

$$E(y_1|y_2,\mathbf{z},r_1) = E(y_1|y_2,\mathbf{z}_1,r_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1). \tag{4.1}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2, \tag{4.2}$$

where $\mathbf{x}_1$ is a general (nonlinear) function of $(y_2,\mathbf{z}_1)$ and r_1 is an omitted factor thought to be correlated with y_2. The first equality in (4.1) imposes at least one exclusion restriction, where a strict subset $\mathbf{z}_1$ of $\mathbf{z}$ appears in $E(y_1|y_2,\mathbf{z},r_1)$. Because $\mathbf{x}_1$ can be any function of $(y_2,\mathbf{z}_1)$, the setup encompasses the case where y_2 should be replaced with $h_2(y_2)$ for $h_2(\bullet)$ is strictly monotonic. In what follows, we take y_2 to be the function of the EEV so that an additive, independent error v_2 is realistic. In fact, we assume that (r_1,v_2) is independent of $\mathbf{z}$ and jointly normal.

With $r_1 \sim Normal(0,\sigma_{r_1}^2)$ it can be shown that the average structural function is

$$ASF(y_2,\mathbf{z}_1) = E_{r_{i1}}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_{i1})] \ = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{r1}) \tag{4.3}$$

$$\boldsymbol{\beta}_{r1} \equiv \boldsymbol{\beta}_1/(1+\sigma_{r_1}^2)^{1/2}, \tag{4.4}$$

where $\mathbf{x}_1$ now denotes fixed values of the arguments. [See Wooldridge (2010, Section 15.7.2).] Fortunately, we can identify the scaled coefficients $\boldsymbol{\beta}_{r1}$, even though $\boldsymbol{\beta}_1$ and $\sigma_{r_1}^2$ are not separately identified.

There is another useful way to obtain the average structural function using the reduced form error v_2. First, by iterated expectations,

$$ASF(y_2, \mathbf{z}_1) = E_{(y_{i2}, \mathbf{z}_i)}\{E[\Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_{i1})|(y_{i2}, \mathbf{z}_i)], \tag{4.5}$$

which means we first find (for fixed $\mathbf{x}_1$) $E[\Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_{i1})|(y_{i2}, \mathbf{z}_i)]$ and then average out over the distribution of $(y_{i2}, \mathbf{z}_i)$. Wooldridge (2005) shows that, under the maintained assumptions,

$$E_{(y_{i2}, \mathbf{z}_i)}\{E[\Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_{i1})|(y_{i2}, \mathbf{z}_i)] \;\; = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{e1} + \gamma_{e1}v_{i2}) \tag{4.6}$$

where $r_{i1} = \gamma_1 v_{i2} + e_{i1}$, $\boldsymbol{\beta}_{e1} = \boldsymbol{\beta}_1/(1 + \sigma_{e_1}^2)^{1/2}$, $\gamma_{e1} = \gamma_1/(1 + \sigma_{e_1}^2)^{1/2}$. Wooldridge (2005) shows that a two-step procedure consistently estimates the scaled coefficients $\boldsymbol{\beta}_{e1}$:

(i) Regress y_{i2} on $\mathbf{z}_i$ and obtain the residuals, $\hat{v}_{i2}$.

(ii) Use Bernoulli QMLE with the probit response function of y_{i1} on $\mathbf{x}_{i1}, \hat{v}_{i2}$ to estimate $\hat{\boldsymbol{\beta}}_{e1}$ and $\hat{\gamma}_{e1}$. (As a practical matter, this can be implemented using so-called “generalized linear models” (GLM) software.)

(iii) The ASF is consistently estimated as

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{e1} + \hat{\gamma}_{e1}\hat{v}_{i2}),$$

and this can be used to obtain APEs with respect to y_2 or $\mathbf{z}_1$.

Rather than use a two-step approach, a joint quasi-LIML approach can be used to consistently estimate $\boldsymbol{\beta}_{r1}$, $\boldsymbol{\delta}_2$, and $\tau_2^2 = E(v_2^2)$. To see how, first, define a binary variable

$$w_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_1 + r_1 + a_1 \geq 0] \tag{4.7}$$

$$D(a_1|y_2, \mathbf{z}, r_1) = Normal(0, 1), \tag{4.8}$$

and note that

$$E(w_1|y_2, \mathbf{z}, r_1) = E(y_1|y_2, \mathbf{z}, r_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1) \tag{4.9}$$

and so, by iterated expectations,

$$E(y_1|y_2,\mathbf{z}) = E(w_1|y_2,\mathbf{z}). \tag{4.10}$$

Now, write

$$w_1 = 1[\mathbf{x}_1\boldsymbol{\beta}_{r1} + (r_1 + a_1)/(1 + \sigma_{r_1}^2)^{1/2} \geq 0] \tag{4.11}$$

$$\equiv 1[\mathbf{x}_1\boldsymbol{\beta}_{r1} + e_1 \geq 0], \tag{4.12}$$

where $e_1 \equiv (r_1 + a_1)/(1 + \sigma_{r_1}^2)^{1/2}$ has a standard normal distribution and is independent of $\mathbf{z}$. Because r_1 is generally correlated with v_2, e_1 and v_2 are generally correlated; let ρ_1 be the correlation. If we assume joint normality of (r_1, v_2) we have exactly the setup for the Rivers-Vuong model, and so

$$E(w_1|y_2,\mathbf{z}) = \Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_{r1} + (\rho_1/\tau_2)(y_2 - \mathbf{z}\boldsymbol{\delta}_2)}{(1-\rho_1^2)^{1/2}}\right] = E(y_1|y_2,\mathbf{z}).$$

What we have shown is that the mean $E(y_1|y_2,\mathbf{z})$ has the exact same form as probit with a continuous EEV. Because the Bernoulli log likelihood is in the linear exponential family, it identifies the parameters in a correctly specifield conditional mean. So we can take

$$q_1(y_1,y_2,\mathbf{z},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = (1-y_1)\left\{1 - \Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_{r1} + (\rho_1/\tau_2)(y_2 - \mathbf{z}\boldsymbol{\delta}_2)}{(1-\rho_1^2)^{1/2}}\right]\right\} + y_1\Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_{r1} + (\rho_1/\tau_2)(y_2 - \mathbf{z}\boldsymbol{\delta}_2)}{(1-\rho_1^2)^{1/2}}\right]$$

and then

$$q_2(y_2,\mathbf{z},\boldsymbol{\theta}_2) = -\log(\tau_2^2)/2 - (y_2 - \mathbf{z}\boldsymbol{\delta}_2)^2/(2\tau_2^2).$$

When we combine $q_1(\cdot)$ and $q_2(\cdot)$, we obtain the usual Rivers-Vuong log likelihood, which is programmed in popular statistical packages. However, we must recognize that y_1 is fractional, and so we are using quasi-MLE. Generally, a fully robust sandwich variance matrix estimator must be used for inference.

We can also allow more flexibility in $D(y_2|\mathbf{z})$ by allowing, say, $Var(y_2|\mathbf{z}) = \exp(\mathbf{z}\boldsymbol{\xi}_2)$, and then using the Gaussian quasi-log likelihood for $D(y_2|\mathbf{z})$ with linear mean and variance $\exp(\mathbf{z}\boldsymbol{\xi}_2)$. Then, we can assume $D(r_1|y_2,\mathbf{z})$ depends only on the standardized error, $(y_2 - \mathbf{z}\boldsymbol{\delta}_2)/\exp(\mathbf{z}\boldsymbol{\xi}_2/2)$.

A similar argument holds when y_2 is binary and follows a probit model:

$$y_2 = 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0] \tag{4.13}$$

$$v_2|\mathbf{z} \sim Normal(0,1) \tag{4.14}$$

If w_1 is defined as in (4.7) and (4.8) we still have the key result in (4.10). Further, from the bivariate probit model,

$$E(w_1|y_2 = 1,\mathbf{z}) = \int_{-\mathbf{z}\boldsymbol{\delta}_2}^{\infty} \Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_{r1} + \rho_1\mathfrak{v}_2}{(1-\rho_1^2)^{1/2}}\right]d\mathfrak{v}_2 = E(y_1|y_2 = 1,\mathbf{z}),$$

and a similar expression holds for $E(y_1|y_2 = 0,\mathbf{z})$. Therefore, for $q_2(y_2,\mathbf{z},\boldsymbol{\theta}_2)$ we use the usual probit log-likelihood and for $q_1(y_1,y_2,\mathbf{z},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ we use the Bernoulli quasi-log likelihood associated with bivariate probit. As a practical matter, bivariate probit software simply needs to allow fractional y_1 and robust inference.

When y_2 is binary, allowing $\mathbf{x}_1$ to be a general function of y_2 and $\mathbf{z}_1$ allows a full set of interactions among y_2 and the exogenous variables $\mathbf{z}_1$. A full switching regression model for fractional responses, where a different source of omitted variables is allowed under the two regimes, is also easily estimated using a standard Bernoulli log likelihood. In this case, we estimate (scaled) coefficients – say $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ – by applying a Heckman self-selection correction to "probit" models for $y_2 = 0$ and $y_2 = 1$. (That is, we again act as if y_1 is binary even though it is fractional.) The argument for why this works is essentially the same as the single regime case. When y_2 is a program indicator, the average treatment effect of the program is estimated as

$$\hat{\tau}_{ate} = N^{-1}\sum_{i=1}^{N}[\Phi(\mathbf{z}_{i1}\hat{\boldsymbol{\beta}}_1) - \Phi(\mathbf{z}_{i1}\hat{\boldsymbol{\beta}}_0)],$$

where the $\mathbf{z}_{i1}$ are the exogenous covariates in the model for y_{i1}, $\hat{\boldsymbol{\beta}}_0$ is obtained using the $y_{i2} = 0$ subsample, and $\hat{\boldsymbol{\beta}}_1$ is obtained using the $y_{i2} = 1$ subsample.

4.2. Exponential Models

For nonnegative responses y_1, including but not restricted to count variables, an omitted variables formulation is

$$E(y_1|y_2,\mathbf{z},r_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1), \tag{4.15}$$

where, again, $\mathbf{x}_1$ contains an intercept and can be any function of $(y_2,\mathbf{z}_1)$. Consider the case where y_2 is binary, as in (4.13) and (4.14), and strengthen the assumptions so that (r_1,v_2) is independent of $\mathbf{z}$ and bivariate normal, with mean zero, $Var(v_2) = 1$, $Var(r_1) = \tau_1^2$, and $Corr(v_2,r_1) = \rho_1$.

Following Terza (1998), it can be shown that

$$E(y_1|y_2,\mathbf{z}) = \exp(\tau_1^2/2 + \mathbf{x}_1\boldsymbol{\beta}_1)\{\Phi(\rho_1 + \mathbf{z}\boldsymbol{\delta}_2)/\Phi(\mathbf{z}\boldsymbol{\delta}_2)\}^{y_2}\{[1-\Phi(\rho_1+\mathbf{z}\boldsymbol{\delta}_2)]/[1-\Phi(\mathbf{z}\boldsymbol{\delta}_2)]\}^{(1-y_2)} \tag{4.16}$$

Because $x_{11} = 1$, only $\tau_1^2/2 + \beta_{11}$ is identified. It is easily seen that this is exactly the intercept that appears in the APEs – see, for example, Terza (2009) and Wooldridge (2010, Chapter 18) – so we just absorb $\tau_1^2/2$ into the intercept β_{11}.

Terza (1998) proposed a two-step nonlinear least squares method, but we can use a quasi-LIML estimator, too. We simply combine the probit log likelihood for $D(y_2|\mathbf{z})$ with, say, the Poisson quasi-log likelihood with conditional mean (4.16). Because the Poisson distribution is a member of the linear exponential family, the discussion from Section 3 shows that we only need the probit model for y_2 to be correctly specified and $E(y_1|y_2,\mathbf{z})$ to have the form (4.16). Computationally, the joint estimator may pose some challenges, but in cases with overidentification or weak instruments it may have better finite-sample properties. In addition, under the null hypothesis $\rho_1 = 0$, the Poisson QMLE of $\boldsymbol{\beta}_1$ has some efficiency properties: it is the asymptotically efficient estimator among estimators that use only the conditional mean assumption if $Var(y_1|y_2,\mathbf{z}_1) = \sigma_1^2 E(y_1|y_2,\mathbf{z}_1)$. Thus, the estimator is more efficient than nonlinear least squares under standard count distributions when $\rho_1 = 0$, and that is likely to be true for nonzero ρ_1, too.

If y_2 is continuous, $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$, and $r_1 = \rho_1 v_2 + e_1$ with $Var(e_1) = \tau_1^2 - \rho_1^2\tau_2^2$, then

$$\begin{aligned} E(y_1|\mathbf{z},y_2) &= E[\exp(e_1)]\exp(\mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2) \\ &= \exp((\tau_1^2 - \rho_1^2\tau_2^2)/2)\exp(\mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2). \end{aligned}$$

As before, the intercept we want to estimate is $\tau_1^2/2 + \beta_{11}$, so we just absorb $\tau_1^2/2$ into β_{11}. Then

$$E(y_1|\mathbf{z},y_2) = \exp(-\rho_1^2\tau_2^2/2 + \mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1(y_2 - \mathbf{z}\boldsymbol{\delta}_2)). \tag{4.17}$$

We can use the Gaussian log likelihood for $D(y_2|\mathbf{z})$, which depends on $\boldsymbol{\delta}_2$ and τ_2^2, along with the Poisson quasi-log likelihood with mean given by (4.17). If $Var(y_2|\mathbf{z}) = \exp(\mathbf{z}\boldsymbol{\xi}_2)$ we can replace τ_2^2 with $\exp(\mathbf{z}\boldsymbol{\xi}_2)$ and v_2 with $(y_2 - \mathbf{z}\boldsymbol{\delta}_2)/\exp(\mathbf{z}\boldsymbol{\xi}_2/2)$.

5. Variable Addition Tests for Endogeneity

The approach we take to testing the null that an EEV is exogenous is similar to Vella (1993), who uses a maximum likelihood framework. In cases where we have entirely specified $D(y_1|y_2,\mathbf{z})$ and $D(y_2|\mathbf{z})$, the approach here reduces to Vella's test in several instances.

To motivate variable addition tests (VATs) when we have not fully specified conditional distributions, consider the case where y_2 is binary and

$$E(y_1|y_2,\mathbf{z}) = E\{\Phi[(1-\rho_1^2)^{-1/2}(\mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2)]|y_2,\mathbf{z}\} \equiv m(\mathbf{x}_1\boldsymbol{\beta}_1,\rho_1,\boldsymbol{\delta}_2) \tag{5.1}$$

where, to simplify notation, we use $\boldsymbol{\beta}_1$ to denote the scaled coefficients that index the ASF, and $\boldsymbol{\delta}_2$ are the parameters in the probit model for y_2. What we need to obtain the score test are the derivatives of $m(\mathbf{x}_1\boldsymbol{\beta}_1,\rho_1,\boldsymbol{\delta}_2)$ with respect to $(\boldsymbol{\beta}_1,\rho_1)$ evaluated at $\rho_1 = 0$. (Under the null hypothesis, $\boldsymbol{\beta}_1$ and $\boldsymbol{\delta}_2$ are estimated using separate procedures.) Taking the derivatives through the integral, it is easily seen that

$$\frac{\partial m(\mathbf{x}_1\boldsymbol{\beta}_1,0,\boldsymbol{\delta}_2)}{\partial\boldsymbol{\beta}_1} = \phi(\mathbf{x}_1\boldsymbol{\beta}_1)\mathbf{x}_1 \tag{5.2}$$

$$\frac{\partial m(\mathbf{x}_1\boldsymbol{\beta}_1,0,\boldsymbol{\delta}_2)}{\partial\rho_1} = E[\phi(\mathbf{x}_1\boldsymbol{\beta}_1)v_2|y_2,\mathbf{z}] = \phi(\mathbf{x}_1\boldsymbol{\beta}_1)E(v_2|y_2,\mathbf{z}) \equiv \phi(\mathbf{x}_1\boldsymbol{\beta}_1)gr_2, \tag{5.3}$$

where

$$gr_2 \equiv E(v_2|y_2,\mathbf{z}) \tag{5.4}$$

is a population generalized residual (PGR) – see, for example, Gourieoux, Monfort, Renault, and Trognon (1987) (GMT). Notice that gr_2 depends on the population parameters. (GMT call gr_2 “generalized error.”) As is well known, the PGR when y_2 follows a probit model is related to the inverse Mills ratio, $\lambda(\cdot) \ = \phi(\cdot)/\Phi(\cdot)$:

$$gr_2 = y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1-y_2)\lambda(-\mathbf{z}_i\boldsymbol{\delta}_2). \tag{5.5}$$

When we plug in the probit estimates, $\hat{\boldsymbol{\delta}}_2$, for each observation i we get a set of N (sample) generalized residuals,

$$\widehat{gr}_{i2} = y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1-y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2). \tag{5.6}$$

Given the structure of the score of the mean function in (5.2) and (5.3), the score test is easily seen to be asymptotically equivalent to a simple omitted variables test of test of gr_2. In other words, we can test $H_0 : \eta_1 = 0$ in the auxiliary equation

$$\text{“}E(y_{i1}|y_{i2}, \mathbf{z}_i) = \Phi(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \eta_1 gr_{i2}),\text{”} \tag{5.7}$$

where we replace $\boldsymbol{\delta}_2$ with the probit estimates. So we can obtain a Wald (robust t) test of $H_0 : \eta_1 = 0$ using the conditional mean function

$$\Phi(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \eta_1 \widehat{gr}_{i2}) \tag{5.8}$$

in a (fractional) probit estimation. We need to make the t statistic robust when y_1 is a fractional response because the implicit variance in the Bernoulli quasi-log likelihood is incorrect . We can use a nonrobust test if y_1 is binary and the probit model for y_1 is correctly specified under H_0. In any case we need not make an adjustment for estimation of $\boldsymbol{\delta}_2$.

When y_2 is continuous with reduced form $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$ where v_2 is independent of $\mathbf{z}$, $gr_2 = v_2$, and then the VAT for the null that y_2 is exogenous is obtained exactly as in (5.8) except that $\widehat{gr}_{i2} = \hat{v}_{i2}$ are just the OLS residuals from y_{i2} on $\mathbf{z}_i$. This is easily derived from the conditional mean function

$$E(y_1|y_2, \mathbf{z}) = \Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_1 + (\rho_1/\tau_2)(y_2 - \mathbf{z}\boldsymbol{\delta}_2)}{(1-\rho_1^2)^{1/2}}\right],$$

obtaining the derivatives with respect to $(\boldsymbol{\beta}_1, \rho_1)$, and then evaluating at $\rho_1 = 0$. Wooldridge (2005) proposed this VAT in the context of control function estimation with fractional response variables.

A general setting for index functions in the context of the linear exponential family starts with

$$y_2 = H(\mathbf{z}, v_2) \tag{5.9}$$

where v_2 is independent $\mathbf{z}$. Further, assume that

$$E(y_1|\mathbf{z}, v_2) = G[\mathbf{x}_1\mathbf{c}(\boldsymbol{\beta}_1, \rho_1) + a(\rho_1, \boldsymbol{\theta}_2)v_2], \tag{5.10}$$

where $\mathbf{x}_1$ is $1 \times K_1$, $G(\bullet)$ is a continuously differentiable function with derivative $g(\bullet)$, and $\mathbf{c}(\boldsymbol{\beta}_1, \rho_1)$ and $a(\rho_1, \boldsymbol{\theta}_2)$ are known functions of the parameters such that

$$\begin{aligned} \mathbf{c}(\boldsymbol{\beta}_1, 0) &= \boldsymbol{\beta}_1 \\ \frac{\partial \mathbf{c}(\boldsymbol{\beta}_1, 0)}{\partial \rho_1} &= \mathbf{0} \\ \frac{\partial \mathbf{c}(\boldsymbol{\beta}_1, 0)}{\partial \boldsymbol{\beta}_1} &= \mathbf{I}_{K_1} \\ a(0, \boldsymbol{\theta}_2) &= 0 \\ \frac{\partial a(0, \boldsymbol{\theta}_2)}{\partial \rho_1} &\equiv \omega_1 \neq 0 \end{aligned}$$

Under the null hypothesis $H_0 : \rho_1 = 0$,

$$E(y_1|y_2, \mathbf{z}) = E(y_1|y_2, \mathbf{z}_1) = G(\mathbf{x}_1\boldsymbol{\beta}_1), \tag{5.11}$$

where $G(\mathbf{x}_1\boldsymbol{\beta}_1)$ is the average structural function under H_0. The test is based on the mean function under the alternative,

$$E(y_1|y_2, \mathbf{z}) = m(\mathbf{x}_1, \boldsymbol{\beta}_1, \rho_1, \boldsymbol{\theta}_2) = E\{G[\mathbf{x}_1\mathbf{c}(\boldsymbol{\beta}_1, \rho_1) + a(\rho_1, \boldsymbol{\theta}_2)v_2]|y_2, \mathbf{z}\}.$$

The derivatives of $m(\mathbf{x}_1, \boldsymbol{\beta}_1, \rho_1, \boldsymbol{\theta}_2)$ with respect to $(\boldsymbol{\beta}_1, \rho_1)$, evaluated at $\rho_1 = 0$, are

$$\frac{\partial m(\mathbf{x}_1, \boldsymbol{\beta}_1, 0, \boldsymbol{\theta}_2)}{\partial \boldsymbol{\beta}_1} = g(\mathbf{x}_1\boldsymbol{\beta}_1)\mathbf{x}_1 \tag{5.12}$$

$$\frac{\partial m(\mathbf{x}_1, \boldsymbol{\beta}_1, 0, \boldsymbol{\theta}_2)}{\partial \rho_1} = E[g(\mathbf{x}_1\boldsymbol{\beta}_1)v_2|y_2, \mathbf{z}] = g(\mathbf{x}_1\boldsymbol{\beta}_1)\omega_1 gr_2, \tag{5.13}$$

where, again, $gr_2 = E(v_2|y_2, \mathbf{z})$ is the population generalized residual. Using the same reasoning as before, we can apply the underlying quasi-MLE for the chosen LEF to the mean function

$$G(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \eta_1 \widehat{gr}_{i2}) \tag{5.14}$$

and use a robust t test of $H_0 : \eta_1 = 0$. The results from Wooldridge (2010, Section 12.4) on two-step M-estimators can be applied to show that replacing $\boldsymbol{\theta}_2$ with $\hat{\boldsymbol{\theta}}_2$ (the M-estimator in the model for y_2) does not affect the limiting distribution under H_0.

There are many ways to extend the previous approach. For example, if y_2 is continuous but heteroskedasticity, we can estimate the moments

$$\begin{aligned} E(y_2|\mathbf{z}) &= \mathbf{z}\boldsymbol{\delta}_2 \\ Var(y_2|\mathbf{z}) &= \exp(\mathbf{z}\boldsymbol{\xi}_2) \end{aligned}$$

using, say, the Gaussian quasi-log likelihood. Then the generalized residual would then be

$$\widehat{gr}_{i2} = \frac{(y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_2)}{\exp(\mathbf{z}_i\hat{\boldsymbol{\xi}}_2/2)}, \tag{5.15}$$

and this can be used in a variable addition test.

If y_2 is a corner solution response following at standard Tobit model, the generalized residual is

$$\widehat{gr}_{i2} = -\hat{\tau}_2 1[y_{i2} = 0]\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) + 1[y_{i2} > 0](y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\delta}}_2), \tag{5.16}$$

where $(\hat{\boldsymbol{\delta}}_2, \hat{\tau}_2^2)$ are the Tobit MLEs. These GRs can be added to a fractional probit estimation, for example, and a simple t test computed.

For an exponential response with y_2 binary,we have $E(y_1|y_2,\mathbf{z})$ in closed form, so we need the derivative of

$$E(y_1|y_2,\mathbf{z}) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1)\{\Phi(\rho_1 + \mathbf{z}\boldsymbol{\delta}_2)/\Phi(\mathbf{z}\boldsymbol{\delta}_2)\}^{y_2}\{[1 - \Phi(\rho_1 + \mathbf{z}\boldsymbol{\delta}_2)]/[1 - \Phi(\mathbf{z}\boldsymbol{\delta}_2)]\}^{(1-y_2)}$$

with respect to ρ_1 evaluated at $\rho_1 = 0$. Computing the derivatives separately for $y_2 = 1$ and

$y_2 = 0$, and then combining terms, we have

$$\frac{\partial m(\mathbf{x}_1, \boldsymbol{\beta}_1, 0, \boldsymbol{\delta}_2)}{\partial \rho_1} = \exp(\mathbf{x}_1\boldsymbol{\beta}_1)[y_2\lambda(\mathbf{z}\boldsymbol{\delta}_2) - (1 - y_2)\lambda(-\mathbf{z}\boldsymbol{\delta}_2)],$$

an expression that follows from the general treatment above. Therefore, after obtaining $\widehat{gr}_{i2}$ from (5.6), we add $\widehat{gr}_{i2}$ along with $\mathbf{x}_{i1}$ to, say, a Poisson QMLE analysis with an exponential mean and compute a robust *t* statistic on $\widehat{gr}_{i2}$. Remember, this allows $\mathbf{x}_{i1}$ to consist of a full set of interactions, $\mathbf{x}_{i1} = (\mathbf{z}_{i1}, y_{i2}\mathbf{z}_{i1})$.

Other extensions to this test may be useful when one suspects a large degree of heterogeneity in an underlying model. If, for example, we start with

$$E(y_1|y_2, \mathbf{z}, a_1, \mathbf{d}_1) = G(a_1y_2 + \mathbf{z}_1\mathbf{d}_1), \qquad (5.17)$$

where $(a_1, \mathbf{d}_1)$ are random coefficients, independent of $\mathbf{z}$, and multivariate normal, the VAT would be to use a quasi-MLE applied to the mean function

$$G(\alpha_1y_{i2} + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \eta_1\widehat{gr}_{i2}y_{i2} + \widehat{gr}_{i2}\mathbf{z}_{i1}\boldsymbol{\psi}_1) \qquad (5.18)$$

and use a joint, robust Wald test of $H_0 : \boldsymbol{\psi}_1 = \mathbf{0}, \eta_1 = 0$. Recall that $\mathbf{z}_{i1}$ includes an intercept, so $\widehat{gr}_{i2}$ appears by itself in this equation and also interacted with the endogenous and exogenous explanatory variables. If y_1 is a binary or fractional response, we can use the probit response function and Bernoulli (quasi-) MLE. If y_1 is nonnegative, such as a count variable, we can use the Poisson quasi-MLE.

In applying the specification tests proposed in this section, a few practical points are worth remembering. First, the only assumption being used is that, under H_0, $E(y_1|y_2, \mathbf{z}) = E(y_1|y_2, \mathbf{z}_1) = G(\mathbf{x}_1\boldsymbol{\beta}_1)$. We do not need any model for y_2 to be correctly specified under H_0. In fact, if y_2 is binary we could, instead of using the generalized residuals obtained from probit, use the OLS residuals obtained from a linear probability model and still obtain a valid test under the null. The reason for preferring a test based on the GRs is that the test is optimal (has highest local asymptotic power) under correct specification of the probit model for y_2. Second, as in any specification testing context, a rejection of the null may occur for many reasons. The variable y_2 may be endogenous, but it could also be that the conditional mean $E(y_1|y_2,\mathbf{z}_1)$ is misspecified. Third, by following the approach proposed in this section, the tests will not reject due to misspecifications of $D(y_1|y_2,\mathbf{z})$ other than $E(y_1|y_2,\mathbf{z}_1)$. Thus, the tests are robust because no auxiliary assumptions are imposed under the null.

6. A General Control Function Approach

The setup in Section 3, illustrated in Section 4, allows for joint and one-step QMLE in a variety of situations, but these methods can be difficult to apply with certain discrete response models for y_1 or discrete EEVs, or both, particularly if we have more than one EEV. Even slight extensions of standard models are difficult to handle if we are wedded to starting with a “structural” model for y_1 and then trying to obtain full MLEs or two-step estimators.

As an example, consider a probit response function with a binary EEV, but were the latter interacts with unobserved heterogeneity:

$$\begin{aligned} E(y_1|y_2,\mathbf{z}_1,a_1,\mathbf{d}_1) &= \Phi(a_1y_2+\mathbf{z}_1\mathbf{d}_1) && (6.1)\\ &= \Phi(\alpha_1y_2+\mathbf{z}_1\boldsymbol{\delta}_1+c_1y_2+\mathbf{z}_1\mathbf{q}_1) && (6.2)\\ y_2 &= 1[\mathbf{z}\boldsymbol{\delta}_2+v_2>0] && (6.3)\end{aligned}$$

where $a_1=\alpha_1+c_1$ and $\mathbf{d}_1=\boldsymbol{\delta}_1+\mathbf{q}_1$. Now, if $(a_1,\mathbf{q}_1,v_2)$ is multivariate normal we could use a joint QMLE by finding $E(y_1|y_2,\mathbf{z})$. But the expectation is not in closed form and the resulting procedure would be computationally intensive.

An alternative approach is suggested by the VATs derived in Section 5 combined with the insights of Blundell and Powell (2003, 2004) and Wooldridge (2005). To describe the approach, we need to review Blundell and Powell (2004) and the slight extension due to Wooldridge (2005). BP study a fully nonparametric situation where

$$y_1 = g_1(y_2, \mathbf{z}_1, u_1) \tag{6.4}$$

for unobservables u_1. The average structural function is

$$ASF(y_2, \mathbf{z}_1) \equiv E_{u_{i1}}[g_1(y_2, \mathbf{z}_1, u_{i1})], \tag{6.5}$$

so that the unobservables are averaged out. Further, BP assume that y_2 (a scalar here for simplicity) has the representation

$$y_2 = g_2(\mathbf{z}) + v_2, \tag{6.6}$$

where (u_1, v_2) is independent of $\mathbf{z}$. Under independence of (u_1, v_2) and the representation $y_2 = g_2(\mathbf{z}) + v_2$,

$$D(u_1|y_2, \mathbf{z}) = D(u_1|v_2). \tag{6.7}$$

Further, as shown by BP (2004), the ASF can be obtained from

$$h_1(y_2, \mathbf{z}_1, v_2) \equiv E(y_1|y_2, \mathbf{z}_1, v_2). \tag{6.8}$$

In particular,

$$ASF(y_2, \mathbf{z}_1) = E_{v_{i2}}[h_1(y_2, \mathbf{z}_1, v_{i2})].$$

Unlike the u_{i1}, for identification purposes we effectively observe the v_{i2} because $v_{i2} = y_{i2} - g_2(\mathbf{z}_i)$, and $g_2(\cdot)$ is nonparametrically identified. (Of course, we can also model $g_2(\cdot)$ parametrically and use standard $\sqrt{N}$-asymptotically normal estimators.) Letting

$$\hat{v}_{i2} = y_{i2} - \hat{g}_2(\mathbf{z}_i) \tag{6.9}$$

denote the reduced form residuals, a consistent estimator of the ASF, under weak regularity conditions, is

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} \hat{h}_1(y_2, \mathbf{z}_1, \hat{v}_{i2}). \tag{6.10}$$

The BP (2004) framework is very general when it comes to allowing flexibility in $g_1(\bullet)$ and $g_2(\bullet)$; in effect, an exclusion restriction is needed in the former and the latter must depend on at least one excluded exogenous variable. Even if one wants to stay within a parametric framework, the BP approach is liberating because it shows that a quantity of considerable interest – the ASF – can be obtained from $E(y_1|y_2, \mathbf{z}_1, v_2)$ without worrying about the structural function $g_1(\bullet)$. In a parametric setting this means that, once $E(y_2|\mathbf{z})$ is modeled and estimated, attention can turn to $E(y_1|y_2, \mathbf{z}_1, v_2)$ or possibly $D(y_1|y_2, \mathbf{z}_1, v_2)$.

Directly modeling $D(y_1|y_2, \mathbf{z}_1, v_2)$ is the approach taken by Petrin and Train (2010) when y_1 is a multinomial response (product choice) and y_2 is replaced with a vector of prices. Starting with standard models for $D(y_1|\mathbf{y}_2, \mathbf{z}_1, \mathbf{u}_1)$ – such as multinomial logit or nested logit – where $\mathbf{u}_1$ includes heterogeneous tastes, leads to complicated estimators. Petrin and Train suggest modeling $D(y_1|\mathbf{y}_2, \mathbf{z}_1, \mathbf{v}_2)$ directly, where $\mathbf{v}_2$ is a vector of reduced form errors: $\mathbf{y}_2 = \mathbf{G}_2(\mathbf{z}) + \mathbf{v}_2$. Given a linear reduced form for $\mathbf{y}_2$, the two-step estimation method is very simple, because the second step is multinomial logit, nested logit, or a mixed logit model.

When the EEVs are continuous, approaches such as that proposed by Petrin and Train (2010) can be viewed as convenient parametric approximations to an analysis that could be made fully nonparametric (subject to practical issues such as number of observations relative to the dimension of the explanatory variables). Unfortunately, when y_2 is discrete, standard models for $D(y_2|\mathbf{z})$ along with structural models for $D(y_1|y_2, \mathbf{z}, \mathbf{u}_1)$, do not generally lead to simple CF estimation. Moreover, models with discrete EEVs are generally nonparametrically identified (for example, Chesher, 2003). Therefore, if we want point estimates of average partial effects when y_2 is discrete, we must rely on parametric assumptions.

As we saw in Section 4, for a wide class of nonlinear models adding the generalized residual produces a test for the null that y_2 is exogenous. What if, as a general strategy, we use generalized residuals as control functions in parametric nonlinear models with the hope that this (largely) solves the endogeneity problem?

It is useful to determine assumptions under which a two-step control function method can produce consistent estimators when the EEVs are discrete. For simplicity take y_2 to be a scalar, and first assume

$$E(y_1|y_2,\mathbf{z},\mathbf{r}_1) = E(y_1|y_2,\mathbf{z}_1,\mathbf{r}_1), \tag{6.11}$$

which imposes the exclusion restriction conditional on heterogeneity $\mathbf{r}_1$. Notice that this condition generalizes the BP approach because it allows for additional unobservables without taking a stand on the exact nature of those unobservables – they could have discreteness, for example. This extension is important for handling models such as fractional responses or nonnegative responses because it is more natural to specify, say, $E(y_1|y_2,\mathbf{z}_1,r_1) = \Phi(\alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + r_1)$ then to write y_1 as a deterministic function of a larger set of unobservables.

Next, let $e_2 = k_2(y_2,\mathbf{z})$ be the proposed control function for some function $k_2(\cdot)$. Under (6.11),

$$E(y_1|y_2,\mathbf{z}_1,\mathbf{r}_1,e_2) = E(y_1|y_2,\mathbf{z}_1,\mathbf{r}_1),$$

so that e_2 is properly excluded from the structural conditional expectation. Further, a key restriction, following BP and Wooldridge (2005), is

$$D(\mathbf{r}_1|y_2,\mathbf{z}) = D(\mathbf{r}_1|e_2). \tag{6.12}$$

In other words, e_2 acts as a kind of sufficient statistic for characterizing the endogeneity of y_2. In the BP setup, $e_2 \equiv v_2 = y_2 - g_2(\mathbf{z})$. In the Heckman linear switching regression framework, $e_2 = gr_2$ suffices, where gr_2 is the generalized residual.

In general, we can verify (6.12) by starting with a generalization of the BP setup by relaxing additivity of v_2:

$$y_2 = \mathbf{g}_2(\mathbf{z},v_2) \tag{6.13}$$

Then, we assume two conditions that imply (6.12):

$$D(\mathbf{r}_1|\mathbf{z},v_2) = D(\mathbf{r}_1|v_2) \tag{6.14}$$

$$D(v_2|y_2,\mathbf{z}) = D(v_2|e_2) \tag{6.15}$$

Condition (6.14) is standard, as it is implied by $(\mathbf{r}_1,v_2)$ independent of $\mathbf{z}$. Condition (6.15) can be shown in some cases where e_2 includes generalized residuals – as in the binary response case, for example.

If we maintain (6.11) and (6.12) then it follows from Wooldridge (2010, Section 2.2.5) that the ASF can be obtained as

$$ASF(y_2,\mathbf{z}_1) = E_{e_{i2}}[h_2(y_2,\mathbf{z}_1,e_{i2})], \tag{6.16}$$

where

$$h_2(y_2,\mathbf{z}_1,e_2) = E(y_1|y_2,\mathbf{z}_1,e_2). \tag{6.17}$$

Asserting that (6.12) holds for discrete y_2 has precedence, although it is typically imposed indirectly. For example, Terza, Basu, and Rathouz (2008) (TBR) effectively use this assumption when $e_2 = y_2 - \Phi(\mathbf{z}\boldsymbol{\delta}_2)$, where y_2 is binary and follows a probit model. In fact, for binary y_1, TBR suppose a parametric model,

$$\begin{aligned} y_1 &= 1[\alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\delta}_1 + u_1 > 0] \\ u_1 &= \rho_1 e_2 + a_1 \\ a_1|(y_2,\mathbf{z}) &\sim Normal(0,\tau_1^2) \end{aligned}$$

[Burnett (1997) actually proposed this approach but without any justification.] Given that the score test uses the generalized residual, and that $E(u_1|y_2,\mathbf{z})$ is linear in the generalized residual (not $e_2 = y_2 - \Phi(\mathbf{z}\boldsymbol{\delta}_2)$), it seems slightly preferred to use the generalized residuals as e_2. It is important to remember that neither can be justified using the usual assumptions for the bivariate probit model and neither is more or less general than the usual bivariate probit assumptions.

Generally, my suggestion is to use convenient parametric models maintaining the key condition (6.12) for an appropriately chosen function e_2, typically a generalized residual. Then parametric models can be applied to estimate the conditional mean $E(y_1|y_2,\mathbf{z}_1,e_2)$. In some cases, we may actually specify a full conditional distribution, $D(y_1|y_2,\mathbf{z}_1,e_2)$ for example, if y_1 is a binary, multinomial, or ordered response. The general method is as follows, assuming a random sample of size N from the population:

1. Estimate a model for $D(y_2|\mathbf{z})$ [or sometimes only for $E(y_2|\mathbf{z})$], where the model depends on parameters $\boldsymbol{\theta}_2$. For the function $e_{i2} = k_2(y_{i2},\mathbf{z}_i,\boldsymbol{\theta}_2)$, define generalized residuals as

$$\hat{e}_{i2} = k_2(y_{i2},\mathbf{z}_i,\hat{\boldsymbol{\theta}}_2). \tag{6.18}$$

2. Estimate a parametric model for $E(y_1|y_2,\mathbf{z}_1,e_2)$ using a quasi-MLE by inserting $\hat{e}_{i2}$ for e_{i2}. Or, if $D(y_1|y_2,\mathbf{z}_1,e_2)$ has been fully specified, use MLE. In either case, let the parameter estimator be $\hat{\boldsymbol{\theta}}_1$.

3. Estimate the ASF as

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} h_1(y_2, \mathbf{z}_1, \hat{e}_{i2}, \hat{\boldsymbol{\theta}}_1) \tag{6.19}$$

where $h_1(y_2, \mathbf{z}_1, e_2, \boldsymbol{\theta}_1) = E(y_1|y_2, \mathbf{z}_1, e_2)$.

Inference concerning $\widehat{ASF}$ can be obtained using the delta method – the particular form is described in Wooldridge (2010, Problem 12.17) – or bootstrapping the two estimation steps.

How might we apply the general CF approach to the problem described in equations (6.1), (6.2), and (6.3)? First, we would not specify (6.1) as the structural conditional mean, but we would assume that (6.12) holds for $e_2 = gr_2$ and use a mean function such as

$$E(y_{i1}|y_{i2}, \mathbf{z}_{i1}, gr_{i2}) = \Phi(\alpha_1 y_{i2} + \mathbf{z}_{i1}\boldsymbol{\delta}_1 + \eta_1 gr_{i2} \cdot y_{i2} + gr_{i2} \cdot \mathbf{z}_{i1}\boldsymbol{\psi}_1). \tag{6.20}$$

In other words, we take a standard functional form that restricts the mean function to the unit interval – in this case the probit function – and add the control function in a fairly flexible way. We get a simple test for the null of exogeneity and, hopefully, a reasonable approximation to the ASF when we average out $\widehat{gr}_{i2}$:

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} \Phi(\hat{\alpha}_1 y_2 + \mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\eta}_1 \widehat{gr}_{i2} \cdot y_2 + \widehat{gr}_{i2} \cdot \mathbf{z}_1\hat{\boldsymbol{\psi}}_1). \tag{6.21}$$

A similar strategy is available if y_2 is a corner solution and follows a Tobit model. In this case, the generalized residual is given in equation (5.16).

An approach based on (6.20) is neither more nor less general than an approach that starts by specifying $D(y_1|y_2, \mathbf{z}_1, \mathbf{u}_1)$ and parametric assumptions in (6.13). While a more structural approach may have more appeal conceptually, it is not nearly as simple as the control function approach based on (6.20). If we are interested in the average structural function, (6.20) is more direct.

The drawback to the CF approach – one that it shares with structural approaches – is that it relies on parametric functional forms. Because the ASF is not parametrically identified when y_2 is discrete, we have few options. Either we can use a parametric structural approach – an example is given in Section 4.1 – the CF approach, change the quantity of interest, or only try to bound specific parameters (such as an average treatment effect). The CF approach proposed here should be viewed as a computationally simple complement to other approaches.

7. Concluding Remarks

I have argued that a general quasi-LIML approach can be used to obtain one-step estimator for nonlinear models with endogenous explanatory variables. This approach leads to estimators that are new for certain kinds of response variables, including a fractional response with a binary endogenous explanatory variable. There are both theoretical and practical issues left to be resolved. For example, in a quasi-MLE framework, are there useful conditions under which the one-step quasi-LIML is asymptotically more efficient than a two-step control function approach? Also, in a nonlinear setting, when might the one-step estimator have less bias than a two-step method (provided there is a consistent two-step estimator available)?

The variable addition tests can be applied in a variety of settings when a generalized residual for the EEV can be computed. These tests are computationally very simple. One issue that needs further study is the best way to obtain tests when $\mathbf{y}_2$ is a vector of EEVs, some of which are discrete.

The CF framework for discrete EEVs proposed in Section 6 can be justified under parametric assumptions – assumptions that are no more or less general than more traditional assumptions. The CF approach leads to simple two-step estimators, simple tests of the null of exogeneity, and straightforward estimation of average partial effects. Unfortunately, unlike in the case where y_2 is continuous, we cannot simply view the parametric assumptions as convenient approximations: they are used to identify the average structural function. Nevertheless, the parametric assumptions might still provide a useful approximation, something that can be studied very simulation.

References

Bekker, P. (1994), “Alternative Approximations to the Distribution of Instrumental Variables Estimators,” *Econometrica* 62, 657-681.

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.

Burnett, N. (1997), “Gender Economics Courses in Liberal Arts Colleges,” *Journal of Economic Education* 28, 369-377.

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

Gourieroux, C., A. Monfort, and A. Trognon (1984), “Pseudo Maximum Likelihood Methods: Theory,” *Econometrica* 52, 681-700.

Gourieroux, C., A. Monfort, E> Renault, and A. Trognon (1987), “Genereralised Residuals,” *Journal of Econometrics* 34, 5–32.

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

Imbens, G.W., and J.M. Wooldridge (2009), “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature* 47, 5–86.

Petrin, A., and K. Train (2010), “A Control Function Approach to Endogeneity in Consumer Choice Models,” *Journal of Marketing Research* 47, 3-13.

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.

Staiger, D., and J.H. Stock, (1997), “Instrumental Variables Regression with Weak Instruments,” *Econometrica* 68, 1055-1096.

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

Terza, J.V. (2009), “Parametric Nonlinear Regression with Endogenous Switching,” *Econometric Reviews* 28, 555-580.

Terza, J.V., A. Basu, and P. J. Rathouz (2008), “Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling,” *Journal of Health Economics* 27, 531-543.

Vella, F. (1993), “A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors,” *International Economic Review* 34, 441-457.

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

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: Cambridge University Press.

Wooldridge J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. Cambridge, MA: MIT Press.