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

Jeff Wooldridge
Michigan State University

October 20, 2011

1. **Introduction**

- Several motivations for this paper.

1. Approaches for handling endogenous explanatory variables (EEVs) for certain kinds of responses (for example, fractional and count) are not widely available – especially for discrete EEVs.

2. In linear models, some evidence that (quasi-) limited information maximum likelihood (LIML) has less bias than two stage least squares (2SLS). Might the same be true in nonlinear contexts?

3. Two-step plug-in estimators are generally inconsistent for quantities of interest, especially with discrete EEVs. (“Forbidden regression.”) So, for example, if the response y_1 is in the unit interval (possibly taking values at the corners), and y_2 is binary, how can we estimate parameters and average partial (treatment) effects? MLE a possibility, but might want something simpler and more robust.

4. Estimating nonlinear models with multiple EEVs, with some discrete, is very difficult using traditional (MLE) approaches. Might some simple control function methods work well?

- Contributions

1. For relatively simple models – for example, a fractional or count response, with a single, binary EEV – show that certain joint quasi-MLEs – "quasi-LIMLs" – are easy to estimate.
2. Derive simple tests of the null of exogeneity. Simple variable addition tests based on generalized residuals. Only require correct specification of the conditional mean under the null.
3. Argue that two-step control function QMLEs using generalized residuals might produce good estimates of average partial effects. The approach is very flexible but does not follow from "standard" assumptions.

2. Example: A Linear Model

- Let y_1 be the response variable, y_2 the endogenous explanatory variable (EEV), and $\mathbf{z}$ the $1 \times L$ vector of exogenous variables (with $z_1 = 1$):

$$y_1 = \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{o1} + u_1, \tag{1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of $\mathbf{z}$. First consider the exogeneity assumption

$$E(\mathbf{z}' u_1) = \mathbf{0}. \tag{2}$$

- Give a random sample, 2SLS is consistent under the rank condition.

- LIML approach. The reduced form for y_2 is a linear projection:

$$y_2 = \mathbf{z}\boldsymbol{\delta}_{o2} + v_2,\ E(\mathbf{z}'v_2) = \mathbf{0}, \tag{3}$$

where $\boldsymbol{\delta}_{o2}$ is $L \times 1$. Write the linear projection of u_1 on v_2, in error form, as

$$\begin{aligned} u_1 &= \rho_{o1}v_2 + e_1 \\ E(v_2e_1) &= 0, \end{aligned} \tag{4}$$

where $\rho_{o1} = \mathrm{E}(v_2u_1)/\mathrm{E}(v_2^2)$ is the population regression coefficient.

- Also know that $\mathrm{E}(\mathbf{z}'e_1) = \mathbf{0}$.

- Plugging (4) into (1):

$$\begin{aligned} y_1 &= \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + \rho_{o1} v_2 + e_1 \qquad (5) \\ E(\mathbf{z}' e_1) &= \mathbf{0}, E(v_2 e_1) = 0 \qquad (6) \end{aligned}$$

- Given a random sample of size N, can use a two-step procedure: (i) Regress y_{i2} on $\mathbf{z}_i$ and obtain the reduced form residuals, $\hat{v}_{i2}$; (ii) Regress

$$y_{i1} \text{ on } y_{i2}, \mathbf{z}_{i1}, \text{ and } \hat{v}_{i2}. \qquad (7)$$

- The OLS estimates from (7) are *control function* (CF) estimates. 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.

- Why use a two-step method, other than computational simplicity?

Can estimate the RF and structural parameters in a single step.

- First, write

$$y_1 = \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{o1} + \rho_{o1}(y_2 - \mathbf{z}\boldsymbol{\delta}_{o2}) + e_1 \tag{8}$$

$$E(\mathbf{z}' e_1) = \mathbf{0},\ E(y_2 e_1) = 0. \tag{9}$$

- Under $E(\mathbf{z}'u_1) = \mathbf{0}$ only, we can write two linear projections:

$$L(y_1|y_2,\mathbf{z}) = \alpha_{o1}y_2 + \mathbf{z}_1\boldsymbol{\delta}_{o1} + \rho_{o1}(y_2 - \mathbf{z}\boldsymbol{\delta}_{o2}) \tag{10}$$

$$L(y_2|\mathbf{z}) = \mathbf{z}\boldsymbol{\delta}_{o2} \tag{11}$$

- Together these identify the parameters $\alpha_{o1}, \boldsymbol{\delta}_{o1}, \rho_{o1}$, and $\boldsymbol{\delta}_{o2}$ because these parameters solve the population problem

$$\min_{\alpha_1,\boldsymbol{\delta}_1,\rho_1,\boldsymbol{\delta}_2} \{E([y_1 - \alpha_1 y_2 - \mathbf{z}_1\boldsymbol{\delta}_1 - \rho_1(y_2 - \mathbf{z}\boldsymbol{\delta}_2)]^2) + E[(y_2 - \mathbf{z}\boldsymbol{\delta}_2)^2]\} \tag{12}$$

- Does not quite lead to LIML under normality because it ignores the variance parameters. So use the quasi-LIML objective function:

$$\begin{aligned}\min_{\alpha_1,\boldsymbol{\delta}_1,\rho_1,\boldsymbol{\delta}_2,\eta_1^2,\tau_2^2} \sum_{i=1}^{n} \{[y_{i1} - \alpha_1 y_{i2} - \mathbf{z}_{i1}\boldsymbol{\delta}_1 - \rho_1(y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)]^2/\eta_1^2 \\ + (y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)^2/\tau_2^2\} + \log(\eta_1^2) + \log(\tau_2^2)\end{aligned} \tag{13}$$

- Can easily show that $\alpha_{o1}, \boldsymbol{\delta}_{o1}, \rho_{o1}$, and $\boldsymbol{\delta}_{o2}$ solve the population analog, and then we can simply *define*

$$\eta_{o1}^2 \equiv E(e_1^2)$$
$$\tau_{o2}^2 \equiv E(v_2^2)$$

- No normality or homoskedasticity in sight. No linear conditional expectations, either. Driven entirely by $E(\mathbf{z}'u_1) = \mathbf{0}$ (and the rank condition). y_2 could be continuous, discrete, or some mixture.

3. Framework for Quasi-LIML for Nonlinear Models

- As an example, suppose y_1 is a binary response and y_2 is continuous:

$$y_1 = 1[\alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{o1} + u_1 \geq 0] \tag{14}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_{o2} + v_2 \tag{15}$$

where (u_1, v_2) is bivariate normal with mean zero and independent of $\mathbf{z}$.

- Can show that

$$P(y_1 = 1|y_2, \mathbf{z}) = \Phi\left[\frac{(\alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{o1} + (\rho_{o1}/\tau_{o2})(y_2 - \mathbf{z}\boldsymbol{\delta}_{o2})}{(1 - \rho_{o1}^2)^{1/2}} \right], \tag{16}$$

where $\tau_{o2}^2 = Var(v_2)$ and $\rho_{o1} = Corr(v_2, u_1)$.

- If we *define* $v_2 = y_2 - \mathbf{z}\boldsymbol{\delta}_{o2}$, where $\boldsymbol{\delta}_{o2}$ is the vector of linear projection parameters, and $\tau_{o2}^2 \equiv E(v_2^2)$, then the Gaussian quasi-log-likelihood function for $D(y_2|\mathbf{z})$ identifies these parameters.
- If we then assume that $D(u_1|y_2,\mathbf{z}) = D(u_1|v_2)$, and that the latter has mean linear in v_2 and is homoskedastic normal, the so-called “IV probit” estimator is still consistent even though the full distributional assumptions do not hold.
- We can go further and allow y_1 to be a fractional response with essentially any distribution, provided $E(y_1|y_2,\mathbf{z})$ has the form in (16).

- General Setup: Let $\boldsymbol{\theta}_{o1}, \boldsymbol{\theta}_{o2}$ be the parameters appearing in the model for (the feature of) $D(\mathbf{y}_1|\mathbf{y}_2, \mathbf{z})$, where only $\boldsymbol{\theta}_{o2}$ appears in (the feature of) $D(\mathbf{y}_2|\mathbf{z})$. If $\boldsymbol{\theta}_{o2}$ 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

$$\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \{E[q_1(\mathbf{y}_1, \mathbf{y}_2, \mathbf{z}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)] + E[q_2(\mathbf{y}_2, \mathbf{z}, \boldsymbol{\theta}_2)]\}. \tag{17}$$

- How should we choose the objective functions to ensure some robustness and, perhaps, efficiency in some cases?
- The joint estimators will be generally as robust as two-step estimators (when the latter are even justified at all).

Asymptotics

- With smooth objective functions, asymptotics is standard. It will often be the case that the scores for the two problems are uncorrelated because $\boldsymbol{\theta}_o$ often solves

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

Then

$$Avar\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_1^{-1}\mathbf{B}_1\mathbf{A}_1 + \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{A}_2.$$

- Further simplifications of the sandwiches are sometimes available.

4. Example: Fractional Response

- Set up endogeneity as an omitted variable problem, and start by assuming y_2 is continuous:

$$E(y_1|\mathbf{z}, y_2, r_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1). \tag{19}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2, \tag{20}$$

where $\mathbf{x}_1$ is a general nonlinear function of $(\mathbf{z}_1, y_2)$, r_1 is an omitted factor thought to be correlated with y_2 but independent of the exogenous variables $\mathbf{z}$.

- The average partial effects in this model are obtained from the "average structural function" (ASF):

$$ASF(\mathbf{x}_1) = E_{r_1}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1)] = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{r1})$$

where

$$\boldsymbol{\beta}_{r1} = \boldsymbol{\beta}_1/(1+\sigma_{r_1}^2)^{1/2}.$$

- These are the only identified parameters, anyway.
- If (r_1, v_2) is jointly normal, a two-step control function method is valid. Note that the distribution of y_1 is not further restricted. under

joint normality:

(i) Regress $h_2(y_{i2})$ on $\mathbf{z}_i$ and obtain the residuals, $\hat{v}_{i2}$.

(ii) Use “probit” of y_{i1} on $\mathbf{x}_{i1}, \hat{v}_{i2}$ to estimate parameters with different scales, say $\hat{\boldsymbol{\beta}}_{e1}$ and $\hat{\gamma}_{e1}$. (Can implement as a “generalized linear model.”)

- The “average structural function” (ASF) is consistently estimated as

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_{e1} + \hat{\gamma}_{e1} \hat{v}_{i2}), \tag{21}$$

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

- What about a quasi-LIML approach? Can show that

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

and so we can plug this mean function into the Bernoulli quasi-log likelihood. This gives $q_1(y_1, y_2, \mathbf{z}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Identify $\boldsymbol{\delta}_2$ and τ_2 using the Gaussian QLL, which gives $q_2(y_2, \mathbf{z}, \boldsymbol{\theta}_2)$.

- The coefficients we estimate on $\mathbf{x}_1$ are those that index the average partial effect.

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

$$\begin{aligned} y_2 &= 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0] \\ v_2|\mathbf{z} &\sim Normal(0,1) \end{aligned}$$

- Can show that $E(y_1|y_2,\mathbf{z})$ has the same form as the response probability in the so-called “bivariate probit” model. For example,

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

- So for $q_2(y_2,\mathbf{z},\boldsymbol{\theta}_2)$ we use the usual probit log-likelihood. For $q_1(y_1,y_2,\mathbf{z},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ we use the Bernoulli QLL associated with bivariate probit.

5. Testing

- Suppose y_2 is binary and (nomially) follows a probit model. y_1 might be fractional, or a count variable, and so on. We do not want to use a full (conditional) distributional assumption for y_1, but we want to test the null that y_2 is exogenous.
- Using the score test is natural. Leads to simple variable-addition tests. The added variable is the generalized residual for y_2.

- For a continuously differentiable function $g(\cdot)$, assume

$$\begin{aligned} E(y_1|y_2,\mathbf{z},r_1) &= g(\mathbf{x}_1\boldsymbol{\beta}_1 + r_1) \\ y_2 &= 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0],\ v_2|\mathbf{z} \sim \ Normal(0,1) \\ r_1 &= \eta_1 v_2 + e_1 \\ & \quad e_1|\mathbf{z},y_2 {\sim} Normal(0,\tau_1^2 - \eta_1^2) \end{aligned}$$

- The null is $H_0 : \eta_1 = 0$. The gradient of the mean function, evaluated under the null estimates, is

$$g^{(1)}(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1) \cdot (\mathbf{x}_{i1}, \hat{r}_{i2})$$

where $g^{(1)}(\cdot)$ is the first derivative and

$$\begin{aligned} \hat{r}_{i2} &= y_{i2}\lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) - (1 - y_{i2})\lambda(-\mathbf{z}_i\hat{\boldsymbol{\delta}}_2) \\ \lambda(\cdot) &= \text{inverse Mills ratio} \end{aligned}$$

- VAT is simple. Estimate the probit model for y_2 and compute the $\hat{r}_{i2}$. Then use a suitable QMLE (Bernoulli, Poisson, gamma, normal) to estimate the mean function (for y_{i1})

$$g(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \gamma_1\hat{r}_{i2})$$

The robust *t* statistic for γ_1 is asymptotically valid under H_0.

- Gives a simple specification test of endogeneity in the context of estimate “treatment effects” for a variety of response variables y_1.
- If y_1 is binary or fractional, use probit or logit or some other regression function in the second stage. Just add $\hat{r}_{i2}0$.
- If y_{i1} is a count (or generally nonnegative), use an exponential function and the Poisson QMLE.
- Can easily extend the test to a nonlinear “switching regression” setup. If y_2 interacts with r_1, variables for VAT are

$$\hat{r}_{i2}, y_{i2} \cdot \hat{r}_{i2}$$

so a two degrees-of-freedom test.

- For a completely general switching regression, also add $\mathbf{z}_{i1} \cdot \hat{r}_{i2}$.

6. Some Radical Suggestions

- First, not so radical. If y_2 is continuous, and r_1 are the unobservables in the "structural" model $E(y_1|y_2,\mathbf{z}_1,r_1)$, just assume $D(r_1|y_2,\mathbf{z}) = D(r_1|v_2)$ for v_2 a residual or standardized residual. Then, can model $D(y_1|y_2,\mathbf{z}_1,v_2)$ or $E(y_1|y_2,\mathbf{z}_1,v_2)$ in a flexible way, without trying to make it consistent with an underlying model such as $y_1 = g_1(y_2,\mathbf{z}_1,u_1)$ for unobservables u_1.

- For example, if y_1 is binary or a fractional response, just use a flexible probit model for $h_1(y_2, \mathbf{z}_1, v_2) = E(y_1|y_2, \mathbf{z}_1, v_2)$, where $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$. So, general functions of v_2, including interactions with elements of $(\mathbf{z}_1, y_2)$. In the end, the average partial effects are obtained by averaging out the $\hat{v}_{i2}$:

$$ASF(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} h_1(y_2, \mathbf{z}_1, \hat{v}_{i2}) \tag{31}$$

Either a control function or quasi-LIML approach can be used to estimate the parameters. We can use “heteroskedastic probit” to make the functional form more flexible. This is a “flexible parametric” approach to Blundell and Powell (2004).

- Similar strategies are available if y_1 is a corner solution, or an ordered response, or a multinomial response. We can model $E(y_2|\mathbf{z}) = \mu_2(\mathbf{z})$ and $E(y_2|\mathbf{z}) = \omega_2(\mathbf{z})$ in flexible ways, use the Gaussian quasi likelihood to identify the parameters, and then assume that $e_2 = (y_2 - \mu_2(\mathbf{z}))/\sqrt{\omega_2(\mathbf{z})}$ is a sufficient statistic for $D(r_1|y_2, \mathbf{z})$, where r_1 are the unobservables in the structural model for y_1.

- How far can we take this approach? A very radical approach to handle discrete y_2 is to assert that one or a few functions of $(y_2, \mathbf{z})$ characterize $D(r_1|y_2, \mathbf{z})$. Very little progress has been made estimating general models with discrete EEVs, unless full parametric assumptions are made. (And then computation is often quite difficult.) Suppose y_2 is binary. If e_2 is a function of $(y_2, \mathbf{z})$ that depends sufficiently on $\mathbf{z}_2$, such that

$$D(r_1|y_2, \mathbf{z}) = D(r_1|e_2), \tag{32}$$

then the ASF (APEs) can be identified and estimated quite generally by estimating $E(y_1|y_2, \mathbf{z}_1, e_2)$, and then average out $\hat{e}_{i2}$.

- Problem: How should we choose e_2? A standardized residual? A generalized residual? Can get some flexibility here, but, generally, e_2 does not appear to exist in traditional formulations.

6. Remaining Issues

- When will the one-step QMLEs be more efficient that two-step QMLEs? [One case, of course, is when the joint estimator is a (conditional) MLE.]
- Are the finite-sample properties of the one-step estimator generally better than two-step estimators? (Weak instrument problem.)
- Asymptotics as the number of instruments grows [Bekker (1994, *Econometrica*)].