Estimating the *Fractional Response Model* with an *Endogenous Count Variable*

Estimating *FRM* with *CEEV*

Hoa Bao Nguyen Minh Cong Nguyen

Michigan State Universtiy American University

July 2009

Econometric Issues and Motivation

- For nonlinear model, accounting for endogeneity is essentially complicated because simple two-stage regression strategies analogous to the Heckman (1979) method are only approximate and inference based on such procedures may lead to wrong conclusions (Wooldridge 2002).
- Many count endogenous explanatory variable (CEEV) have been treated as continuous endogenous variable and therefore increasing the value in the CEEV by 1 (e.g., from c_k to c_{k+1}) at different interesting values of the CEEV has roughly same effect on the dependent variable.
- There is no routine for estimation of the fractional response model (FRM) with CEEV. Full-Information Maximum Likelihood approach is possible but it requires major computations and time consumption that it is usually used in binary response model or count model (Romeu and Vera-Hernandez 2005, Weiss 1999, Terza 1998, Heckman 1978).

Contribution

- This presentation will describe the implementation of the Quasi-Maximum Likelihood (QMLE) techniques with control function approach (Blundell and Powell 2003, 2004) to estimate such FRM with CEEV (based on the theoretical paper by Hoa Nguyen and Jeffrey Wooldridge 2009).
- Properties of this estimator and other estimators will be compared using Monte Carlo simulations.
- Stata has several commands for continuous outcome with endogenous continuous or dummy explanatory variable (ivprobit, ivtobit, treatreg, ssm or gllamm). However, there are currently no commands for a CEEV in a nonlinear model. The **frcount** command will set a base for estimating nonlinear model with CEEV.

Model Specification

- The conditional mean is expressed as:

$$E(y_1|y_2, \mathbf{z}, a_1) = \Phi(\alpha_1 y_2 + \mathbf{z}_1\delta_1 + \eta_1 a_1)$$

- y_2 is the count endogenous variable with the conditional mean:

$$E(y_2|\mathbf{z}, a_1) = \exp(\mathbf{z}\delta_2 + a_1)$$

- $y_2|\mathbf{z}, a_1$ has a Poisson distribution while $y_2|\mathbf{z}$ is NBII distributed.
- $\mathbf{z}, \mathbf{z}_1$ are $1 \times K$ and $1 \times L$ vector of exogenous explanatory variables $(L > K)$
- a_1 is independent of $\mathbf{z}$ and $\exp(a_1) = c_1 \sim \text{Gamma}(\delta_0, \delta_0)$

Model Specification

- We are interested in:
  $E(y_1|y_2,\mathbf{z}) = \int_{-\infty}^{+\infty} \Phi(\alpha_1 y_2 + \mathbf{z}_1\delta_1 + \eta_1 a_1) f(a_1|y_2,\mathbf{z}) da_1$
  $= \mu(\theta_1; y_2,\mathbf{z})$ where $\boldsymbol{\theta}_1 = (\alpha_1, \delta_1', \eta_1)'$ and $\mathbf{g}_i = (y_{2i}, \mathbf{z}_{1i}, a_{1i})$
- The conditional mean $E(y_1|y_2,\mathbf{z})$ cannot simply have the closed-form solution so we need to use numerical approximation. The numerical routine for integration of unobserved heterogeneity in the conditional mean equation is based on the Adaptive Gauss–Hermite Quadrature. Once the evaluation has been done, the numerical values can be passed on to a maximizer in order to find the QMLE $\hat{\boldsymbol{\theta}}_1$
- The QMLE of $\boldsymbol{\theta}_1$ is obtained from the maximization problem:
  $\underset{\theta_1 \in \Theta}{Max} \sum_{i=1}^{n} l_i(\boldsymbol{\theta}_1)$
  where the Bernoulli log-likelihood function is given by:
  $l_i(\boldsymbol{\theta}_1) = y_{i1} \ln \mu_i + (1 - y_{i1}) \ln(1 - \mu_i)$
  with μ_i is evaluated with numerical approximation.

Model Estimation

- Two-step Estimation Procedure:
  - Estimate the reduced form of CEEV by using maximum likelihood of y_{i2} on $\mathbf{z}_i$ in the NB model. Obtain the estimated parameters $\hat{\delta}_2$ and $\hat{\delta}_0$.
  - Use the QMLE of y_{i1} on y_{i2}, z_{i1} to estimate α_1, δ_1 and η_1 with the approximated conditional mean. Approximate the conditional mean using the estimated parameters in the first step and by using the Adaptive Gauss-Hermite method. Obtain estimated parameters $\hat{\alpha}_1$, $\hat{\delta}_1$ and $\hat{\eta}_1$.
- Standard errors in the second stage is adjusted for the first stage estimation and computed by using delta method.

$$\sqrt{N}(\hat{\delta}_2 - \delta_2) = N^{-1/2} \sum_{i=1}^{N} \mathbf{r}_{i2} + o_p(1)$$

$$\sqrt{N}(\hat{\theta}_1 - \theta_1) = \mathbf{A}_1^{-1}(N^{-1/2} \sum_{i=1}^{N} \mathbf{r}_{i1}(\theta_1; \delta_2)) + o_p(1)$$

$$\hat{\mathbf{r}}_{i1}(\theta_1; \delta_2) = \hat{\mathbf{s}}_i(\theta_1; \delta_2) - \hat{\mathbf{F}}_1 \hat{\mathbf{r}}_{i2}(\delta_2)$$

$$\widehat{Avar(\hat{\theta}_1)} = \frac{1}{N} \hat{\mathbf{A}}_1^{-1} \left(N^{-1} \sum_{i=1}^{N} \hat{\mathbf{r}}_{i1} \hat{\mathbf{r}}_{i1}' \right) \hat{\mathbf{A}}_1^{-1}$$

Average Partial Effects

- We are interested in partial effects of the explanatory variables in non-linear models in order to get comparable magnitudes.
  - The conditional mean model is:
  $E(y_1|y_2, \mathbf{z}, a_1) = \Phi(\alpha_1 y_2 + \mathbf{z}_1\delta_1 + \eta_1 a_1)$
  - The APE is obtained, for given $\mathbf{z}_1, y_2$, by averaging the partial effect across the distribution of a_1 in the population
- For a continuous z_1, $APE = E_{a_1}[\delta_{11}\phi(\alpha_1 y_2 + \mathbf{z}_1\delta_1 + \eta_1 a_1)]$
  $= \delta_{11}\int_{-\infty}^{+\infty}\phi(\alpha_1 y_2 + \mathbf{z}_1\delta_1 + \eta_1 a_1)f(a_1)da_1$

$$\widehat{APE} = \hat{\delta}_{11}N^{-1}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\phi(\hat{\alpha}_1 y_{2i} + \mathbf{z}_{1i}\hat{\delta}_1 + \hat{\eta}_1 a_1)f(a_1)da_1$$

- For a count y_2, for y_2 going from c_k to c_{k+1}

$$APE = \Phi(\alpha_1 c_{k+1} + \mathbf{z}_1\delta_1 + \eta_1 a_1) - \Phi(\alpha_1 c_k + \mathbf{z}_1\delta_1 + \eta_1 a_1)$$

$$\widehat{APE} = N^{-1}\sum_{i=1}^{N}\left(\begin{array}{l}\int_{-\infty}^{+\infty}\Phi(\hat{\alpha}_1 c_{k+1} + \mathbf{z}_{1i}\hat{\delta}_1 + \hat{\eta}_1 a_1)f(a_1)da_1 \\ -\int_{-\infty}^{+\infty}\Phi(\hat{\alpha}_1 c_k + \mathbf{z}_{1i}\hat{\delta}_1 + \hat{\eta}_1 a_1)f(a_1)da_1\end{array}\right)$$

Implementation of the Command

- Syntax

The general syntax of the command is as follows:
frcount depvar varlist [if] [in], endog(varname) iv(varlist) quad(#) maximize_options

- Details:
  - endo (varname) specifies that endogenous variable be included in varname
  - iv(varlist) specifies that instrument(s) be included in varlist
  - quad(#) sets the number of quadrature
  - maximize_options are passed to QMLE or NLS.

Theses methods have to be indicated clearly.

Monte Carlo Simulation

- Data generating process:

- The CEEV y_2 is generated from a Poisson distribution with conditional mean:

$$\lambda = \mathbf{E}(\mathbf{y}_2|\mathbf{z}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}_1) = \mathbf{exp}(\mathbf{0.01} * \mathbf{x}_1 + \mathbf{0.01} * \mathbf{x}_2 + \mathbf{2} * \mathbf{iv} + \mathbf{a}_1)$$

using independent draws of the normal variables
$iv \sim N(0, 0.2^2), x_1 \sim N(0, 0.005^2), x_2 \sim N(0, 0.01^2)$ and
$exp(a_1) \sim Gamma(1, 1/\delta_0)$ where $\delta_0 = 2$.

- We generate the dependent variable y_1 by first drawing a binomial random variable x with n trials and probability p and then $y_1 = x/n$. In this simulation, we use $n = 100$ and

$$\mathbf{p} = \mathbf{E}(\mathbf{y}_1|\mathbf{y}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}_1) = \Phi(\mathbf{0.1} * \mathbf{x}_1 + \mathbf{0.1} * \mathbf{x}_2 - \mathbf{0.1} * \mathbf{y}_2 + \mathbf{0.5} * \mathbf{a}_1)$$

- We run 500 replications for three sample sizes 500, 1000 and 5000 respectively.

Table 1: Simulation Results

- We report sample means and sample standard deviations of these 500 estimates. Table 1 show the results of simulations for parameter α_1 estimated by QMLE and three alternative estimators for different sample size.
- Table of Results

Simulation Result						
Sample Size	**500**		**1000**		**5000**	
Method	Mean	(S.d.)	Mean	(S.d.)	Mean	(S.d.)
OLS	.0124	.0044	.0122	.0030	.0122	.0013
2SLS	-.0389	.0162	-.0386	.0107	-.0382	.0050
QMLE	-.1025	.0124	-.1010	.0089	-.1000	.0036
NLS	-.1025	.0124	-.1011	.0089	-.0999	.0036

Conclusion

- The QMLE method produces unbiased and more efficient estimates compared to three alternative estimators.
- The QMLE estimates are significant and APEs are available to get different magnitudes for increasing values of the CEEV. Other traditional estimates only produce the same partial effect for increasing values of the CEEV.

Example

```
. frcount y1 x1 x2, endog(y2) iv(iv) quad(50) qmle

Getting Initial Values:
Iteration 0:  log likelihood = -845.94337
Iteration 1:  log likelihood = -845.33452
Iteration 2:  log likelihood =  -845.3345
Resetting the parameters...
Iteration 0:  log likelihood =  -848.3366
Iteration 1:  log likelihood =  -848.2526
Iteration 2:  log likelihood =  -848.2526
Resetting the parameters...
Iteration 0:  log likelihood = -848.44551
Iteration 1:  log likelihood = -848.44551

Fitting Quasi-MLE Model:
Iteration 0:  log likelihood = -679.63122
Iteration 1:  log likelihood = -674.69828
Iteration 2:  log likelihood = -674.69328
Iteration 3:  log likelihood = -674.69327
```

Quasi-MLE Fractional Response Model — Number of obs = 1000; Corr(y1, y1hat) = .0665616

Log likelihood: -674.69327 — Pseudo R2 = .0656221

y1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y2	-.0940061	.0053054	-17.72	0.000	-.1044046	-.0836077
x1	1.315443	2.302827	0.57	0.568	-3.198015	5.828902
x2	1.63701	1.187422	1.38	0.168	-.6902948	3.964314
eta1	.4940986	.0296782	16.65	0.000	.4359305	.5522667

Future Work

- Run simulations for different values of η_1.
- Adopt Simulated Maximum Likehood method in order to get approximation and compare with the Adaptive Gauss-Hermite method.
- Simulate data with censored fractional response variable and compare the QML estimator of the FRM and the Tobit estimator.

Thank You

THANK YOU!