EXPLORING ALTERNATIVES FOR ESTIMATING SYSTEMS OF EQUATIONS WITH MULTIPLE CENSORED VARIABLES:

Farm Output Supply and Input Demand

Carlos Arias and Federico Perali

Carlos Arias, Dep. of Economics, University of Oviedo, Avda del Cristo sn, 33071 Oviedo, Spain. E- mail: arias@econo.uniovi.es; Phone: 34-985104879; Fax: 34-985104871

Federico Perali, Dept. of Economics, University of Verona, Via dell'Artigliere, 19, 37129 Verona, Italy. E-mail: wicaro@chiostro.univr.it; Phone: 39-458098486; Fax: 39-458098529

1999 AAEA ANNUAL MEETING
AUGUST 8-11
NASHVILLE, TN

EXPLORING ALTERNATIVES FOR ESTIMATING SYSTEMS OF EQUATIONS WITH MULTIPLE CENSORED VARIABLES:

Farm Output Supply and Input Demand

Carlos Arias and Federico Perali[1]

ABSTRACT:

This paper explores two alternatives for estimating systems of equations with multiple censored variables: Maximum Simulated Likelihood and a two-step technique that seems to be well suited for large samples. The empirical part of the paper estimates a system of cost, cost shares and revenue shares equations of Italian farms using both approaches.

Subject code: 18 Research Methods/Econometrics/Statistics

[1] Carlos Arias, Dep. of Economics, University of Oviedo, Avda del Cristo sn, 33071 Oviedo, Spain. E- mail: arias@econo.uniovi.es; Phone: 34-985104879; Fax: 34-985104871

Federico Perali, Dept. of Economics, University of Verona, Via dell'Artigliere, 19, 37129 Verona, Italy. E-mail: wicaro@chiostro.univr.it; Phone: 39-458098486; Fax: 39-458098529

1. Introduction

There are many examples in economics of models that can be represented by systems of equations with several censored variables: systems of demand equations where some consumers choose not to buy several of the goods in the system (Wales and Woodland, 1983; Lee and Pitt, 1986; Phaneuf, 1999) or systems of input demand and supply equations where firms choose not to produce several of the outputs or not to use several of the inputs in the system (Lee and Pitt, 1984; Huffman, 1988).

The latter example is the object of interest of the present paper since we claim that the estimation of a system of input demand and output supply at a high level of detail is very important for the microsimulation and welfare analysis of farm policy. For example, the policy analysis of the cereal market would not be as relevant if the model cannot explain the choice associated to the production of tender and durum wheat. In Italy, the first is produced in the North, the latter in the South and an aggregate treatment of generic wheat would not allow policy makers to target sound regional policies. In agricultural samples, it is common that a substantial number of farms do not produce the full set of outputs and do not use the full set of inputs. In this case, we have a system of equations with censored variables.

The methodological part of the present paper describes two alternative approaches for estimation of systems of equations with multiple censored variables. The first one, is the estimation of the system of equations by simulated maximum likelihood. The main problem of this approach is the need to evaluate high dimensional integrals. We propose to compute these integrals using a probability simulation method. The second one, proposes a two step methodology that seems promising for large samples, or for relatively small samples and large systems of stacked equations. In the first step, the system of linear Tobit equations is

estimated equation by equation for relatively small random samples on which we implement the jackknife technique to obtain consistent parameter estimates while reducing the computational burden. The second step uses a minimum distance estimator for imposing the parametric constraints required by economic theory.

The empirical part of the paper estimates a system that includes a cost function, the input demand shares and revenue share equations of Italian cereal farms using both approaches. The method of simulated maximum likelihood is considered the exact method, while the two stage Tobit system estimation is the approximation that most closely reproduces the same data generating process. This empirical example provide a nice setting for an interesting comparison of both techniques. Our experimental setup has been chosen also with the objective to learn about the consistency properties of the proposed estimators and the optimal stopping rules of the simulation procedure.

The structure of the paper is as follows: Section 2 reviews theoretical issues related to cost and input demand equations wit censoring, Section 3 discusses the two estimation methods for systems of equations with multiple censored variables, Section 4 presents an empirical application to the sample of Italian cereal farms.

2. Theoretical issues

Cost functions can be used to measure the welfare impact on farms of changes in input prices or in output quantities. The welfare impact can be different depending on the exogenous characteristics of the farm describing exogenous characteristics. The modified cost function can be written as:

$$C(\mathbf{w}, \mathbf{y}, \mathbf{d}) = \min_{\mathbf{x}} \{ \mathbf{wx} \text{ st } F(\mathbf{x}, \mathbf{y}, \mathbf{d}) = 0 \} \quad (1)$$

where **w** is a vector of input prices, **y** is a vector of outputs, **d** is a vector of demographic characteristics, **x** is a vector of inputs and F is a transformation function with the usual properties. The structure of the minimization problem implies that the cost function is homogeneous of degree one on input prices. Two additional properties of the cost function are of interest for estimation purposes. First, by Shephard's lemma we have that:

$$x_i(\mathbf{w}, \mathbf{y}, \mathbf{d}) = \frac{\partial C(\mathbf{w}, \mathbf{y}, \mathbf{d})}{\partial w_i} \tag{2}$$

where, x_i is the i-th component of the input vector and w_i is the price of that input. These input demand functions are homogeneous of degree zero on input prices. Second, if farmer are maximizing profit, we have that:

$$p_j = \frac{\partial C(\mathbf{w}, \mathbf{y}, \mathbf{d})}{\partial y_j} \tag{3}$$

It is common practice to estimate the cost function together with the equations implied by Shephard's lemma and the profit maximization condition (Huffman, 1988). The properties derived from the optimization structure of the model can be tested/imposed as well. Our methodological approach assumes that the zero realizations are the outcome of physical or technological constraints thus justifying a Tobit structure.

3. Methodological issues

In this section, we review two feasible methods of estimation for systems of equations with multiple censored variables: maximum simulated likelihood and a minimum distance estimator.

3.1. Maximum Simulated Likelihood

The likelihood function of a system of equations in the case in which all endogenous variables are above the censoring levels is given by:

$$L_1 = f(u_1, \ldots, u_m), \tag{4}$$

where the u_i's are the random disturbances of the system of equations and f is the probability density function of a multivariate normal function with mean zero and variance **Ω**. The likelihood function for an observation in which the n first endogenous variables out of m are censored is:

$$\begin{aligned} L_2 &= \int_{-\infty}^{c1} \ldots \int_{-\infty}^{cn} f(u_1, \ldots, u_m)\, du_1 \ldots du_n \\ &= f_1(u_{n+1}, \ldots, u_m) \int_{-\infty}^{c1} \ldots \int_{-\infty}^{cn} g(u_1, \ldots, u_n \mid u_{n+1}, \ldots, u_m)\, du_1 \ldots du_n \end{aligned} \tag{5}$$

where, f_1 is the marginal probability density function of the uncensored portion and g is the conditional marginal density function.

Expression (5) represents a portion of the likelihood function with an n-dimensional definite integral. Under the common assumption of multivariate normality of the disturbances of the system this integral does not have a closed form solution. Therefore, estimating the system of equations by maximum likelihood requires an efficient method for evaluating the high dimensional definite integrals. Maximum Simulated Likelihood (MSL) consists on simulating rather than calculating these integrals using probability simulation methods.

Probability simulation methods are based on the fact that the integral of interest represents the probability of an event in a population. Lerman and Manski (1981) propose generating a pseudo-random sample of observations from the relevant population and using the relative frequency of the event in the sample to approximate the integral of interest. This simulation method is called a "crude frequency simulator" and it was improved in several subsequent papers. Stern (1992) explains the importance of smoothness in a probability simulator and proposes an smooth alternative to the "crude frequency simulator". Geweke

(1989) and Borsh-Saupan and Hajivassiliou (1993) proposed the GHK simulator. Hajivassiliou et al. (1996) find that the GHK probability simulator outperforms all other methods by keeping a good balance between accuracy and computational costs.

The GHK simulator computes the value of the integral:

$$\Pr(\mathbf{a} < \mathbf{u} < \mathbf{b}) = \int_{\mathbf{a}}^{\mathbf{b}} g(\mathbf{u})d\mathbf{u} \tag{6}$$

where, **u** is a random vector distributed multivariate normal with mean **0** and variance **Ω** and g is the density function of the random vector **u**. The starting point is that:

$$\Pr(\mathbf{a} < \mathbf{u} < \mathbf{b}) = \Pr(\mathbf{a} < \mathbf{Le} < \mathbf{b}) \tag{7}$$

where, **L** is the lower triangular Cholesky factor of **Ω**, such that **LL'**=**Ω**, and **e** is a random vector of independent standard normal variables. The right hand side of expression (2) is easier to simulate than the probability in the left hand side due to the triangular structure of the constrains defined by **Le**. The intervals defining the event in the right hand side of expression (7) can be written as:

$$\begin{aligned} a_1 &< l_{11}e_1 < b_1 \\ a_2 &< l_{12}e_1 + l_{22}e_2 < b_2 \\ &\ldots \\ a_n &< l_{1n}e_1 + \ldots + l_{nn}e_n < b_n \end{aligned} \tag{8}$$

where l_{ij}, a_i and b_i are the corresponding elements of L, **a** and **b**. Arranging terms in (8) the event in (7) can be decomposed into the following events:

$$
\begin{aligned}
A_1 &= \left\{ \frac{a_1}{l_{11}} < e_1 < \frac{b_1}{l_{11}} \right\} \\
A_2 &= \left\{ \frac{a_2 - l_{12} e_1}{l_{22}} < e_2 < \frac{b_2 - l_{12} e_1}{l_{22}} \right\} \\
&\ldots \\
A_n &= \left\{ \frac{a_n - l_{1n} e_1 - \ldots - l_{n-1n} e_{n-1}}{l_{nn}} < e_n < \frac{b_n - l_{1n} e_1 - \ldots - l_{n-1n} e_{n-1}}{l_{nn}} \right\}
\end{aligned} \tag{9}
$$

Expression (9) shows the recursive nature of the constraints that affect the random vector **e**. As a result, the probability of interest can be written as:

$$
\Pr(a < Le < b) = \Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_1, A_2)\ldots\Pr(A_n|A_1, \ldots, A_{n-1}) \tag{10}
$$

The idea behind the GHK simulator is that expression (10) can be difficult to calculate but can be simulated instead. Therefore, the GHK simulator can be written as:

$$
\tilde{\Pr}(a < Le < b) = \frac{1}{R} \sum_{r=1}^{R} \Pr(A_1)\Pr(A_2|e_{1r})\Pr(A_3|e_{1r}, e_{2r})\ldots\Pr(A_n|e_{1r}, \ldots, e_{n-1r}) \tag{11}
$$

where the e_{ir}´s are drawn sequentially from independent standard normal distributions truncated by expression (9) and R is the number of simulations. The truncated random variables e_{ir} can be generated smoothly using the integral transform theorem (Ross, 1988). Once the e_{ir}´s are drawn, the terms in the product are calculated as:

$$
\Pr(A_i | e_{1r}, e_{2r}, \ldots, e_{i-1r}) = \Phi\left(\frac{b_i - l_{i1} e_{1r} - \ldots l_{i-1i} e_{i-1r}}{l_{ii}} \right) - \Phi\left(\frac{a_i - l_{i1} e_{1r} - \ldots l_{i-1i} e_{i-1r}}{l_{ii}} \right) \tag{12}
$$

where Φ is the cumulative distribution function of a standard normal distribution function.

Borsch-Saupan and Hajivassiliou (1993) proved that the probability simulator in (11) is an unbiased estimator of the true probability.

3.2. Two steps

In a statistical sense, the present exercise is somewhat akin to the comparison of a full information maximum likelihood and a two stage least squares estimation. Within the context of a censored system of demand equations, our two stage approach proposes to estimate first each demand equation in the corresponding linear reduced form, without structural restrictions, using the jackknife technique both with the aim of reducing the large sample/large system problem and as a device to construct the variance covariance matrix of the truncated cross-equations error terms. Our object of interest is the expected revenue share conditional on both the non negative realization of the own share and another share pairwise:

$$\begin{aligned} &E(w_i \mid w_i \geq 0, w_j \geq 0) = f_i(z,\pi_i) + E(\varepsilon_i \mid \varepsilon_i \geq -f_i(z,\pi_i), \varepsilon_j \geq -f_i(z,\pi_i)) \\ &= f_i(z,\pi_i) + \frac{\sigma_i}{F(h_i, k_j \rho_{ij})}\left(\varphi(h_i)\left[1-\phi(k_j^*)\right] + \rho_{ij}\varphi(k_j)\left[1-\phi(h_j^*)\right]\right) \end{aligned} \tag{13}$$

where π are the reduced form parameters and z are exogenous explanatory variables, k and h are the standardized $f(z,\pi)$ functions in the different regimes, ϵ are the censored error terms, and ρ_{ij} are the cross-equations correlation coefficient. The reduced form of the estimated equations $f(z,\pi)$ is linear, as the structural specifications do, and does not incorporate the structural cross-equations restrictions. The variance covariance matrix of the complete set of parameters γ, including the matrix of correlation coefficient, is obtained using the jackknife estimator:

$$\Sigma_u = \frac{r}{dm}\sum_{s \in S_n}\left(\gamma_{us} - \frac{1}{m}\sum_{s \in S_n}\gamma_{us}\right)\left(\gamma_{us} - \frac{1}{m}\sum_{s \in S_n}\gamma_{us}\right)' \tag{14}$$

We then recover the structural demand parameters using minimum distance estimation and imposing the cross-equations theory restrictions as also previously done by Blundell *et al.*1993 and Browning and Meghir 1992 for the estimation of large demand systems:

$$\gamma_r = \arg\ min_{\gamma_r} \left\{ \left[\gamma_u - F(\gamma_r) \right]' \Sigma_u^{-1} \left[\gamma_u - F(\gamma_r) \right] \right\} \tag{15}$$

where Σ_u is a consistent estimate of the variance-covariance matrix of the unrestricted parameters γ_u obtained from the jackknife technique. Note that the computational algorithm that estimates the unrestricted parameters γ_u along with its covariance matrix $\sum_u$ in the first stage, and then recovers the structure in the second stage, is an indirect feasible generalized least square procedure. To the extent that it generates consistent and asymptotically efficient estimates, it provides a computationally convenient alternative to the full-information maximum likelihood estimation method.

3. Empirical Application

In the present paper, we estimate a system composed by a modified translog cost function (Lewbel, 1985) and its derivatives with respect to input prices and output quantities. The translog cost function modified via a simple translating can be written as:

$$\begin{aligned} \ln C = {} & \alpha_0 + \sum_{i=1}^{6} \alpha_i \ln y_i + \sum_{i=1}^{6}\sum_{j=1}^{6} \alpha_{ij} \ln y_i \ln y_j + \sum_{r=1}^{4} \beta_r \ln w_r + \sum_{r=1}^{4}\sum_{s=1}^{4} \beta_{rs} \ln w_r \ln w_s \\ & + \sum_{r=1}^{4}\sum_{i=1}^{6} \gamma_{ri} \ln w_r \ln y_i + \sum_{i=1}^{6} \rho_i \ln y_i m(\mathbf{d}) + \sum_{r=1}^{4} \sigma_r \ln w_r m(\mathbf{d}) \end{aligned} \tag{16}$$

where, y_i denotes the amount of output i produced, w_r is the price of input r and m(**d**) is a function of demographic characteristics that can be written as:

$$m(\mathbf{d}) = \sum_{k=1}^{5} \delta_k d_k \tag{17}$$

Using Shephard's lemma, the derivatives of the cost function with respect to the natural

logarithm of input prices can be written as:

$$s_r = \beta_r + \sum_{s=1}^{4} \beta_{rs} \ln w_s + \sum_{i=1}^{6} \gamma_{ri} \ln y_i + \sigma_r m(\mathbf{d}) \quad (18)$$

where, s_r is the cost share of input r. Using the profit maximization condition, the derivative of the cost function with respect to the natural logarithm of output quantities can be written as:

$$rs_i = \alpha_i + \sum_{j=1}^{6} \alpha_{ij} \ln y_j + \sum_{r=1}^{6} \gamma_{ri} \ln w_r + \rho_i m(\mathbf{d}) \quad (19)$$

where, rs_i is the revenue share of output i.

The homogeneity property of the cost function implies the following parametric restriction:

$$\sum_{r=1}^{4} \beta_r = 1, \ \sum_{s=1}^{4} \beta_{rs} = 0, \ \sum_{r=1}^{4} \sigma_r = 1 \quad (20)$$

Cost shares have to add-up to 1. This property implies the parametric restrictions in (20) plus the following parametric restriction:

$$\sum_{r=1}^{4} \gamma_{ri} = 0 \quad (21)$$

The system of equations formed by expressions (17) to (19) was estimated using data from a sample of 311 cereal farms in Italy. Data come from ISMEA, a socioeconomic survey of Italian agriculture designed on the basis of a theoretical model (Caiumi and Perali, 1997).

The outputs produced by the farms are wheat, durum wheat, corn, other grains, forage and other crops. The inputs used are family labor, land and structure, capital and materials. We include variables in the analysis that control for the location of the farm in the country: north west, northeast, center and south. We also assume that land is fixed in the short run. As a consequence we treat land size as an exogenous attribute of the farm. We model the farmer's decision related to the best combination of outputs to adopt. We expect that our

approach towards modeling heterogeneity can properly take into account the fact that some of the outputs are not chosen because of physical constraints. As an example, durum wheat is not produced in the North because is not fit to the local meteorological conditions; while in the south both options are feasible. This should be captured both by the structure of our likelihood function and by the design of the matrix of correlation coefficients in the two step approach. Given the initial stage of the research, we maintain a level of aggregation across inputs sufficient to avoid truncation in the input side. However, both truncated output decisions and continuous input choices are made jointly. Not all farms in the sample produce all outputs. In fact, as shown in table 1, there are no farms producing some of all outputs. As a result, we have a system of equations with several censored endogenous variables that we estimated using the two methods outlined in section 3. The results obtained until present are encouraging. We deem that it is still too early to present them formally. So far, we know that both approaches are feasible, robust and we can add great flexibility to our model specification. The models compare well under a statistical point of view, so we expect to have similar power in terms of economic reasonableness and prediction.

Maximum Simulated Likelihood is a time consuming method even with today computers as the size of either the sample or the system of equations gets relatively large. The two step estimator of the system of tobit equations, on the other hand, is not constrained by either sample or system size. As an example, based on our experience, it is difficult to fit in a computer with 384 Mb of RAM a sample larger than about 3000 observations within the estimation of a system of seven equations (that is, about 21000stacked observations)even with no truncation. Considering that the MSL procedure adopted here runs one observation at the time, the problem can become very cumbersome.

Tab. 1. Descriptive Statistics - Cereal Sector of the Italian Agriculture - 1995 ISMEA data

Variable Name	Label	Mean	Std.Dev.	Min	Max
s_wheat	output revenue share wheat	0.053	0.135	0.000	1.317
s_durwe	output revenue share durum wheat	0.227	0.388	0.000	2.815
s_corn	output revenue share corn	0.195	0.356	0.000	2.080
s_grains	output revenue share other gr	0.135	0.361	0.000	2.696
s_forage	output revenue share forage	0.145	0.428	0.000	5.149
s_others	output revenue share other	0.212	0.285	0.000	2.554
lq_wheat	log quantity wheat	1.477	2.443	0.000	7.090
lq_durwe	log quantity durum wheat	2.682	2.909	0.000	8.006
lq_corn	log quantity corn	2.636	3.209	0.000	8.875
lq_grain	log quantity grain	2.449	2.879	0.000	9.616
lq_forag	log quantity forage	2.800	3.198	0.000	10.393
lq_other	log quantity other	4.070	3.205	0.000	9.687
s_lf	input share family labor	0.579	0.218	0.014	0.949
s_t	input share land and structure	0.117	0.083	0.004	0.522
s_k	input share capital	0.040	0.037	0.000	0.214
s_m	input share all materials	0.231	0.145	0.014	0.795
rlh_wagf	log family wage per hour (000 Lire)	2.494	0.180	2.020	3.461
rlh_land	log price of land&structure / labor unit (000 lire)	0.619	1.081	-2.543	3.689
rlh_cap	log price of capital / labor unit (000 lire)	-0.620	1.324	-6.410	3.462
rlh_oth	log price of material / labor unit (000 lire)	1.346	1.099	-1.787	4.966
l_tc	log total cost	11.144	0.689	8.776	14.733
No	North - West	0.261	0.440	0.000	1.000
En	North - East	0.138	0.346	0.000	1.000
Centro	Center	0.238	0.427	0.000	1.000
Sud	South	0.267	0.443	0.000	1.000

5. References

Blundell, R., Pashardes, S., and G. Weber: "What Do We Learn about Consumer Demand Patterns from Micro Data?," *American Economic Review*, 83, (1993): 570-597.

Browning, M., and C. Meghir: "The Effects of Male and Female Labor Supply on Commodity Demands," *Econometrica*, 59 (1991): 925-951.

Börsh-Saupan, A. and V. Hajivassiliou. "Smooth Unbiased Multivariate Probability Simulators for

Maximum Likelihood Estimation of Limited Dependent Variable Models," *Journal of Econometrics*, 58(1993):347-68.

Caiumi, A. and F. Perali : "Female Labor Force Participation: A comparison between Urban and Rural Households in Italy," *American Journal of Agricultural Economics*, 79, (1997), 595-601.

Geweke, J. "Efficient Simulation from the Multivariate Normal Distribution Subject to Linear Inequality Constraints and the Evaluation of Constraint Probabilities," Duke University, Durham, N.C: Mimeo, (1989).

Hajivassiliou, V., D. McFadden and P. Ruud. "Simulation of Multivariate Normal Rectangle Probabilities and their Derivatives: Theoretical and Computational Results," *Journal of Econometrics,* 72 (1996):85-134.

Huffman, W. E. "An Econometric Methodology for Multiple-Output Agricultural Technology: An Application of Endogenous Switching Models," in *Agricultural Productivity Measurement and Explanation* ed. Susan M. Capalbo and Antle, J., Resources for the Future, the Johns Hopkins University Press, (1988).

Lee, L.F. and M. M. Pitt. "Microeconometric Models of Concumer and Producer Decisions with Limited Dependent Variables," University of Minnesota, Department of Economics: Discussion Paper, (1984).

Lee, L.F. and M. M. Pitt. "Microeconometric Demand Systems with Binding Nonnegativity Constraints: The Dual Approach," *Econometrica,* 54(1986):1237-42.

Lerman, S. and C. Manski. "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in: *Structural Analysis of Discrete Data with Econometric Applications*. Edited by Manski C.F. D. and McFadden. Cambridge-Massachusetts: The MIT press, (1981).

Lewbel, A. "A Unified Approach to incorporating demographic or other effects into demand systems", *Review of Economic Studies*, 70 (1985): 1-18.

Phaneuf, D. J. "A Dual Approach to Modeling Corner Solutions in Recreation Demand," *Journal of Environmental Economics and Management*, 37 (1999): 85-105.
Ross, S.M., *A First Course in Probability*, Macmillan, New York, (1988).

Stern, S. "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models," *Econometrica*, 60 (1992): 943-52.

Wales, T.J. and A. D. Woodland. "Estimation of Consumer Demand Systems with Binding Non-Negative Constraints," *Journal of Econometrics*, 21(1983):263-85.