Available online at www.sciencedirect.com

SCIENCE DIRECT®

Agricultural Economics 31 (2004) 97–106

AGRICULTURAL ECONOMICS

www.elsevier.com/locate/agecon

A revised Tobit procedure for mitigating bias in the presence of non-zero censoring with an application to milk-market participation in the Ethiopian highlands

Garth Holloway [a,*], Charles Nicholson [b], Chris Delgado [c], Steve Staal [d], Simeon Ehui [e]

[a] *Reader in Agricultural Economics and Statistics, Department of Agricultural and Food Economics, University of Reading, 4 Earley Gate, Whiteknights Road, P.O. Box 237, Reading RG6 6AR, UK*

[b] *Senior Research Associate, Department of Applied Economics and Management, Cornell University, Ithaca, New York, USA*

[c] *Senior Research Fellow, International Food Policy Research Institute and International Livestock Research Institute, Washington, DC, USA*

[d] *Economist, Livestock Policy Analysis Program, International Livestock Research Institute, Nairobi, Kenya*

[e] *Economist, The World Bank, Washington, DC, USA*

Received 10 October 2002; received in revised form 10 January 2003; accepted 11 December 2003

Abstract

Fixed transactions costs that prohibit exchange engender bias in supply analysis due to censoring of the sample observations. The associated bias in conventional regression procedures applied to censored data and the construction of robust methods for mitigating bias have been preoccupations of applied economists since Tobin [Econometrica 26 (1958) 24]. This literature assumes that the true point of censoring in the data is zero and, when this is not the case, imparts a bias to parameter estimates of the censored regression model. We conjecture that this bias can be significant; affirm this from experiments; and suggest techniques for mitigating this bias using Bayesian procedures. The bias-mitigating procedures are based on modifications of the key step that facilitates Bayesian estimation of the censored regression model; are easy to implement; work well in both small and large samples; and lead to significantly improved inference in the censored regression model. These findings are important in light of the widespread use of the zero-censored Tobit regression and we investigate their consequences using data on milk-market participation in the Ethiopian highlands.

JEL classification: C24; O12; Q01; Q12

Keywords: Revised Tobit procedure; Milk-market development

1. Introduction

Often, non-negligible fixed costs are associated with market transactions and for some economic agents these costs prohibit exchange. When fixed costs are prohibitive, conventional supply analysis contains bias due to censoring of the sample observations. The associated bias in conventional regression procedures and the construction of robust methods for mitigating such bias have been preoccupations of applied economists since Tobin (1958). However, bias resulting from incorrectly assuming that the true point of censoring in the sample is zero appears to have gone largely unnoticed. The objective of this

* Corresponding author. Tel.: +44 118 378 6775; fax: +44 118 975 6467.
E-mail address: garth.holloway@reading.ac.uk (G. Holloway).

0169-5150/$ – see front matter
doi:10.1016/j.agecon.2003.12.001

paper is to study this second source of bias and propose procedures for mitigating it. Ignoring the true point of censoring imparts significant inaccuracy to estimates of censored regression parameters and we present three alternative techniques for mitigating this bias using Bayesian procedures. The bias-mitigating procedures are based on modifications of the key step that facilitates Bayesian estimation of the censored regression model. These procedures are straightforward to implement; work well in both small and large samples; and lead to significantly improved inference in the censored regression model. These findings are important in light of the widespread application of the Tobit regression in which the zero-censoring assumption is applied. The preferred and conventional procedures are compared and contrasted in an application to milk-market participation in the Ethiopian highlands.

2. Conventional, censored regression and three alternative procedures

Greene (1993, p. 691) lists a number of diverse situations in which the Tobit model has been applied, including household purchasing decisions, extramarital behaviour, labour supply and criminal activity. In agricultural economics, important classes of applications involve commodity supply decisions (Mundlak, 2002), disequilibrium models (Fair and Jaffee, 1972), production economics (Paris, 1992), and development economics (Goetz, 1992). In the latter category a key interest lies in expanding the density of market participation (Stiglitz, 1989). The agricultural economics literature is replete with examples in which the censored regression has been usefully adapted. Recent examples that have appeared in this *Journal* include Woldehanna *et al*. (2000), Angulo *et al*. (2001) and Kosarek *et al*. (2001), to mention a few. The basic model structure which this literature applies is

$$z_i = x_i\beta + \varepsilon_i, \tag{1}$$

where z_i denotes a latent economic quantity of interest; $x_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ik})$ denotes a vector of characteristics associated with the latent z_i; $\beta \equiv (\beta_1, \beta_2, \ldots, \beta_k)'$ depicts the relationship between the characteristics and the latent z_i; ε_i denotes random error, which, we assume, is normally distributed with mean zero and variance σ^2, that is, $\varepsilon_i \sim N(0, \sigma^2)$; and for each $i = 1, 2, \ldots, N$, we observe

$$y_i = \max\{z_i, 0\}. \tag{2}$$

Eqs. (1) and (2) comprise the standard Tobit model in which the point of censoring—henceforth, π is assumed to be zero. The situation we are interested in is when Eq. (1) is applied but, instead of Eq. (2) governing the censoring of the data, they are instead governed by the rule

$$y_i = z_i \quad \text{if } y_i \geq \pi \text{ and } y_i = 0, \text{ otherwise}. \tag{3}$$

We refer to Eqs. (1) and (2) as the *conventional model* and refer to Eqs. (1) and (3) as the *true model*. The two models are, of course, identical when π equals zero and, although it may be possible to infer the exact point of censoring in rare situations, usually the value of π will not be known *a priori*.

When π is random and unknown, three principal issues arise. The first issue is the magnitude of bias arising from the practice of incorrectly assuming that π equals zero when, in fact, π is greater than zero. The second issue is the derivation of procedures that may mitigate any bias, and the third issue is the gainful understanding of the economic implications of the bias.

Even when the censoring point is actually known, ordinary least squares applied to the censored data leads to biased and inefficient parameter estimates (Tobin, 1958). Classical procedures for correcting this bias rely on one of two approaches. A first approach is to correct, using iterative procedures, the bias in least-squares estimation. A second approach relies on maximum-likelihood techniques and local approximation, for example Newton–Raphson. As an alternative, following seminal work on the Bayesian censored regression (Chib, 1992), we employ a data-augmented, Gibbs-sampling algorithm to simulate draws for the parameters from their intractable joint posterior distribution. In Chib (1992) this approach is compared to more conventional procedures and is shown to lead to accurate estimates of regression parameters.

In the remainder of the paper we pursue the Bayesian approach to estimation. There are four main justifications. First, as demonstrated in Chib (1992), the non-informative Bayesian and sampling theory approaches lead to estimates that are very similar, both in terms of their locations and their scales. Second, the Bayesian approach is conceptually appealing and, we feel, somewhat simpler to implement. Third, in view of the paucity of applications in agricultural economics exploiting the Bayesian approach, our demonstration has the ancillary appeal of highlighting the power of an under-exploited technique in solving a problem with a considerable heritage in agriculture. Fourth, the development of alternatives to the traditional zero-censored regression relies on an idea embedded in the crucial step enabling application of the Gibbs sampler to the censored regression model.

Gibbs sampling (Gelfand and Smith, 1990) and data augmentation (Tanner and Wong, 1987) are parts of a broader set of techniques in Bayesian inference known as Markov-chain, Monte-Carlo (MCMC) methods.[1] Examples of their application to censored-, discrete-, and truncated-regression problems are Albert and Chib (1993), George and McCulloch (1993) and Dorfman (1996, 1997, 1998).

Because the ideas underlying improvements to the traditional approach rely on an understanding of the Gibbs algorithm, it is useful to examine its application in general terms. Note that the censored regression framework in Eqs. (1) and (2) can, instead, be written

$$\mathbf{z} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\mathrm{i}}, \tag{4}$$

where $\mathbf{z}\equiv(z_1, z_2, \ldots, z_N)'$; $\mathbf{x}\equiv(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$, $\mathbf{x_1}\equiv(x_{11}, x_{21}, \ldots, x_{N1})'$, $\mathbf{x_2}\equiv(x_{12}, x_{22}, \ldots, x_{N2})'$, $\ldots, \mathbf{x_K}\equiv(x_{1K}, x_{2K}, \ldots, x_{NK})'$; $\boldsymbol{\beta}\equiv(\beta_1, \beta_2, \ldots, \beta_K)'$; $\boldsymbol{\varepsilon}\equiv(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)' \sim N(\mathbf{0}_N, \sigma^2\mathbf{I}_N)$; and we observe $y_i = \max\{z_i, 0\}$. Ordering the data so that the first N_1 observations are the observed, positive, quantities and the remaining $N_2(= N–N_1)$ observations correspond to the censored data, we write $\mathbf{y} \equiv (\mathbf{y_1}', \mathbf{y_2}')'$, where $y_1 \equiv (y_1, y_2, \ldots, y_{N1})'$ are the observations associated with positive quantities and $\mathbf{y_2} \equiv (y_{N1+1}, y_{N1+2}, \ldots, y_N)'$ are the censored data. Also, in $\mathbf{z} \equiv (\mathbf{z_1}', \mathbf{z_2}')'$, $\mathbf{z_1} \equiv (z_1, z_2, \ldots, y_{N1})' = \mathbf{y_1}$ and $\mathbf{z_2} \equiv (z_{N1+1}, z_{N1+2}, \ldots, z_N)' \neq \mathbf{y_2} = (0, 0, \ldots, 0)'$. The interchange $\mathbf{z_1} = \mathbf{y_1}$ is purely for notational convenience.

Simplification of the model is now made possible by working with the latent $\mathbf{z}_2$ rather than the observed $\mathbf{y}_2$. The essential recognition (Chib, 1992, p. 88) is that the data-augmented posterior distributions conditioned by the complete data ($\mathbf{y}$), and the observed data ($\mathbf{y}_1$), converge in distribution. The former distribution is difficult to work with because it involves censoring, but no censoring is involved in the latter formulation, making it easy to characterise in terms of its fully conditional component forms. Conditioned by the regression parameters, the latent, dependent variable has a normal distribution, truncated to be negative. Conditioned by the latent data and the error standard deviation, the regression parameters have a normal distribution; and, conditioned by the data and the regression parameters, the error standard deviation has an inverse gamma distribution.

The MCMC approach to estimation samples sequentially from these three sets of fully conditional distributions and, in so doing, simulates draws from the marginal posterior densities of interest. The algorithm is implemented as follows:

$$\begin{array}{ll}
\text{Step 1}: & \text{Select starting values for the regression}\\
& \text{coefficients and the error variance.}\\
\text{Step 2}: & \text{For each of the observations in the}\\
& \text{censored part of the data draw a normal}\\
& \text{random variable truncated according to the}\\
& \text{appropriate censor value.}\\
\text{Step 3}: & \text{Draw the regression coefficients from a}\\
& \text{multivariate normal distribution.}\\
\text{Step 4}: & \text{Draw the error variance from an inverse}\\
& \text{gamma distribution.}\\
\text{Step 5}: & \text{Repeat steps 1–4 for a ‘burn-in’ phase until}\\
& \text{convergence is achieved.}\\
\text{Step 6}: & \text{Repeat steps 1–4 and collect the outputs}\\
& \text{of the respective draws.}
\end{array} \tag{5}$$

Details of the distributions in question are contained in Appendix A. An important feature of this approach is that the outputs in the last step can be used to plot histograms, compute means and standard deviations, or estimate any desired posterior characteristics of interest (Gelfand and Smith, 1990).

An important step in the revised procedure involves specification of the sampling interval for the censor value, π. We develop three alternative approaches to constraining the choice of π. To do so, it is useful to denote the censored observations through a generic symbol, and we use $\mathbf{c} \equiv \{i,\ y_i = 0\}$ to denote the censor set.

[1] Readable introductions to Gibbs sampling, data augmentation and the Metropolis–Hastings method (of which Gibbs sampling is a special case) are Casella and George (1992) and Tanner (1993) and Chib and Greenberg (1995).

2.1. Alternative one

The first approach to the censored regression relies on the logic that the *minimum* of the observed supply values defines a *maximum* for the censor value. In particular, an upper bound on π is the minimum of the strictly positive sample quantities, or, the minimum of the set $\{y_i, i \notin \mathbf{c}\}$. By similar reasoning, because the observed net supplies can never be negative, a logical lower bound on π is zero. In other words, logic constrains the feasible choice for π to the closed interval

$$\pi \in [0, \min\{y_i, i \notin \mathbf{c}\}]. \tag{6}$$

This interval provides the basis for an estimation algorithm in which the true point of censoring is permitted to vary, but vary only within the range of values that lie below the minimum of the observed, positive quantities. Implementation requires three modifications to the algorithm in Eq. (5). First, a starting value for the true point of censoring, π, must be added in the first step. We recommend using the minimum of the uncensored observations as this value. Second, instead of zero, the draws for the latent data are now truncated to be less than the value π. Third, a draw for π is appended as an additional step in the algorithm. In the absence of additional information, we recommend making this draw from a uniform distribution.

2.2. Alternative two

Two problems arise with the interval in Eq. (6). First, estimates of π, although improved from assuming $\pi = 0$, are likely to be quite imprecise in the event that the true censor value is large. Second, neither the upper nor the lower bound for π in Eq. (6) is tied in any way to estimates of the Tobit regression. A second interval for estimating π arises from considering probit estimation on the zero-one, discrete outcome data. The rationale for this approach is that no censoring is involved in probit regression. The typical probit regression estimates a linear relationship between a truncated-normal random variable and a linear combination of the covariates, including a constant term. The Bayesian approach to estimation is outlined in Albert and Chib (1993, Eqs. (3)–(6), p. 671) and is almost identical to the Tobit algorithm in Eq. (5). There are two differences. First, due to the well-known problem that the probit regression is identified only up to a scale-normalized transformation of the linear function (see, for example, Greene, 1993, Section 21.3.2, p. 642), a parameter restriction must be imposed on the latent-variable regression. The typical restriction is to peg the error variance at a specified value, usually assumed to be one. Second, latent quantities are estimated for both partitions of the data, with the set of latent quantities pertaining to non-participants constrained to be negative and the remaining quantities pertaining to the participating observations constrained to be positive. For later reference, let $\{v_i \leq 0, i \in \mathbf{c}\}$ denote the latent quantities corresponding to the censored observations and let $\{v_i \geq 0, i \notin \mathbf{c}\}$ denote those corresponding to the uncensored observations. Reconsidering both of these aspects of probit estimation leads ultimately to an improved range for the censor value, π.

If instead of a forced separation at zero, the latent quantities could be transformed to produce a separation point endogenously, as part of the estimation itself, this separation point would provide a natural alternative to the bounds presented in Eq. (6). Using this fact, we derive estimates of the censor value by forcing the probit draws to mimic the actual observed data. However, in order to do so we require the latent data to possess the same error variance as the data in the Tobit regression. With this variance modification at hand, regressing the probit model through the origin, using the Tobit variance as its fixed scale, we derive a potentially improved range for the censor value, namely the interval

$$\pi \in [\max\{v_i, i \in \mathbf{c}\}, \min\{y_i, i \notin \mathbf{c}\}]. \tag{7}$$

Comparing the intervals in Eqs. (6) and (7), the latter interval is suspected to provide possible improvement due to the fact that the lower bound in Eq. (7) is estimated from the given covariates. Conceivably, this range could result in a latent draw $\max\{v_i, i \in \mathbf{c}\}$ that is negative, but this will only be the case when the estimate of the propensity to participate by the agent 'closest' to participation (that is, the largest latent value) is negative. In this case the interval defined by Eq. (7) provides less precise information than the one in Eq. (6). However, such an occurrence is unlikely to arise whenever fixed cost constrains participation by making positive latent quantities economically infeasible.

2.3. Alternative three

The link to the Tobit regression in Eq. (7) is an important one, but it is indirect. A third approach that provides a more direct linkage is to use the maximum values of the latent draws for the censored observations—in other words, draws from the Tobit regression itself. This approach is a natural extension of the logic developed for the use of the probit latent variables and consists of drawing values for π from the interval

$$\pi \in [\max\{z_i, i \in \mathbf{c}\}, \min\{y_i, i \notin \mathbf{c}\}]. \tag{8}$$

This approach links the interval for the censor value directly to the Tobit regression and, thus, eliminates the need for probit estimation.

2.4. Experimental evidence

With three alternatives to the conventional Tobit available, there appears to be considerable scope for deriving 'improvements' over the traditional model and a possibility for identifying a 'preferred procedure' from the pool of available alternatives. These questions are pursued in the context of some fairly comprehensive experiments using simulated data and a wide range of censor values. Space limits reporting the experiments and their results (details are available upon request), except to say that two features of the exercise are particularly noteworthy. First, some fairly definitive conclusions emerge from the experiments, including a consistent ranking among the three alternatives. Compared to the conventional Tobit regression, in which the censor value is assumed to be zero; the first, alternative procedure (Eq. (6)) generates considerable improvements in estimation accuracy; compared to the first alternative, estimation accuracy is further enhanced by combining the probit and Tobit models (Eq. (7)); and a further improvement is evident when the draws for the censor value are restricted by the latent data generated solely from the Tobit regression (Eq. (8)). Second, these experiments suggest a clear candidate for comparison with the traditional model and raise considerable scope for empirical enquiry. Thus, the interval in Eq. (8) is the one applied in the empirical application that follows.

3. Empirical application

As noted earlier, assessing factors influencing market participation in developing countries is a common application of the conventional model (Goetz, 1992). Where data on marketable surplus—that quantity of food product not consumed by the household itself—are available, a standard entry-analysing procedure regresses marketable surplus on a set of relevant household characteristics. Because nonparticipation is often at issue, some data are censored and Tobit estimation is relevant. The estimated regression is capable of identifying the subset of covariates that impact the entry decision and, also, predicting the levels of these covariates that are required for entry. Our concern lies in the extent to which the bias arising in this situation leads to significant biases in policy recommendations; leads, in turn, to false inferences about reform; and leads, therefore, to incorrect prescriptions for economic policy. The bias arises from a basic analogy to the theory of the firm: when fixed costs are relevant, there exist finite, non-negligible, quantities of marketable surplus (net supply quantities) below which household participation in the relevant market becomes infeasible.

Household-level data on milk sales in the Ethiopian highlands are used to compare the conventional model (that is, the zero-censored regression) and the true model (the Tobit regression that allows for a non-zero censor value). In the highlands, significant transactions costs prohibit entry for many households and recorded data on milk-market participation are therefore censored. Identifying the levels of covariates that influence the entry decision is one relevant objective, but principal interest lies in characterising the minimum efficient scale of operations for households to participate in the market.

Domestic dairy production has potential to generate income and employment on a large scale in the peri-urban areas of sub-Saharan Africa.[2] However, growth in dairying by small-holder farmers in peri-urban areas has been limited by transactions costs

[2] For the purpose of present discussion, the term 'peri-urban' defines those locations in geographical proximity to a major urban area (such as Addis Ababa) from which fluid milk and other dairy products are feasibly supplied to urban markets. The term is defined by Staal (1995).

inherent in production and marketing of dairy products. The nature of milk—perishable and bulky—and the possibility of transforming fluid milk into less perishable product forms, explain many of the transactions costs associated with the production and marketing of milk and dairy products in sub-Saharan Africa. Previous studies (see, for example, Debrah and Anteneh (1991)) provide evidence on the behavioural implications of transactions costs in dairying and draw attention to two features of the peri-urban structure in sub-Saharan markets. First, only a small proportion of African dairy production is marketed. Second, outlets for marketed surplus differ significantly in terms of a number of important structural features that seem to be related to quantity of throughput. These observations are indicative of features associated with relatively high transactions costs (Staal et al., 1997).

Fixed costs associated with fluid milk market entry include acquisition of productive resources, primarily cattle; search costs to determine the potential size and temporal variability of market outlets; and transport costs, including time spent traveling to market.[3] In general, these quantities are not markedly affected by the quantity of product sold. In the Ethiopian highlands, the informal fresh milk market involves direct delivery of raw milk by producers to consumers in the immediate neighborhood, as well as sales to itinerant traders or individuals in nearby towns (Debrah and Anteneh, 1991). However, fresh milk sales by small-holder farmers have historically been important only near formal milk marketing facilities such as government milk collection enterprises. Prior to the last 5 years, the vast majority of milk produced outside urban centers in Ethiopia was processed into cooking butter and a local variety of cottage cheese. Surpluses above household consumption were typically sold to traders or to other households in local markets.

Milk marketing cooperatives established by a development project in two regions of Ethiopia since 1996 have provided an additional outlet for fluid milk sales by peri-urban producers. These so-called 'milk groups' buy milk from members and non-members, process it, and sell the derivative products to traders and local consumers. The groups have been promoted as 'organisational innovations' that encourage market participation through reduction of transactions costs in dairy marketing. Previous studies note the effects of fixed costs such as ownership of higher-producing crossbred dairy cows and time required for transporting milk to market on the decision by small-holder producers to sell milk to the groups (Holloway et al., 2000). They motivate a more detailed analysis of the role of fixed costs in market participation decisions by examining the implications of non-zero censoring of the sales data.

Data were collected from households in two 'peasant associations' in two regions of the Ethiopian highlands. In each region fluid milk sales occurred in a small proportion of the households. A sample of 36 households was selected in each group, stratified according to ownership of crossbred cows, selling activity, and the physical distance that the household resided from the milk group location (Table 1). During June 1997 baseline household surveys were administered to 72 households. From June 1997 to October 1997 data on milk allocation and marketing, and significant events occurring in the cattle herd were collected, including features such as births, deaths, purchases, sales, illness and cow feeding practices. This study focuses on the responses concerning milk allocation and marketing from a subset containing 68 households. Data on milk sales in the 7 days prior to three survey visits yield a total of $(68 \times 7 \times 3 =)$ 1428 observations. Because of the panel nature of the data, the opportunity arises to exploit a hierarchical formulation for the model and implement, what some may argue, is a more appropriate formulation for the empir-

Table 1
Characteristics of dairy-producing households by participation status

Covariate	Participation status	
	Yes	No
Number of crossbred cows	1.41 (0.99)	0.49 (0.69)
Number of local cows	1.26 (1.03)	1.42 (1.12)
Time to the milk group, minutes	35.16 (18.76)	45.53 (29.94)
Years schooling household head	1.96 (4.01)	1.90 (3.24)
Extension visits in previous year	3.19 (3.59)	0.78 (1.66)

[3] While it is natural to include the acquisition of land among those fixed resources that contribute to transactions costs, this inclusion is somewhat less appropriate in the Ethiopian context. In Ethiopia land endowments are assigned almost exclusively by the state. Although there is some rental and exchange of land, allocations are usually based on non-economic factors. The primary fixed resource is cattle—cross-bred cattle in particular.

ical model. However, one feature of the data prohibits a hierarchical treatment. This feature is that the only covariates exhibiting significance are the 'stock' variables that reference each household and, by definition, remain fixed at each of the temporal sample points. This feature leads, in turn, to a repeated-measures problem that complicates unnecessarily and detracts from the main objective of demonstrating the significance of relaxing the zero-value censor assumption. We circumvent this issue simply by ignoring the panel structure of the data. In this case, we treat the individual sales quantities as though they were independent and identically distributed draws from the same normal distribution. Alternative treatments that deal explicitly with the repeated-measures problem are unlikely to yield additional insights about the significance of relaxing the zero-censoring assumption.

4. Results

The results of the conventional and alternative estimation procedures applied to the Ethiopian data are reported in Table 2. In both formulations the endogenous variable is strongly dependent on cross-bred cow ownership; local-breed cow ownership; minutes return time to walk bucketed fluid milk to the milk cooperative; and extension visitation (which is the number of times in the preceding 12 months that the household was visited by an extension agent discussing production and marketing techniques). The results are less significantly dependent on education. Sales are affected positively by the two livestock variables, schooling and extension visitation; and are affected negatively by the distance covariate. In addition, both models demonstrate an extremely good degree of fit, with most of the 168 non-censored observations and most of the 1260 censored observations predicted to lie within the correct, respective intervals.

Table 2
Posterior mean estimates under alternative censoring assumptions

	Assumption	
	$\pi = 0$	$\pi \neq 0$
Censor value	0.00	0.98 (0.02)
Number of crossbred cows	4.52 (0.33)	3.76 (0.27)
Number of local cows	1.87 (0.25)	1.54 (0.21)
Time to the milk group, minutes	−0.06 (0.01)	−0.05 (0.01)
Years schooling household head	0.29 (0.08)	0.21 (0.06)
Extension visits previous year	0.96 (0.10)	0.74 (0.09)
Constant	−12.67 (1.03)	−9.31 (0.89)
Square error of regression	5.34 (0.28)	4.20 (0.26)
Censored observations		
Predicted $\geq \pi$	27.34 (7.09)	29.02 (7.16)
Predicted $\leq \pi$	1232.66 (7.09)	1230.98 (7.16)
Uncensored observations		
Predicted $\geq \pi$	62.60 (2.85)	62.73 (2.83)
Predicted $\leq \pi$	105.40 (2.85)	105.27 (2.83)
Estimated additional milk requirements		
Mean requirement	10.04 (3.06)	8.02 (2.48)

Note: Numbers in parentheses are standard deviations in the Gibbs sample.

Considerable interest centers on the impacts of revising the conventional-model estimates with the revised procedure. Generally speaking, there are three noteworthy observations in the comparison between the two reports. First, the allowance for fixed costs through non-zero censoring appears to affect markedly the reports of the constant terms and the reports of the covariate coefficients. The constant term in the revised-Tobit formulation is significantly smaller in absolute value than the constant term in the traditional formulation and the coefficient estimates are also smaller in absolute value terms than those reported in the traditional formulation. Hence, third, relaxing the zero-censoring assumption appears to 'flatten' the estimated Tobit model, implying that an upwards bias exists in the response estimates in the traditional formulation. In short, ignoring fixed costs appears to lead to inflated Tobit regression parameter estimates.

These results are interesting in themselves, but additional interest lies in the impacts that these biases may have for policy prescriptions, notably those prescriptions centered on the estimation of minimum resource requirements—sometimes termed 'distances'—for market participation to occur. These reports are made in Fig. 1, where we plot the milk requirements across the 1260 'non-participants' estimated at each round of the Gibbs sequence. The distribution under the revised procedure has the higher mode and the distribution under the traditional procedure has the longer tail. The distribution under the traditional procedure is less compacted. Both its mean of 10.04 liters of milk per household per day and its standard deviation of 3.06 are greater than those obtained under the revised

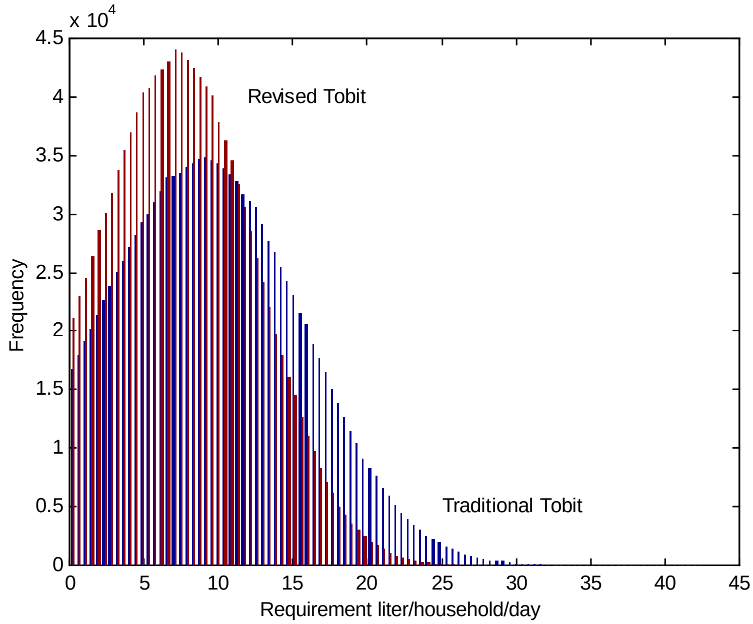

Fig. 1. Distribution of milk requirements under alternative censoring assumptions.

procedure, namely 8.02 and 2.48 liters per household per day, respectively. The difference between means represents an increase over the estimate obtained using the revised procedure in the order of about 25%, or approximately two liters of milk per household per day. Average daily supply among participating households is in the order of about 1.5 l. Hence, the difference between the two estimates is quite substantial. Consequently, we conclude that one impact of ignoring the non-zero censoring value in the Tobit regression is to significantly overstate estimates of market participation requirements.

5. Conclusions

In this paper we draw into question the conventional practice of assuming zero to be the true point of censoring in Tobit regression. When it is not the case, this assumption appears to impute significant bias to estimates of regression parameters. We consider the implications of this bias by applying the traditional and a revised procedure to a milk marketing example in which fixed costs affect participation and selling decisions. The estimates obtained from the two models—one in which the censor value is assumed to be zero and one in which the censor value is estimated endogenously from the data—are different. Generally speaking, applying the zero censoring assumption in the data appears to bias upwards absolute values of regression parameter estimates as well as reports of minimum resource levels required for market entry. The consequences for development policy are especially important: In devising policy prescriptions, restricting the true point of censoring to be zero may significantly overstate the levels of resources that are required for market entry.

The example presented in this paper is but one of a possibly vast set of applications in which generalising the censoring assumption may prove important. Further work should assess the robustness of our findings to a broader set of empirical contexts in which policy prescriptions derived from Tobit regression may be sensitive to the censoring assumption.

Acknowledgements

We are indebted to a large number of colleagues, too numerous to list individually, who shared detailed thoughts about the value of the revised procedures during the period that the manuscript itself was under revision. We are also grateful to the two anonymous referees for comments that led to improvements. Full responsibility for error remains with the authors.

Appendix A

Under a non-informative prior, the complete posterior distribution of the parameters (β, σ) and the latent data (z_2) has the following, conditional components (Zellner, 1996):

$$\begin{aligned} & z_i|\beta, \sigma \sim \text{truncated-normal}(\hat{z}_i, V_{z_i}), \quad i \in \mathbf{c}, \\ & \beta|\sigma, z \sim \text{normal}(\hat{\beta}, V_\beta), \\ & \sigma|z, \beta \sim \text{inverted-gamma}(\upsilon, s^2), \end{aligned} \tag{A.1}$$

where $\hat{z}_i \equiv x_i\beta$, $V_{z_i} \equiv \sigma^2$, $\hat{\beta} \equiv (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{z}$, $\mathbf{V}_\beta \equiv \sigma^2(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{z}$, $\upsilon \equiv N$ and $s^2 \equiv (\mathbf{z}\text{-}\mathbf{x}\beta)'(\mathbf{z}\text{-}\mathbf{x}\beta)/\upsilon$.

Details of the algorithm required to implement the Tobit regression are as follows:

$$\begin{aligned}
&\text{Step 1: Select starting values } \beta^{(s)} \text{ and } \sigma^{(s)}.\\
&\text{Step 2: For each } i \in \mathbf{c},\ \text{draw } z_i^{(s)} \sim \text{normal}(\hat{z}_i, V_{z_i}),\ \text{truncated so that } z_i^{(s)} \le 0 \text{ where } \hat{z}_i \text{ and } V_{z_i} \text{ are defined with}\\
&\qquad \text{respect to } \beta^{(s)} \text{ and } \sigma^{(s)} \text{ above and we define } \mathbf{z}_2^{(s)} \equiv \{z_i^{(s)}, i \in \mathbf{c}\} \text{ and, correspondingly, } \mathbf{z}^{(s)} \equiv (\mathbf{z_1}, \mathbf{z_2}^{(s)}).\\
&\text{Step 3: Draw } \beta^{(s+1)} \sim \text{normal}(\hat{\beta}, V_\beta),\ \text{where } \hat{\beta} \text{ and } V_\beta \text{ are defined with respect to } \mathbf{z}_2^{(s)} \text{ and } \sigma^{(s)},\ \text{above.}\\
&\text{Step 4: Draw } \sigma^{(s+1)} \text{ inverted-gamma}(\upsilon, s^2),\ \text{where } s^2 \text{ is defined with respect to } \mathbf{z}^{(s)} \text{ and } \beta^{(s+1)},\ \text{above.}\\
&\text{Step 5: Repeat steps 2–4 until } s \text{ equals some predetermined number, say } S_1,\ \text{within which convergence is}\\
&\qquad \text{achieved.}\\
&\text{Step 6: Repeat steps 2–4 until } s \text{ equals some predetermined number, say } S_2,\ \text{and collect output}\\
&\{\mathbf{z}_2^{(s)} s = 1, 2, \ldots, S_2\}, \{\beta^{(s)} s = 1, 2, \ldots, S_2\} \text{ and } \{\sigma^{(s)} s = 1, 2, \ldots, S_2\}.
\end{aligned} \tag{A.2}$$

Appendix B

The algorithm above is easily modified to incorporate a draw for the censor value, π, by making three amendments. First, in step 1, add a starting value for $\pi^{(s)}$. Second, in step 2, draw the latent variables truncated so that $z_i^{(s)} \le \pi^{(s)}$, where $\pi^{(s)}$ denotes the draw in step 1. Third, append an additional step between steps 4 and 5, drawing $\pi^{(s+1)} \sim \text{uniform}(\underline{\pi}, \bar{\pi})$, where $\underline{\pi}$ and $\bar{\pi}$ are defined, alternatively, by text Eqs. (6)–(8). Finally, in step 6, above, collect additional output $\{\pi^{(s)}$ $s = 1, 2, \ldots, S_2\}$.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. 88, 669–679.
Angulo, A., Gil, J., Gracia, A., 2001. The demand for alcoholic beverages in Spain. Agric. Econ. 26, 71–84.
Casella, G., George, E., 1992. Explaining the Gibbs sampler. Am. Stat. 46, 167–174.
Chib, S., 1992. Bayes inference in the Tobit censored regression model. J. Econ. 52, 79–99.
Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. Am. Stat. 49, 327–335.
Debrah, S., Anteneh, B., 1991. Dairy Marketing in Ethiopia: Markets of First Sale and Producers' Marketing Patterns. ILCA Research Report 19. International Livestock Center For Africa, Addis Ababa, Ethiopia.
Dorfman, J.H., 1996. Modeling multiple adoption decisions in a joint framework. Am. J. Agric. Econ. 78, 547–557.
Dorfman, J.H., 1997. Bayesian Economics Through Numerical Methods: A Guide to Econometrics and Decision-Making with Prior Information. Springer, New York.
Dorfman, J.H., 1998. Bayesian composite qualitative forecasting: hog prices again. Am. J. Agric. Econ. 80, 543–551.
Fair, R., Jaffee, D., 1972. Methods of estimation for markets in disequilibrium. Econometrica 40, 497–514.
Gelfand, A., Smith, A., 1990. Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. 85, 972–985.
George, E., McCulloch, R., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88, 881–889.
Goetz, S., 1992. A selectivity model of household food marketing behavior in sub-Saharan Africa. Am. J. Agric. Econ. 74, 444–452.
Greene, W., 1993. Econometric Analysis, second ed. Macmillan, New York.
Holloway, G., Nicholson, C., Delgado, C., Ehui, S., Staal, S., 2000. Agroindustrialization through institutional innovation: transactions costs, cooperatives and milk-market development in the east African highlands. Agric. Econ. 23, 279–288.
Kosarek, J., Garcia, P., Morris, M., 2001. Factors explaining the diffusion of hybrid maize in Latin America and the Caribbean region. Agric. Econ. 26, 267–280.

Mundlak, Y., 2002. Production and supply: theory and estimation. In: Gardner, B., Rausser, G. (Eds.), Handbook of Agricultural Economics. Elsevier, Amsterdam.

Paris, Q., 1992. The von Liebig hypothesis. Am. J. Agric. Econ. 74, 1019–1028.

Staal, S., 1995. Peri-Urban Dairying and Public Policy in Ethiopia and Kenya: A Comparative Institutional and Economic Analysis, Ph.D. dissertation. University of Florida, Gainesville, USA.

Staal, S., Delgado, C., Nicholson, C., 1997. Smallholder dairying under transactions costs in east Africa. World Dev. 25, 779–794.

Stiglitz, J., 1989. Markets, market failures and development. Am. Econ. Rev. 79, 197–203.

Tanner, M., 1993. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer-Verlag, New York.

Tanner, M., Wong, W., 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82, 528–550.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24–36.

Woldehanna, T., Lansink, A., Peerlings, J., 2000. Off-farm work decisions on Dutch cash crop farms and the 1992 Agenda 2000 CAP Reforms. Agric. Econ. 22, 163.

Zellner, A., 1996. An Introduction to Bayesian Inference in Econometrics, Wiley classics library ed. Wiley, New York.