Journal of Econometrics 51 (1992) 79–99. North-Holland

Bayes inference in the Tobit censored regression model*

Siddhartha Chib
University of Missouri, Columbia, MO 65211, USA

Received December 1989, final version received September 1990

We consider the Bayes estimation of the Tobit censored regression model with normally distributed errors. A simple condition for the existence of posterior moments is provided. Suitable versions of Monte Carlo procedures based on symmetric multivariate-t distributions, and Laplacian approximations in a certain parametrization, are developed and illustrated. Ideas involving data augmentation and Gibbs sampling [cf. Tanner and Wong (1987) and Gelfand and Smith (1990)] are also developed. The methods are compared in two examples with diffuse priors, and various combinations of sample sizes and degrees of censoring.

1. Introduction

This paper is concerned with the Bayes estimation of the well-known censored regression model introduced by Tobin (1958). In that seminal work, Tobin shows that if the dependent variable in a regression model is censored, say from below at zero, and the data set contains several such observations, the method of least squares is no longer appropriate. Over the years a number of estimation procedures have been discussed [cf. Maddala (1983), Amemiya (1984, 1985), and Greene (1990)]. However, little or no discussion of Bayesian methods is available although Carriquiry et al. (1987) and Sweeting (1987) deal with a related model that arises in biomedical applications.

The purpose of this paper is two-fold: first to develop suitable Bayesian approaches, and second to compare the efficacy of the different methods

*An earlier version of the paper was presented at the conference on 'Statistical Multiple Integration' at Humboldt State University. Comments from Jim Albert, John Geweke, Walid Huwaidi, Dale Poirier, Arnold Zellner, as well as two anonymous referees and an Associate Editor have significantly improved the paper and are gratefully acknowledged.

available. Not unlike the Probit model discussed in Zellner and Rossi (1984), the Tobit likelihood, even one based on the normality assumption, precludes an analytical treatment. Posterior moments must, therefore, be computed by numerical methods that deal with integration in multiple dimensions. Monte Carlo integration with importance sampling that has been successfully applied in several recent Bayesian applications [cf. Kloek and Van Dijk (1978), Zellner and Rossi (1984), Geweke (1986, 1989)] is again of direct use in this problem.

Another numerical approach, recently developed by Tierney and Kadane (1986), Kass, Tierney, and Kadane (1989), with applications considered in Sweeting (1987), Albert (1988), and Poirier (1988), amongst others, converts the multiple integration problem into a maximization problem. The resulting approximations, referred to as Laplacian approximations because they involve the use of Laplace's method for integrals, are particularly suitable for the Tobit model, given that large samples are sometimes available and because the likelihood times the prior can be readily maximized in the Tobin reparametrization.

Finally, much of the difficulty of analysing the Tobit model, and actually other censored problems as we will report elsewhere, can be overcome by using the ideas of data augmentation [cf. Tanner and Wong (1987)]. The method involves augmenting the censored data by negative imputations, and then dealing with the posterior w.r.t. the complete data through the iterative Gibbs sampler. The Gibbs sampler, which is recently discussed in Gelfand and Smith (1990), combined with data augmentation, provides an easy way to produce posterior moments even when the degree of censoring is large, as in most applications in labor economics.

The paper is organized as follows. Section 2 contains preliminaries about the Tobit model, while section 3 presents a general result about the prior–posterior analysis. Monte Carlo integration and Laplacian approximations are discussed in section 4, while the data augmentation algorithm is developed in section 5. The final section illustrates the efficacy of the methods with two examples involving diffuse priors, and various combinations of sample sizes and degrees of censoring.

2. Model and other preliminaries

We are concerned with the standard Tobit model

$$y_i^* = x_i'\beta + \varepsilon_i, \qquad \varepsilon_i \sim \text{iid } \mathrm{N}(0, \tau^{-2}),$$

$$y_i = \max\{y_i^*, 0\}, \qquad i = 1, 2, \ldots, n, \tag{1}$$

where y_i^* is a latent random variable which is observed as y_i if it is positive,

and is otherwise observed as equal to zero. The regression structure is defined by the covariates, $x_i: k \times 1$ (a k-vector), and the parameter vector $\beta \in R^k$. The error, ε_i, is independent normal with mean zero and precision (inverse of the variance) $\tau^2 > 0$.

Suppose that a sample of size n is available in which n_0 observations are censored and $n_1 (= n - n_0)$ observations are positive. The likelihood function of β, τ^2, due to independence of the $\{y_i\}$, is then given by

$$l(\beta, \tau^2) = \prod_{i \in C} [1 - \Phi(x_i'\beta\tau)](2\pi)^{-n_1/2}(\tau^2)^{n_1/2} e^{-\tau^2 \|y_1 - X_1\beta\|^2/2}$$

$$\equiv l_0(\beta, \tau^2) l_1(\beta, \tau^2), \qquad \beta, \tau^2 \in \Theta = R^k \times (0, \infty), \tag{2}$$

where $C = \{j: y_j = 0,\ j = 1, \ldots, n\}$ is the index set for the censored data, Φ is the standard normal cdf, $y_1: n_1 \times 1$ are the nonzero observations, $X_1: n_1 \times k$ is the matrix of explanatory variables corresponding to y_1, $\|\cdot\|$ denotes the Euclidean norm, and l_0 and l_1 define the two distinct parts of the likelihood. The term l_0 is the joint probability of the censored data, given independence, and the fact that from (1), $\Pr(y_i = 0) = \Pr(y_i^* < 0) = 1 - \Phi(x_i'\beta\tau)$.

Maximum likelihood (ML) estimates of the unknown parameters β and τ^2 can be obtained via an iterative process such as the Newton–Raphson. The easiest way to avoid the cumbersome derivatives is to use the Tobin reparameterization $(\beta, \tau^2) \to (\alpha = \beta\tau,\ \tau = (\tau^2)^{1/2})$, find the ML estimates in the $\psi = (\alpha, \tau)$ space, and then transform back to the original parameterization. The derivatives of $L(\psi) = \log l(\alpha, \tau)$ required for this process are given by

$$L_\alpha = X_0'A_0 + X_1'(\tau y_1 - X_1\alpha),$$

$$L_\tau = n_1\tau^{-1} - y_1'(\tau y_1 - X_1\alpha),$$

$$L_{\psi\psi} = \operatorname{diag}(-X_0'B_0X_0, -n_1\tau^{-2}) - Z_1'Z_1, \tag{3}$$

where $X_0: n_0 \times k$ is the data on the censored observations, $Z_1 = (X_1, -Y_1)$, A_0 and B_0 are both $n_0 \times n_0$ diagonal matrices with typical element given by λ_i and $\lambda_i(\lambda_i - x_i'\alpha)$, respectively, where $\lambda_i = \phi_i/(1 - \Phi_i)$ is the hazard function of the normal distribution, and ϕ_i and Φ_i are the standard normal pdf and cdf evaluated at $x_i'\alpha$, $i \in C$. It is easy to show that each element of B_0 is positive. Letting $\hat{\alpha}_{\mathrm{ML}}$ and $\hat{\tau}_{\mathrm{ML}}$ denote the ML estimates, it follows from the invariance property that the ML estimates of the original parameters can be obtained as $\hat{\beta}_{\mathrm{ML}} \equiv \hat{\alpha}_{\mathrm{ML}}\hat{\tau}_{\mathrm{ML}}^{-1}$ and $\hat{\tau}_{\mathrm{ML}}^2 = (\hat{\tau}_{\mathrm{ML}})^2$, respectively. The sampling

variance of these estimates may, for example, be approximated by

$$\hat{\Omega}_{\mathrm{ML}} = \hat{J}\left[-L_{\psi\psi}(\hat{\alpha}_{\mathrm{ML}}, \hat{\tau}_{\mathrm{ML}})\right]^{-1}\hat{J}', \tag{4}$$

where $J: (k+1)\times(k+1)$ is the Jacobian matrix given by

$$J = \begin{bmatrix} \tau^{-1}I_k & -\alpha\tau^{-2} \\ 0 & 2\tau \end{bmatrix}$$

and $\hat{J}$ is J evaluated at $\hat{\psi}$.

3. A general result

Suppose that $\pi(\theta)$, a proper or improper probability density function, is used to denote available prior information about the random vector, $\theta = (\beta, \tau^2)$. Although it is not necessary to impose any special restrictions on the prior, it is convenient to work with densities that combine with the likelihood function of the uncensored observations. In that spirit, substantive prior information can be modeled through a normal-gamma distribution, similar to Zellner and Rossi (1984), where a normal prior is interpreted as defining the approximate conjugate prior for the approximate likelihood function. Likewise, diffuse prior information can be modeled by a uniform independent distribution for β and $\log\tau^2$, i.e., $\pi(\beta, \tau^2) \propto \tau^{-2}$, which can also be motivated by drawing an analogy to the uncensored case [cf. Sweeting (1987)].[1]

Our interest is in the posterior of θ and the posterior expectation of real-valued functions $g(\theta)$, defined by

$$\pi(\theta|y) = K^{-1}\pi(\theta)l(\theta),$$

$$\mathrm{E}(g(\theta)|y) = \int_{\Theta} g(\theta)\pi(\theta)l(\theta)\,\mathrm{d}\theta \Big/ \int_{\Theta} \pi(\theta)l(\theta)\,\mathrm{d}\theta, \tag{5}$$

where $K = \int_{\Theta}\pi(\theta)l(\theta)\,\mathrm{d}\theta$ is the normalizing constant. Examples of $g(\theta)$ include $g_1(\theta) = \beta_j[1 - \xi_i\phi(\xi_i)/\Phi(\xi_i) - \phi(\xi_i)^2/\Phi(\xi_i)^2]$ and $g_2(\theta) = \Phi(\xi_i)g_1(\theta) + \bar{y}_p g_3(\theta)$, where $\xi_i = x_i'\beta\tau$, $\bar{y}_p = x_i'\beta + \tau^{-1}\phi(\xi_i)/\Phi(\xi_i)$, and $g_3(\theta) = \phi(\xi_i)\beta_j\tau$. The first of these is $\partial\bar{y}_p/\partial x_{ij}$, the change in the expected value of y_i w.r.t. x_{ij} given that $y_i > 0$, and the second is $\partial\bar{y}_i/\partial x_{ij}$ [cf. McDonald and Moffit (1980)].

[1] From the information matrix [cf. Amemiya (1973, p. 1007)] Jeffreys' prior contains the extra term, $\det(X'DX)^{1/2}$, where D is diagonal with elements $\{\Phi_i + \phi_i\lambda_i - \phi_i x_i'\beta\tau\}$. This reduces to the uniform prior for β if D is approximately constant over the main region dictated by the likelihood function.

It is clear from (2) that, due to the term $l_0(\beta_1, \tau^2)$, the posterior in (5) will not produce analytical moments, for any prior. Nonetheless, it is possible to state a general result that yields conditions for the existence of posterior moments. The proof is straightforward and is omitted.

Theorem. Let $\pi(\theta)$ *be any prior pdf of* θ *satisfying the condition* $\int_\Theta \pi(\theta) l_1(\beta, \tau^2)\, d\beta\, d\tau^2 < \infty$. *Let* $q(\theta|y_1)$ *denote the posterior of* θ *given* y_1, *i.e.,* $q(\theta|y_1) \propto \pi(\theta) l_1(\theta)$. *Let* $q(\beta|y_1)$, $q(\tau^2|y_1)$, *and* $q(\beta|y_1, \tau^2)$ *denote the associated marginal and conditional pdfs. Then the posterior of* θ *given y exists, and*

$$\text{(i)} \qquad \pi(\theta|y) = l_0(\beta, \tau^2) q(\theta|y_1)/K,$$

$$\text{(ii)} \qquad \pi(\beta|y) = K(\beta) q(\beta|y_1)/K,$$

$$\text{(iii)} \qquad \pi(\tau^2|y) = K(\tau^2) q(\tau^2|y_1)/K, \tag{6}$$

where

$$K(\beta) = \int \pi(\tau^2) l(\theta)\, d\tau^2 / \int \pi(\tau^2) l_1(\theta)\, d\tau^2$$

and

$$K(\tau^2) = \int l_0(\theta) q(\beta|y_1, \tau^2)\, d\beta.$$

One consequence of this result is that, if the expectation of $g(\theta)$ under q exists, then so does its expectation under the full posterior as $\int_\Theta |g(\theta)| \pi(\theta|y)\, d\theta$ is bounded by $\int_\Theta |g(\theta)| q(\theta|y)\, d\theta$, given $l_0(\theta) \leq 1$ uniformly over Θ. In addition, the result shows how the posteriors based on the positive observations are modified by the censored data. Finally, if there is no censoring, $\pi(\theta|y) = q(\theta|y)$, and an exact analysis of the posterior is straightforward in contrast to the classical case where the exact sampling distribution of the ML estimates remains intractable.

4. Estimation by Monte Carlo integration and Laplace approximations

4.1. Monte Carlo integration with importance sampling

By far the most direct method of evaluating the integrals in (5) is through Monte Carlo integration with importance sampling, replacing those integrals by suitable sample averages as we describe next. Define the numerator and denominator of (5) as

$$\mathrm{E}(g(\theta)|y) \equiv \bar{g} = \eta/K,$$

where $g(\theta)$ is an integrable function of θ and $K > 0$. The key idea in Monte Carlo integration is that η and K can be written as $\int_\Theta w(\theta)g(\theta)h(\theta)\,d\theta$ and $\int_\Theta w(\theta)h(\theta)\,d\theta$, respectively, where $h(\theta)$ is the so-called importance density that is similar to $\pi(\theta|y)$ and from which synthetic draws of θ can be easily made, and $w(\theta) \equiv \pi(\theta)l(\theta)/h(\theta)$.[2] The integrals are estimated as follows.

Step 1: Simulate $\{\theta_i\}_1^N$, an iid sample from $h(\theta)$, where N is a large number, say 10^4.

Step 2: Calculate the weights $w(\theta_j) = \pi(\theta_j)l(\theta_j)/h(\theta_j)$, $j = 1, \ldots, N$.

Step 3: Estimate η by $\hat{\eta} = \sum[w(\theta_j)g(\theta_j)]/N$, K by $\hat{K} = \sum[w(\theta_j)]/N$, and $\bar{g}$ as $\bar{g}_N = \hat{\eta}/\hat{K}$, where the sum runs from 1 to N.

Step 4: Retain the estimate if $RNE \equiv \text{var}[g(\theta)|y]/(N \text{ var}^h[\bar{g}_N])$ is close to unity, where $\text{var}^h[\bar{g}_N] = N^{-1}K^{-2}\int_\Theta (g(\theta) - \bar{g})^2 w(\theta)^2 h(\theta)\,d\theta$ is the square of the *numerical standard error of the estimate*, up to order $O(N^{-2})$, where $O(N^{-2})$ is a term that converges to zero at the rate N^{-2}. Revise the importance density otherwise and go to step 1.

Now if we were to use (6), we can express $\bar{g}$ as the ratio of two expectations w.r.t. $q(\theta|y_1)$, i.e., $\int_\Theta g^*(\theta)q(\theta|y_1)\,d\theta / \int_\Theta l_0(\beta, \tau^2)q(\theta|y_1)\,d\theta$, where $g^*(\theta) = g(\theta)l_0(\beta, \tau^2)$. Does this imply that $q(\theta|y_1)$ can be used as an importance density? Unfortunately, the answer is in the negative. As Geweke (1989) and others have emphasized, the right importance density should mimic $\pi(\theta|y)$ with the ratio of the two densities essentially constant in θ. This condition is not satisfied by $q(\theta|y_1)$ as can be easily seen from (6).

There is another problem with this importance density that is best illustrated with the one-dimensional case ($k = 1$) where $\beta = \mu$, to avoid confusion, and τ^2 is fixed. Because large values of μ imply greater prior probability of a positive observation, the posterior based on only the uncensored data will be, a posteriori, concentrated at the mean, away from zero; the posterior probability that $\mu < 0$ under $q(\mu|y_1)$ will be underestimated. If we think of fixing the prior and the expected proportion of censored data, and let n become large, the match between $q(\mu|y_1)$ and $\pi(\mu|y)$ will further erode due to the inconsistency of the sample mean based on the positive data. As a result, for a fixed N, the variance of the estimate in (9) diverges as $n \to \infty$. Equivalently, to maintain a constant *RNE*, N must increase without bound as the sample size goes to infinity. Similar statements can be made for the general vector case, although now the configuration of the x's and the

[2]Typically the numerator of w is multiplied by an *arbitrary* constant, d [cf. Geweke (1989)]. If d is chosen to approximate K^{-1}, then the numerator and denominator are directly comparable, both being densities. Inserting this factor is critical; otherwise w goes to zero pointwise, as n becomes large. In our examples, d is the inverse of the maximum of the likelihood, although any other similar choice will suffice.

dependence with y will determine the subset of the parameter space in which q *given* y_1 is concentrated. This region will again imply low posterior probabilities of censoring and the same problems mentioned earlier.

It is therefore preferable to take the approach that is used in Zellner and Rossi (1984) and Geweke (1986), termed the *exact* Monte Carlo method. Under the usual boundedness conditions on the sampling process generating the covariates, if the sample size is large or if the prior is diffuse, the posterior of β given τ^2 will be approximately quadratic around the ML estimate, while that of τ^2 will be scaled chi-squared [cf. Sweeting (1987)]. From these approximations we can deduce that the multivariate-t density given by

$$h(\theta) = f_T\left(\theta | \hat{\theta}_{ML}, \hat{\Omega}_{ML}, \nu\right), \tag{7}$$

where $\hat{\theta}_{ML}$ is the ML estimate of θ, $\hat{\Omega}_{ML}$ is its covariance matrix, and ν is the degrees-of-freedom parameter chosen conservatively to produce thick tails is likely to be an adequate importance function.[3]

Another device that is interesting in the context of the large samples encountered in applications is a *bias-corrected* approach. Although the method is also anchored on (7), the idea now is to compute $\bar{g}_N$ with limited replications, say with $N = 50$, and then to correct the estimate by subtracting the bias of the estimate. The bias can be calculated by noting that both $\hat{\eta}$ and $\hat{K}$ are unbiased estimates of η and K, respectively, i.e., $E^h(\hat{\eta}) = \eta$ and $E^h(\hat{K}) = K$, where the expectation is w.r.t. the density $h(\theta)$. However, if we assume the regularity conditions under which $\text{var}^h[\bar{g}_N]$ exists, i.e., (1) $\text{var}^h(w(\theta)) < \infty$ and (2) $\text{var}^h(g(\theta)w(\theta)) < \infty$ [(1) is satisfied if $E^h(w(\theta)^2) < \infty$ and (2) if $E^h(g(\theta)^2 w(\theta)^2) < \infty$], then we show in appendix A through the delta method that $\bar{g}_N$ is biased and its finite-sample bias is given by

$$E^h[\bar{g}_N] - \bar{g} = -N^{-1}K^{-2}\int_\Theta (g(\theta) - \bar{g})(w(\theta) - K)$$

$$\times w(\theta)h(\theta)\,d\theta + O(N^{-2}). \tag{8}$$

This expression for the finite-sample bias is not specific to the Tobit model, and is generally unimportant if the number of replications is relatively large.

[3]Formally, the tails of the likelihood along the jth β axis behave as $C\exp(-\beta_j^2)[1 + o(1)]$ as $|\beta_j|$ becomes large, where C is some constant and the $o(1)$ term goes to zero, and reflects the order of l_0. By integrating over τ^2 the tails can be shown to behave like that of the Student-t density.

4.2. Laplacian approximations

In this subsection we discuss the use of Laplacian approximations [cf. Tierney and Kadane (1986)] to evaluate the expectation in (5). We propose that $\mathrm{E}[g(\theta)|y]$ be evaluated from the $\psi = (\alpha, \tau)$ parameterization, since the concavity of the log-likelihood $L(\psi)$ [cf. Olsen (1978)] ensures that the prior times likelihood will have a single interior mode (if one exists) as long as the prior is not excessively sharp. This regularity condition is, of course, sufficient for the proper application of Laplace's method. We should therefore deal with the expectation

$$\mathrm{E}[f(\psi)|y] = \int_{\Psi} f(\psi)\pi(\psi)l(\psi)\,\mathrm{d}\psi \Big/ \int_{\Psi} \pi(\psi)l(\psi)\,\mathrm{d}\psi, \tag{9}$$

where $\pi(\psi) = \pi(\theta)|J|^{1/2}$, J is the Jacobian of the transformation, and $\Psi \subseteq R^p$ is the parameter space. In the reference prior case $\pi(\psi) \propto \tau^{-1}$. Examples of $f(\psi)$ include $f(\psi) = \alpha_j/\tau$ which gives us β_j, $f(\psi) = (\tau)^2$ which gives τ^2, etc.

In its simplest form, Laplace's method states that under regularity conditions I and $\hat{I}$, defined by

$$I = \int_{\Psi} \exp(nh(\psi))\,\mathrm{d}\psi, \qquad \psi \in R^p,$$

and

$$\hat{I} = (2\pi/n)^{p/2}\left|\Sigma(\hat{\psi})\right|^{1/2} \exp\left(nh(\hat{\psi})\right), \tag{10}$$

differ by terms of order $\mathrm{O}(n^{-1})$, where $\hat{\psi} = \operatorname{argmax} h(\psi)$ and $\Sigma(\hat{\psi})$ is the *inverse* of the *negative* of the Hessian of h at $\hat{\psi}$. A double application of (10) to each of the integrals in (9) yields accurate estimates of the posterior moments. The method works as follows.

Step 1: Maximize $\tilde{L}(\psi) = [L(\psi) + \log \pi(\psi)]/n$, and obtain $\hat{\psi}$, the posterior mode; $\tilde{\Sigma}$, the inverse of the *negative* Hessian of $\tilde{L}$; and $\tilde{L}(\hat{\psi})$.

Step 2: Maximize $L^*(\psi) = \tilde{L}(\psi) + n^{-1}\log f(\psi)$, whenever $f(\cdot)$ is a positive function and the logarithm is defined, and obtain $\psi^* = \operatorname{argmax} L^*(\psi)$; Σ^*, the inverse of the *negative* Hessian of L^* at the maximum; and $L^*(\psi^*)$.

Step 3: Estimate the posterior moment in (9) by the *second-order approximation*

$$|\Sigma^*(\psi^*)|^{1/2} \exp(nL^*(\psi^*)) \Big/ \left|\tilde{\Sigma}(\hat{\psi})\right|^{1/2} \exp\left(n\tilde{L}(\hat{\psi})\right). \tag{11}$$

As shown in Kass, Tierney, and Kadane (1990), the relative error of the approximation is $O_p(n^{-2})$. The conditions that ensure their error calculations, the Laplace regularity conditions, are essentially the same as those imposed by Amemiya (1973) to obtain the asymptotic normality of the ML estimates, namely that $\tilde{L}(\psi) - \tilde{L}(\hat{\psi}_{ML})$ converges in probability uniformly over ψ to zero, and $\hat{\psi}_{ML} = \psi + O_p(n^{-1/2})$ as $n \to \infty$.

When $f(\psi)$ is not positive, step 2 is applied to $f(\psi) + c$, where c is a large positive constant which is then subtracted from the result. In other cases, dealing with $-f(\psi)$ solves the problem.

The approximation in (11) turns out to be extremely easy to apply to the Tobit model. If we assume the reference prior for ψ, then $\log \pi(\psi) = -\log \tau$ and the quantities $\hat{\psi}$ and $\tilde{H}(\hat{\psi})$ are obtained from (3) by simply replacing n_1 by $n_1 - 1$. For the numerator, depending on the function $f(\psi)$, simple modifications to the equations in (3) have to be made. For example, if $f(\psi) = \alpha_j/\tau$, $\alpha_j > 0$, then those equations are changed by (i) replacing n_1 by $n_1 - 2$, (ii) adding e_j: $k \times 1$ to L_α where e_j is a vector with jth component $1/\alpha_j$ and zero elsewhere, (iii) adding $-e_j e_j'$ to $L_{\alpha\alpha}$. The corresponding second-order approximation of the variance of a positive function is

$$V(f(\psi)|y) = \left\{\hat{E}\left(f(\psi)^2\right) - \left[\hat{E}\, f(\psi)\right]^2\right\}\left(1 + O_p(n^{-2})\right), \tag{12}$$

where $\hat{E}$ is the expectation approximation given in (11). Again, this approximation is easy to code at least for simple functions $f(\psi)$.

A case can also be made in our application for sometimes employing the simpler *first-order approximation* which yields a relative error of $O_p(n^{-1})$. If the derivatives of $\log f(\psi)$ are sufficiently complicated, the increase in error can be tolerated especially when the sample size is somewhat large. These approximations, which are derived by applying Laplace's method to integrals of the type $I = \int_\Psi f(\psi)\exp(nh(\psi))\,d\psi$, give

$$E(f(\psi)|y) = f(\hat{\psi})\left\{1 + O_p(n^{-1})\right\},$$

$$V(f(\psi)|y) = f_\psi(\hat{\psi})'\left(n^{-1}\tilde{\Sigma}(\hat{\psi})\right) f_\psi(\hat{\psi})\left\{1 + O_p(n^{-1})\right\}. \tag{13}$$

Finally we consider the computation of marginal posterior densities. The simplest cases concern the margins of ψ, for example, τ. Now letting $\tilde{L}(\alpha, \tau) = L(\alpha, \tau) + \log \pi(\alpha, \tau)$, $\hat{\alpha}(\tau) = \operatorname{argmax}_\alpha \tilde{L}(\alpha, \tau)$, $\tilde{\Sigma}(\hat{\psi}) = [-\tilde{L}_{\psi\psi}(\hat{\psi})]^{-1}$, and $\tilde{\Sigma}(\tau) = [-\tilde{L}_{\alpha\alpha}(\hat{\alpha}(\tau), \tau)]^{-1}$, the marginal density is [cf. Kass, Tierney, and Kadane (1989)]

$$\hat{\pi}(\tau|y) = c^{-1}|\tilde{\Sigma}(\tau)|^{1/2}\exp\left(\tilde{L}(\hat{\alpha}(\tau), \tau)\right), \tag{14}$$

where the normalizing constant $c = (2\pi)^{1/2}|\tilde{\Sigma}(\hat{\psi})|^{1/2}\exp(\tilde{L}(\hat{\psi}))$. The margins of a component of α, say α_j, can be found similarly. The marginal posterior of some nonlinear function of ψ, say $s(\psi)$, can be also obtained. Let $\hat{\psi}(\gamma) = \text{argmax}\ \tilde{L}(\psi)$ subject to $s(\psi) = \gamma$, s_ψ be the gradient of s evaluated at $\hat{\psi}(\gamma)$, and $\tilde{\Sigma}$ the inverse of the negative Hessian of $\tilde{L}(\psi)$ at $\hat{\psi}(\gamma)$. Then

$$\hat{\pi}(\gamma|y) = c^{-1}|\tilde{\Sigma}|^{1/2}\left(s'_\psi \tilde{\Sigma} s_\psi\right)^{-1/2}\exp\left(\tilde{L}\left(\hat{\psi}(\gamma)\right)\right), \tag{15}$$

where c is the normalizing constant given in (14). The latter result can be used to find the marginal posteriors of θ. If we let $s(\psi) = \alpha_j/\tau$, we get β_j. The maximum of $\tilde{L}(\psi)$ subject to $\alpha_j/\tau = \gamma$ is accomplished simply by substituting $\alpha_j = \gamma\tau$ into $\tilde{L}(\psi)$ and optimising over a reduced parameter space (cf. appendix B).

5. Data augmentation and Gibbs sampling

We now consider an altogether Monte Carlo approach that is based on the recently developed ideas in Tanner and Wong (1987) and Gelfand and Smith (1990).[4] The data augmentation strategy of the former and the Gibbs sampler of the latter can be combined to yield an elegant solution to the censored data problem. Results provided in Gelfand and Smith (1990) can be used to show that the resulting method possesses very appealing theoretical properties and that the joint density of the draws converges to the true density in the sup norm as $t \to \infty$, an improvement over the L^1 convergence achieved in the Tanner and Wong algorithm.

The essential idea is quite simple. Suppose that along with the censored observations y_0, we have available (hypothetically) the corresponding latent data, say $z: n_0 \times 1$. By definition, all components of z must be negative. It is clear that in this case $\pi(\theta|y, z) = \pi(\theta|y_1, z)$, i.e.,

$$\theta|y_0, y_1, z \overset{d}{=} \theta|y_1, z, \tag{16}$$

where $\overset{d}{=}$ denotes equality in distribution. For *any prior* that combines with the *complete* data likelihood, the distribution on the right-hand side of (16) can be calculated analytically since no censoring is involved. Although z is not observed, a method that is based on simulating z is available.[5] Note from (1) that given y and θ, $z = (z_i)$ is an independent collection of random

[4]The discussion in this section was stimulated by the comment of one of the referees that this avenue could be explored.

[5]Our approach involves simulating the censored data unlike the EM algorithm where the censored data is replaced by the estimates of their conditional expected values, $x'_i\beta + \tau\lambda_i$, $i \in C$, where now $\lambda_i = -\phi_i/(1 - \Phi_i)$.

variables distributed as truncated normal with support $(-\infty, 0)$ and pdf

$$f(z_i|y,\theta) = f_N\left(z_i|x_i'\beta, (\tau^2)^{-1}\right)\Big/\left(1-\Phi(x_i'\beta\tau)\right), \qquad -\infty < z_i < 0, \quad i \in C. \tag{17}$$

The important point is that the data-augmented posterior, $\pi(\theta|y, z)$, and the conditional pdf of the latent data, $f(z_i|y,\theta)$, are both available in tractable form. These conditional pdf's, the inputs for the Gibbs sampler, enable us to recursively simulate the entire posterior distribution of θ.

Let $y_z = (z', y_1')'$: $n \times 1$ be the collection with y_0 replaced by z, and let $\hat{\beta}_z = (X'X)^{-1}X'y_z$ be the corresponding OLS estimate. Also if $\pi(\beta, \tau^2) \propto \tau^{-2}$, then it follows from (16) that $\pi(\beta|y, z, \tau^2) = f_N(\beta|\hat{\beta}_z, \tau^{-2}(X'X)^{-1})$ and $\pi(\tau^2|y,\beta,z) = f_G(\tau^2|n/2, \|y_z - X\beta\|^2/2)$. If $\theta^{(0)} = (\beta^{(0)}, \tau^{2(0)})$ is the starting value of θ, then the Gibbs algorithm for the Tobit model is defined through the simulation of

$$\begin{aligned} z_i \quad & \text{from} \quad f\left(z_i|y,\theta^{(0)}\right), \qquad i \in C, \\ \beta^{(1)} \quad & \text{from} \quad \pi\left(\beta|y, z^{(1)}, \tau^{2(0)}\right), \\ \tau^{2(1)} \quad & \text{from} \quad \pi\left(\tau^2|y, z^{(1)}, \beta^{(1)}\right), \end{aligned} \tag{18}$$

where $z^{(1)}$ is the $n_0 \times 1$ vector of simulated z's.[6] After t cycles of the sequence (18), we obtain one simulated draw, $(z^{(t)}, \beta^{(t)}, \tau^{2(t)})$, from the joint distribution $(\theta, z)|y$. Repeating the above t cycles M times produces M iid draws $(z_j^{(t)}, \beta_j^{(t)}, \tau_j^{2(t)})$, $j = 1, \ldots, M$, from the posterior distribution. Practical implementation of (18) is easy even if the values of M and t are large. We point out that for the problem at hand, the ML estimate based on the data, $(0, y_1)$, is the most appropriate choice for $\theta^{(0)}$, the starting value of the iterations.

Based on the M generated draws, posterior moments of any function $g(\theta)$ are computed in the usual way as sample averages. The marginal posterior density of a component of θ, say β, is estimated by the finite mixture of normal densities

$$\hat{\pi}(\beta|y) = M^{-1}\sum_{j=1}^{M} \pi\left(\beta|y, z_j^{(t)}, \tau_j^{2(t)}\right), \tag{19}$$

where the pdf on the right-hand side is the conditional density of β, given y,

[6] To sample from the truncated normal in (17), a rejection method can be employed. A less costly one-for-one draw is $x_i'\beta + \tau^{-1}\Phi^{-1}(U\Phi(x_i'\beta\tau))$ where U is a distributed as standard uniform [cf. Devroye (1986)].

z, and τ^2 [cf. (18)]. In the same way, the marginal of τ^2 is approximated as the average of gamma densities, $\pi(\tau^2|y, \beta, z)$, over the simulated β and z.

6. Examples

In this section we report how the various methods work in the context of two examples chosen to reflect some of the main features that arise in the applications of this model. The choice of the prior pdf is mostly tangential to our focus and the prior is, therefore, assumed to be of negligible importance relative to the sample size. The effects of different combinations of sample size and degrees of censoring are investigated. Since the true data-generating process (DGP) is known, it becomes possible to observe the adequacy of the proposed methods. All the computations are undertaken with GAUSS on an 80386, 16 Mhz 3 MB personal computer.

Example 1. For samples of size $n = 100$ and $n = 1000$, we analyze models with 40% and 60% degrees of censoring. The DGP is given by

$$y_i^* = \beta_1 + \beta_2 x_{2i} + \varepsilon_i, \tag{20}$$

where $\beta = (-9, 1)'$ and $\beta = (-13, 1)'$ for 40% and 60%, respectively, and $\varepsilon_i \sim \mathrm{N}(0, (16)^{-1})$. Data on y in this model is generated by replicating the exogenous variables in Judge et al. (1988, pp. 800) in batches of 20 and using (20).

Example 2. Here the degrees of censoring are 50% and 70% and the samples are of sizes $n = 200$ and $n = 400$. The DGP is taken from Wales and Woodland (1980) and is given by

$$y_i^* = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \tag{21}$$

$$x_{2i} = \gamma_2 z_{2i} + \gamma_3 z_{3i} + u_i,$$

where x_{3i}, z_{2i}, and z_{3i} are iid Unif$(-1, 1)$ random variables, $u_i \sim \mathrm{N}(0, 1.312)$, $\gamma_2 = \gamma_3 = 1$, and the parameters of interest are given by $\beta = (-1.1854, 1, 10)'$ and $\beta = (-1.1854, 1, 1)'$ for 50% and 70% censoring, respectively, and $\tau^2 = 0.6428$.

For all eight data sets, as functions g to be estimated we consider $g(\theta) = \beta_j$, $g(\theta) = \tau^2$, $g(\theta) = \partial \bar{y}_p / \partial x_2$, $g(\theta) = \partial \bar{y} / \partial x_2$. In the ψ parameterization used in the Laplacian calculations, the functions are given in an obvious way as $f(\psi) = \alpha_j / \tau$, $f(\psi) = \tau^2$, etc. We take the prior to be $\pi(\theta) \propto \tau^{-2}$ and $\pi(\psi) \propto \tau^{-1}$ in the θ and ψ parameterization, respectively. For comparison, we also calculate the ML estimates of all the parameters by the

Newton–Raphson method. The convergence of both the ML and second-order Laplace estimates is achieved in no more than five iterations starting from the least squares estimates based on the noncensored observations. The parameter β_1, being negative in both examples, is estimated as $g(\theta) = -\beta_1$; the negative of the result is the estimate of β_1. This method works in our examples since all the posterior probability of β_1 is concentrated on the negative half-line. As far as the Gibbs sampler (GS) in (18) is concerned, we have let $M = 200$ and $t = 50$, although smaller values of t between 10 and 20 may also be adequate [cf. Gelfand et al. (1989)].

In the case of the exact Monte Carlo estimates, $N = 10{,}000$ draws are used from the importance densities discussed in section 3; the bias-corrected (BC) estimates are based on $N = 50$ draws. The exact and bias-corrected calculations are based on the multivariate-t importance pdf given in (7), where the degrees of freedom is chosen to ensure that *RNE* is close to unity and that no negative draws of τ^2 are obtained.

To provide an idea of the computational burden, execution times for the computations in table 1 for $n = 100$ are: MLE = 7 seconds, Laplace = 16 seconds, BC = 8 seconds, GS = 34 minutes, Exact = 26 minutes, and for $n = 1000$: MLE = 20 seconds, Laplace = 40 seconds, BC = 2 minutes, GS = 3 hours, Exact = 2.5 hours. Given the nature of the Gibbs algorithm, it must be pointed out that the time involved in the GS calculations is an increasing function of the degree of censoring. However, even though the GS and exact estimates are relatively time-intensive, they do not require difficult coding and all functions of interest can be evaluated in one program run.

Next, the marginal posterior densities are calculated for a few of the parameters using the Tierney–Kass–Kadane method. The GS and exact method yields densities that are quite similar to those from (14) and (15) and are hence not reported. The marginal pdf's are computed on a 60- to 70-point grid. The resulting function values are numerically integrated using a 10-point Gaussian quadrature rule, and it is determined that renormalization is not necessary in any of the cases considered. Maximum computation time for a single marginal pdf is about 10 minutes.

The posterior moments for functions of interest are reported in tables 1–4. We note that all the different methods, the Laplace, BC, GS, and exact Bayes, yield very good estimates which, for some parameters and for all levels of censoring, are quite close to the true values. The difference between the MLE and the posterior mean indicates a certain degree of skewness in the posterior distributions, although as the sample size increases the estimates become roughly similar. It should be noted that the posterior variance from the exact Bayes calculations and the Gibbs sampler are usually larger than those from the other methods. Thus, if one uses these approaches, one is not likely to make over-precise statements about the parameters. The numerical

Table 1

Example 1: MLE, Bayesian posterior mean and variance,[a] *40% censoring*.

	β_1 (−9)	β_2 (1)	τ^2 (0.0625)	$\partial \bar{y}_p/\partial X_2$ (0.453)[b]	$\partial \bar{y}/\partial X_2$ (0.646)[b]
			$n = 100$		
MLE[c]	−6.841	0.896	0.075	0.486	0.680
	(1.244)	(6.9×10^{-3})	(1.86×10^{-4})	(2.15×10^{-3})	(3.6×10^{-3})
Laplace[d]	−6.942	0.902	0.075	0.485	0.679
	(1.297)	(7.2×10^{-3})	(1.87×10^{-4})	(2.16×10^{-3})	(3.6×10^{-3})
GS[e]	−7.044	0.907	0.072	0.48	0.673
	(1.365)	(7.9×10^{-3})	(1.6×10^{-4})	(2.3×10^{-3})	(3.9×10^{-3})
BC[f]	−7.027	0.909	0.071	0.483	0.675
	(1.557)	(7.3×10^{-3})	(1.79×10^{-4})	(2.03×10^{-3})	(3.34×10^{-3})
Exact[g]	−7.020	0.907	0.073	0.484	0.676
	$[1.31 \times 10^{-2}]$	$[9.65 \times 10^{-4}]$	$[1.4 \times 10^{-4}]$	$[4.95 \times 10^{-4}]$	$[6.56 \times 10^{-4}]$
	(1.349)	(7.52×10^{-3})	(1.82×10^{-4})	(2.19×10^{-3})	(3.77×10^{-3})
	{0.786}	{0.808}	{0.929}	{0.894}	{0.876}
			$n = 1000$		
MLE[c]	−8.309	0.951	0.068	0.446	0.636
	(0.165)	(8.7×10^{-4})	(1.63×10^{-5})	(1.81×10^{-4})	(3.6×10^{-4})
Laplace[d]	−8.322	0.952	0.068	0.446	0.636
	(0.166)	(8.74×10^{-4})	(1.64×10^{-5})	(1.81×10^{-4})	(3.6×10^{-4})
GS[e]	−8.301	0.950	0.067	0.445	0.634
	(0.168)	(9.0×10^{-4})	(1.54×10^{-5})	(1.76×10^{-4})	(3.47×10^{-4})
BC[f]	−8.299	0.951	0.067	0.446	0.636
	(0.184)	(8.12×10^{-4})	(1.47×10^{-5})	(1.57×10^{-4})	(3.04×10^{-4})
Exact[g]	−8.334	0.953	0.067	0.447	0.636
	$[4.08 \times 10^{-3}]$	$[2.97 \times 10^{-4}]$	$[3.95 \times 10^{-5}]$	$[1.33 \times 10^{-4}]$	$[1.87 \times 10^{-4}]$
	(0.168)	(8.9×10^{-4})	(1.59×10^{-5})	(1.79×10^{-4})	(3.54×10^{-4})
	{1.009}	{1.009}	{1.019}	{1.012}	{1.012}

[a]Posterior variance is in parentheses; numerical standard errors in square brackets; relative numerical efficiency in curly brackets.
[b]Obtained at the mean of x.
[c]Using the Newton–Raphson method.
[d]Second-order approximation; derivatives using (13).
[e]Based on $M = 200$ and $t = 50$.
[f]Based on $N = 50$ draws from (7) with $\nu = 15$.
[g]Based on $N = 10{,}000$ draws from (7) with $\nu = 15$.

standard errors and *RNE* associated with the exact estimates suggest that the symmetric multivariate-t importance pdf is quite adequate for this model.

The marginal posteriors in figs. 1–3 for some of the parameters in example 2 reveal some interesting results. When the sample size is 200, the skewness in the posteriors is quite evident, although it is much less for α_2, and tends to disappear as the sample increases. Second, for any given sample size, the

Table 2

Example 1 (continued): MLE, Bayesian posterior mean and variance,[a] *60% censoring*.

	β_1 (-13)	β_2 (1)	τ^2 (0.0625)	$\partial \bar{y}_p/\partial X_2$ (0.249)[b]	$\partial \bar{y}/\partial X_2$ (0.266)[b]
			$n = 100$		
MLE[c]	-11.951	0.976	0.066	0.272	0.323
	(3.826)	(1.66×10^{-2})	(2.36×10^{-4})	(7.49×10^{-4})	(2.12×10^{-3})
Laplace[d]	-12.285	0.996	0.066	0.273	0.323
	(4.118)	(1.78×10^{-2})	(2.37×10^{-4})	(7.64×10^{-4})	(2.14×10^{-3})
GS[e]	-12.741	1.027	0.063	0.279	0.323
	(5.023)	(2.16×10^{-2})	(2.51×10^{-4})	(8.46×10^{-4})	(2.28×10^{-3})
BC[f]	-12.731	1.023	0.060	0.279	0.323
	(5.561)	(2.2×10^{-2})	(2.49×10^{-4})	(6.59×10^{-4})	(2.02×10^{-3})
Exact[g]	-12.512	1.008	0.063	0.275	0.320
	$[2.72 \times 10^{-2}]$	$[1.74 \times 10^{-3}]$	$[1.71 \times 10^{-4}]$	$[3.30 \times 10^{-4}]$	$[5.43 \times 10^{-4}]$
	(4.249)	(1.84×10^{-2})	(2.26×10^{-4})	(8.25×10^{-4})	(2.18×10^{-3})
	{0.574}	{0.608}	{0.773}	{0.758}	{0.739}
			$n = 1000$		
MLE[c]	-12.816	0.996	0.064	0.252	0.274
	(0.482)	(2.02×10^{-3})	(2.35×10^{-5})	(6.83×10^{-5})	(2.01×10^{-4})
Laplace[d]	-12.857	0.999	0.064	0.252	0.274
	(0.486)	(2.03×10^{-3})	(2.35×10^{-5})	(6.84×10^{-5})	(2.01×10^{-4})
GS[e]	-12.857	0.999	0.064	0.253	0.275
	(0.498)	(2.11×10^{-3})	(2.35×10^{-5})	(6.56×10^{-5})	(1.84×10^{-4})
BC[f]	-12.850	0.999	0.063	0.253	0.276
	(0.559)	(2.08×10^{-3})	(2.26×10^{-5})	(5.01×10^{-5})	(2.10×10^{-4})
Exact[g]	-12.877	1.001	0.064	0.252	0.274
	$[7.03 \times 10^{-3}]$	$[4.56 \times 10^{-4}]$	$[4.79 \times 10^{-5}]$	$[8.33 \times 10^{-5}]$	$[1.41 \times 10^{-4}]$
	(0.486)	(2.06×10^{-3})	(2.31×10^{-5})	(6.97×10^{-5})	(2.00×10^{-4})
	{0.983}	{0.991}	{1.006}	{1.004}	{1.006}

[a]Posterior variance is in parentheses; numerical standard errors in square brackets; relative numerical efficiency in curly brackets.
[b]Obtained at the mean of x.
[c]Using the Newton–Raphson method.
[d]Second-order approximation; derivatives using (13).
[e]Based on $M = 200$ and $t = 50$.
[f]Based on $N = 50$ draws from (7) with $\nu = 15$.
[g]Based on $N = 10{,}000$ draws from (7) with $\nu = 15$.

marginal posteriors for 70% censoring display a pronounced location difference and thicker tails relative to the posteriors with 50% censoring.

The important points illustrated by these examples concern the improvement achieved in small samples by the Bayes estimates over the ML estimate. That this occurs with a diffuse prior on the parameters is all the more significant. Next, the second-order and the GS estimates are quite close to

Table 3

Example 2: MLE, Bayesian posterior mean and variance,[a] *50% censoring*.

	β_1 (-1.1854)	β_2 (1.0)	β_3 (10.0)	τ^2 (0.6429)	$\partial\bar{y}_p/\partial X_2$ (0.195)[b]	$\partial\bar{y}/\partial X_2$ (0.151)[b]
			$n = 200$			
MLE[c]	-0.747 (4.93×10^{-2})	1.053 (8.14×10^{-3})	9.284 (1.56×10^{-1})	0.718 (1.05×10^{-2})	0.248 (9.09×10^{-4})	0.249 (3.45×10^{-3})
Laplace[d]	-0.764 (4.91×10^{-2})	1.054 (8.29×10^{-3})	9.308 (1.59×10^{-1})	0.719 (1.05×10^{-2})	0.248 (9.13×10^{-4})	0.250 (3.45×10^{-3})
GS[e]	-0.763 (4.30×10^{-2})	1.051 (9.11×10^{-2})	9.307 (1.69×10^{-1})	0.699 (1.06×10^{-2})	0.248 (8.60×10^{-4})	0.251 (2.96×10^{-3})
BC[f]	-0.771 (6.67×10^{-2})	1.067 (8.10×10^{-3})	9.283 (2.07×10^{-1})	0.691 (1.11×10^{-2})	0.253 (1.27×10^{-4})	0.257 (4.55×10^{-3})
Exact[g]	-0.773 $[2.42 \times 10^{-3}]$ (5.10×10^{-2}) (0.871)	1.056 $[9.93 \times 10^{-4}]$ (8.65×10^{-3}) (0.877)	9.328 $[4.25 \times 10^{-3}]$ (1.58×10^{-1}) (0.875)	0.696 $[1.03 \times 10^{-3}]$ (1.02×10^{-2}) (0.961)	0.249 $[3.16 \times 10^{-4}]$ (9.17×10^{-4}) (0.918)	0.250 $[5.99 \times 10^{-4}]$ (3.33×10^{-3}) (0.928)
			$n = 400$			
MLE[c]	-0.916 (2.64×10^{-2})	0.990 (3.61×10^{-3})	9.592 (7.43×10^{-2})	0.705 (4.72×10^{-3})	0.320 (7.25×10^{-4})	0.418 (2.33×10^{-4})
Laplace[d]	-0.923 (2.65×10^{-2})	0.990 (3.64×10^{-3})	9.603 (7.48×10^{-2})	0.706 (4.72×10^{-3})	0.320 (7.25×10^{-4})	0.418 (2.33×10^{-3})
GS[e]	-0.939 (2.48×10^{-2})	0.994 (3.41×10^{-3})	9.621 (7.19×10^{-2})	0.703 (4.73×10^{-3})	0.319 (7.60×10^{-4})	0.414 (2.38×10^{-3})
BC[f]	-0.921 (3.30×10^{-2})	0.997 (3.33×10^{-3})	9.573 (9.26×10^{-2})	0.693 (4.95×10^{-3})	0.323 (9.74×10^{-4})	0.421 (3.12×10^{-3})
Exact[g]	-0.926 $[1.67 \times 10^{-3}]$ (2.64×10^{-2}) (0.947)	0.991 $[6.18 \times 10^{-4}]$ (3.63×10^{-3}) (0.950)	9.606 $[2.81 \times 10^{-3}]$ (7.53×10^{-2}) (0.954)	0.696 $[6.88 \times 10^{-4}]$ (4.67×10^{-3}) (0.987)	0.319 $[2.70 \times 10^{-4}]$ (7.08×10^{-4}) (0.971)	0.416 $[4.81 \times 10^{-4}]$ (2.24×10^{-3}) (0.968)

[a] Posterior variance is in parentheses; numerical standard errors in square brackets; relative numerical efficiency in curly brackets.
[b] Obtained at the mean of x. For $n = 400$, true = (0.288, 0.351).
[c] Using the Newton–Raphson method.
[d] Second-order approximation; derivatives using (13).
[e] Based on $M = 200$ and $t = 50$.
[f] Based on $N = 50$ draws from (7) with $\nu = 15$.
[g] Based on $N = 10{,}000$ draws from (7) with $\nu = 15$.

Table 4

Example 2 (continued): MLE, Bayesian posterior mean and variance,[a] *70% censoring*.

	β_1 (-1.1854)	β_2 (1.0)	β_3 (1.0)	τ^2 (0.6429)	$\partial\bar{y}_p/\partial X_2$ $(0.202)^b$	$\partial\bar{y}/\partial X_2$ $(0.165)^b$
			$n = 200$			
MLE[c]	-0.879 (2.72×10^{-2})	0.902 (1.25×10^{-2})	0.611 (3.34×10^{-2})	1.023 (4.14×10^{-2})	0.189 (3.39×10^{-4})	0.163 (5.93×10^{-4})
Laplace[d]	-0.905 (2.90×10^{-2})	0.915 (1.31×10^{-2})	0.626 (3.41×10^{-2})	1.025 (4.15×10^{-2})	0.190 (3.47×10^{-4})	0.163 (5.98×10^{-4})
GS[e]	-0.938 (3.06×10^{-2})	0.936 (1.29×10^{-2})	0.617 (3.75×10^{-2})	0.993 (4.4×10^{-2})	0.192 (3.0×10^{-4})	0.161 (4.99×10^{-4})
BC[f]	-0.903 (2.79×10^{-2})	0.916 (9.62×10^{-2})	0.591 (3.72×10^{-2})	0.991 (4.96×10^{-2})	0.193 (3.95×10^{-4})	0.168 (6.93×10^{-4})
Exact[g]	-0.937 $[2.53 \times 10^{-3}]$ (3.04×10^{-2}) {0.475}	0.930 $[1.58 \times 10^{-3}]$ (1.39×10^{-2}) {0.557}	0.638 $[2.44 \times 10^{-3}]$ (3.61×10^{-2}) {0.606}	0.964 $[2.36 \times 10^{-3}]$ (3.89×10^{-2}) {0.698}	0.192 $[2.38 \times 10^{-4}]$ (3.85×10^{-4}) {0.680}	0.163 $[2.91 \times 10^{-4}]$ (6.17×10^{-4}) {0.729}
			$n = 400$			
MLE[c]	-0.968 (1.64×10^{-2})	0.887 (6.62×01^{-3})	0.827 (2.09×10^{-2})	0.845 (1.48×10^{-2})	0.181 (1.81×10^{-4})	0.150 (2.84×10^{-4})
Laplace[d]	-0.983 (1.7×10^{-2})	0.894 (6.78×10^{-3})	0.834 (2.14×10^{-2})	0.846 (1.48×10^{-2})	0.181 (1.83×10^{-4})	0.150 (2.85×10^{-4})
GS[e]	-1.012 (1.74×10^{-2})	0.903 (6.12×10^{-3})	0.851 (2.47×10^{-2})	0.812 (1.37×10^{-2})	0.183 (1.74×10^{-4})	0.149 (3.07×10^{-4})
BC[f]	-1.000 (2.01×10^{-2})	0.899 (5.43×10^{-3})	0.813 (2.04×10^{-2})	0.816 (1.72×10^{-2})	0.183 (2.35×10^{-4})	0.152 (4.04×10^{-4})
Exact[g]	-1.001 $[1.58 \times 10^{-3}]$ (1.76×10^{-2}) {0.705}	0.902 $[9.47 \times 10^{-4}]$ (6.96×10^{-3}) {0.776}	0.840 $[1.67 \times 10^{-3}]$ (2.26×10^{-2}) {0.810}	0.821 $[1.29 \times 10^{-3}]$ (1.43×10^{-2}) {0.859}	0.183 $[1.49 \times 10^{-4}]$ (1.89×10^{-4}) {0.851}	0.150 $[1.81 \times 10^{-4}]$ (2.87×10^{-4}) {0.876}

[a] Posterior variance is in parentheses; numerical standard errors in square brackets; relative numerical efficiency in curly brackets.
[b] Obtained at the mean of x. For $n = 400$, true = (0.197, 0.155).
[c] Using the Newton–Raphson method.
[d] Second-order approximation; derivatives using (13).
[e] Based on $M = 200$ and $t = 50$.
[f] Based on $N = 50$ draws from (7) with $\nu = 15$.
[g] Based on $N = 10{,}000$ draws from (7) with $\nu = 15$.

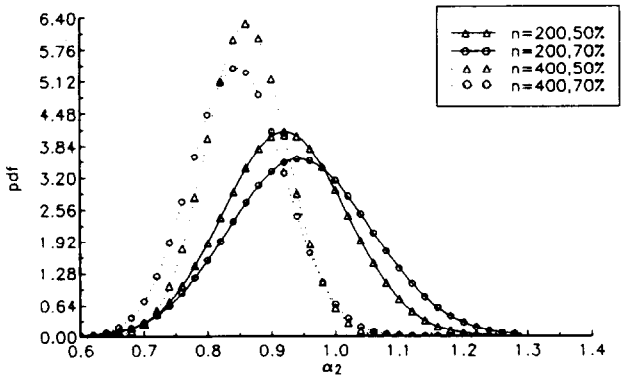

Fig. 1. Pdf of α_2: example 2.

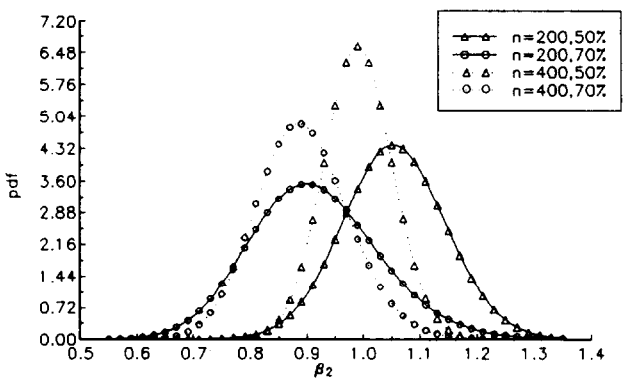

Fig. 2. Pdf of β_2: example 2.

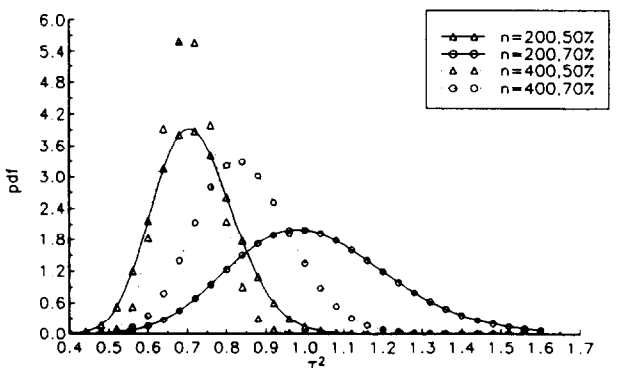

Fig. 3. Pdf of τ^2: example 2.

the exact Bayes estimates in samples as large as 400. One can thus argue that the Tierney–Kass–Kadane and data augmentation procedures applied in this paper may provide very useful alternatives to the exact Monte Carlo method.

We conclude by providing remarks about some applications of our ideas and some open problems. First, we mention that certain other functions of interest, namely, prediction probabilities (for, e.g., the probability that the first out-of-sample observation is censored) and predictive moments, can be easily computed. Second, it is possible to deal with the outlier detection and outlier accommodation issues. This problem is considered in Chib and Tiwari (1989) in the context of Tobit models with nonnormal errors. As far as open problems are concerned, it remains to be seen how the analysis of multiple-equation Tobit models may be conducted. We feel that the Tierney–Kadane approximations, perhaps employed with numerical derivatives and data augmentation ideas, should provide an adequate approach.

Appendix A

The bias term in (8) is derived as follows. From a Taylor's series of $f(\hat{\eta}, \hat{K}) = \hat{\eta}/\hat{K}$ around (η, K), we get

$$\hat{\eta}/\hat{K} = \bar{g} + f_{\eta}(\hat{\eta} - \eta) + f_K(\hat{K} - K) + \tfrac{1}{2}\left(f_{KK}(\hat{K} - K)^2 + 2f_{\eta K}(\hat{\eta} - \eta)(\hat{K} - K) \right) + R_N,$$

where the remainder R_N is $O(N^{-2})$, $f_{KK} = 2\bar{g}K^{-2}$, and $f_{\eta K} = -K^{-2}$. Taking expectations w.r.t. $h(\theta)$ gives

$$\begin{aligned} \mathrm{E}^h(\hat{\eta}/\hat{K}) &= \bar{g} + N^{-1}\bar{g}K^{-2}\int_{\Theta}(w(\theta) - K)^2 h(\theta)\,\mathrm{d}\theta \\ &\quad - N^{-1}K^{-2}\int_{\Theta}(w(\theta) - K)(g(\theta)w(\theta) - \eta)h(\theta)\,\mathrm{d}\theta \\ &\quad + O(N^{-2}) \\ &= \bar{g} + N^{-1}K^{-2}\int_{\Theta}(w(\theta) - K) \\ &\quad \times [\bar{g}(w(\theta) - K) - (g(\theta)w(\theta) - \eta)]h(\theta)\,\mathrm{d}\theta \\ &\quad + O(N^{-2}), \end{aligned}$$

which yields (8) if the expression in square brackets is simplified using the fact that $\eta = \bar{g}K$.

Appendix B

We consider the problem of maximizing $\tilde{L}(\psi)$ subject to $\alpha_j/\tau = \gamma$ [cf. (15)]. Let $\underset{\sim}{\alpha}: (k-1)\times 1$ denote the vector of parameters excluding α. Similarly, let $\underset{\sim}{X} = (\underset{\sim}{x}_1', \ldots, \underset{\sim}{x}_n')' = (\underset{\sim}{X}_0', \underset{\sim}{X}_1')'$ denote the matrix of exogenous variables without the jth column and $x_j = (x_{j0}, x_{j1})$: $n \times 1$. Also let $n_1 = \dim(y_1) - 1$. Then substituting $\alpha_j = \gamma\tau$ into $\tilde{L}(\psi)$ we obtain

$$\underset{\sim}{L}(\underset{\sim}{\alpha}, \tau) = \Sigma_0 \log[1 - \Phi(w_i)] + n_1 \log \tau - \tfrac{1}{2}\Sigma_1(\tau \underset{\sim}{y}_i - \underset{\sim}{x}_i'\underset{\sim}{\alpha})^2,$$

where $w_i = \underset{\sim}{x}_i'\underset{\sim}{\alpha} + x_{ij}\gamma\tau$, $i \in C$, and $\underset{\sim}{y}_i = (y_i - x_{ij}\gamma)$, $i \in U$, where U is the set of indices of the uncensored data. Letting $\underset{\sim}{y}_1: n_1 \times 1$ denote the vector with components $\underset{\sim}{y}_i$, the derivatives are now given by

$$\underset{\sim}{L}_{\underset{\sim}{\alpha}} = -\Sigma_0 \phi(w_i)\underset{\sim}{x}_i/(1 - \Phi(w_i)) + X_1'(\tau \underset{\sim}{y}_1 - \underset{\sim}{X}_1\underset{\sim}{\alpha}),$$

$$\underset{\sim}{L}_{\tau} = -\Sigma_0 \phi(w_i)(x_{ij}\gamma)/(1 - \Phi(w_i)) + n_1\tau^{-1} - \underset{\sim}{y}_1'(\tau \underset{\sim}{y}_1 - \underset{\sim}{X}_1\underset{\sim}{\alpha}),$$

$$\underset{\sim\sim}{L}_{\underset{\sim}{\alpha}\underset{\sim}{\alpha}} = -\underset{\sim}{X}_0'\underset{\sim}{B}_0\underset{\sim}{X}_0 - \underset{\sim}{X}_1'\underset{\sim}{X}_1,$$

$$\underset{\sim}{L}_{\underset{\sim}{\alpha}\tau} = \underset{\sim}{X}_0'\underset{\sim}{B}_0 x_{j0}\gamma + \underset{\sim}{X}_1'\underset{\sim}{y}_1,$$

$$\underset{\sim}{L}_{\tau\tau} = -\Sigma_0(\gamma x_{ij})^2 a_i - n_1\tau^{-2} = -(x_{j0}\gamma)B_0(x_{j0}\gamma) - n_1\tau^{-2},$$

where $\underset{\sim}{B}_0: n_0 \times n_0$ is a diagonal matrix with typical elements $\underset{\sim}{b}_i = \lambda(w_i)(\lambda(w_i) - w_i)$, where $\lambda(w_i) = \phi(w_i)/(1 - \Phi(w_i))$. A Newton–Raphson procedure based on these derivatives yields the maximizing values of $\underset{\sim}{\alpha}$ and τ. The value of α_j is then obtained through the constraint.

References

Albert, James H., 1988, Computational methods using a Bayesian hierarchical generalized linear model, Journal of the American Statistical Association 83, 1037–1044.

Amemiya, Takeshi, 1984, Tobit models: A survey, Journal of Econometrics 24, 3–61.

Amemiya, Takeshi, 1985, Advanced econometrics (Harvard University Press, Cambridge, MA).

Carriquiry, Alicia L., Daniel Gianola, and Rohan L. Fernando, 1987, Mixed-model analysis of a censored normal distribution with reference to animal breeding, Biometrics 43, 929–939.

Chib, Siddhartha and Ram C. Tiwari, 1989, On a Bayesian approach to outlier detection and residual analysis, Manuscript.

Devroye, Luc, 1986, Non-uniform random variate generation (Springer-Verlag, New York, NY).

Gelfand, Alan E. and Adrian F.M. Smith, 1990, Sampling based approaches to calculating marginal densities, Journal of the American Statistical Association 85, 398–409.

Gelfand, Alan E., Susan E. Hills, Amy Racine-Poon, and Adrian F.M. Smith, Illustration of Bayesian inference in normal data models using Gibbs sampling, Manuscript.

Geman, Stuart and Donald J. Geman, 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Geweke, John, 1986, Exact inference in the inequality constrained normal linear regression model, Journal of Applied Econometrics 1, 127–141.
Geweke, John, 1989, Bayesian inference in econometric models using Monte Carlo integration, Econometrica 57, 1317–1339.
Greene, William H., 1990, Econometric analysis (Macmillan, New York, NY).
Judge, George C., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee, 1988, Introduction to the theory and practice of econometrics (Wiley, New York, NY).
Kass, Robert E., Luke Tierney, and Joseph B. Kadane, 1988, Asymptotics in Bayesian computation, in: Jose M. Bernardo, Morris H. DeGroot, Dennis V. Lindley, and Adrian F.M. Smith, eds., Bayesian statistics: 3 (Oxford University Press, Oxford) 261–278.
Kass, Robert E., Luke Tierney, and Joseph B. Kadane, 1990, The validity of posterior expansions based on Laplace's method, in: Seymour Geisser, James S. Hodges, S. James Press, and Arnold Zellner, eds., Bayesian and likelihood methods in statistics and econometrics (North-Holland, Amsterdam) 473–487.
Kloek, Teun and Herman K. Van Dijk, 1978, Bayesian estimates of equation system parameters: An application of integration by Monte Carlo, Econometrica 46, 1–20.
Maddala, G.S., 1983, Limited-dependent and qualitative variables in econometrics (Cambridge University Press, New York, NY).
McDonald, John F. and Robert A. Moffitt, 1980, The uses of Tobit analysis, Review of Economics and Statistics 62, 318–321.
Poirier, Dale J., 1988, Bayesian diagnostic testing in the general linear normal regression model, in: Jose M. Bernardo, Morris H. DeGroot, Dennis V. Lindley, and Adrian F.M. Smith, eds., Bayesian statistics: 3 (Oxford University Press, Oxford) 725–732.
Sweeting, Trevor J., 1987, Approximate Bayesian analysis of censored survival data, Biometrika 74, 809–916.
Tanner, Martin A. and Wing-Hung Wong, 1987, The calculation of posterior distributions by data augmentation (with discussion), Journal of the American Statistical Association 82, 528–550.
Tierney, Luke and Joseph B. Kadane, 1986, Accurate approximations for posterior moments and marginal densities, Journal of the American Statistical Association 81, 82–86.
Tierney, Luke, Robert E. Kass, and Joseph B. Kadane, 1989, Approximate marginal densities of nonlinear functions, Biometrika 76, 425–433.
Tobin, James, 1958, Estimation of relationships for limited dependent variables, Econometrica 26, 24–36.
Wales, Terence J. and Alan D. Woodland, 1980, Sample selectivity and the estimation of labor supply functions, International Economic Review 21, 437–468.
Zellner, Arnold and Peter E. Rossi, 1984, Bayesian analysis of dichotomous quantal response models, Journal of Econometrics 25, 365–393.