A BAYESIAN ANALYSIS OF ENDOGENOUS SWITCHING MODELS FOR COUNT DATA

Hideo Kozumi*

This paper considers the count model with endogenous switching proposed by Terza (1998) from a Bayesian point of view. We consider Markov chain Monte Carlo methods to estimate the parameters of the model. Furthermore, an extension is made to handle the case of non-normality and model determination is discussed. Our approach is illustrated with both simulated and real data sets.

Key words and phrases: Bivariate t distribution; Count data; Endogenous switching; Markov chain Monte Carlo; Poisson model; Unobserved heterogeneity.

1. Introduction

The Poisson model has been widely used in the literature (see, e.g., Hausman et al. (1984); Kumar and Shih (1978); Okoruwa et al. (1988)) in count data regression. Although the Poisson model is useful for count data analysis, count data often exhibit “non-Poisson” features such as overdispersion, excess zeros, etc. The Poisson model is apparently inadequate for the examination of count data with such features, and because of this, several authors have extended it: count models with unobserved heterogeneity (Cameron and Trivedi (1986); Gourieroux and Visser (1997)), truncated count models (Grogger and Carson (1991)), censored count models (Terza (1985)), hurdle count models (Mullahy (1986)) and zero inflated count models (Lambert (1992)). Winkelmann and Zimmermann (1995) and Winkelmann (1997) give excellent surveys of these models.

Since the work by Heckman (1976), the problem of sample selection has been examined by many researchers, both theoretically and empirically. As Heckman’s (1976) analysis is based on a linear regression model, there are extensions to other econometric models such as binary choice models. Recently, several count models dealing with sample selection have been proposed. Crepon and Duguet (1997) propose a count model with endogenous censoring. In their model, individual counts are censored depending on the outcome of a binary variable. They also suggest to use the simulated maximum likelihood method from Gourieroux and Monfort (1991) to estimate parameters. Terza (1998) discusses a count model with endogenous switching, in which a binary variable is incorporated in the mean number of counts. For the estimation of parameters, he proposes a non-linear weighted least squares (NWLS) estimator as well as a two-stage method of moments estimator. A count model with endogenous reporting has also been considered by Winkelmann (1998).

*Graduate School of Economics, Hokkaido University Kita 9 Nishi 7, Kita-ku, Sapporo 060-0809, Japan.

From a Bayesian point of view, Winkelmann (1996) deals with underreported count data, and Chib et al. (1998) examine a panel count model with random effects. This paper considers a count model with selectivity in a Bayesian framework. Among the count models with selectivity, we confine ourself to the model that utilizes endogenous switching considered in Terza (1998). Although Terza (1998) proposed a NWLS estimator, this procedure resulted in anomalous estimates as shown in his numerical analysis. Thus we propose a Bayesian approach for the estimation of the model using Markov chain Monte Carlo (MCMC) methods.

Count models with selectivity consist of a distribution of counts and a selection mechanism. The selection mechanism is usually based on a normality assumption. However, as pointed out by Goldberger (1983), parameter estimates are sensitive to departures from normality. This fact stimulates us to extend the count model to non-normal cases. In this paper, we consider a non-normal case by assuming that the selection mechanism is modeled with a $t$ distribution.

The rest of the paper is organized as follows. In Section 2 we summarize the count model with endogenous switching and obtain a joint posterior distribution. Section 3 discusses computational strategy of the MCMC method, and the extension to a non-normal case is also considered. In Section 4 our approach is illustrated using both simulated and real data sets. Finally, brief conclusions are given in Section 5.

2. Endogenous switching model

Let us consider a collection of count dependent variables $\{y_i\}_{i=1}^n$ and covariates $\{x_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$, where $x_i$ is a $k \times 1$ vector and $z_i$ is an $l \times 1$ vector. The set of covariates may be disjoint or overlapping. The count model with endogenous switching proposed by Terza (1998) is defined as follows. The probability distribution of $y_i$ is given by

$$f(y_i \mid c_i, u_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!},$$

where

$$\lambda_i = \exp(x_i'\alpha + \beta c_i + u_i),$$

(2.1)

and $u_i$ is unobserved heterogeneity. In (2.1) $c_i$ represents a binary variable and is characterized as

$$c_i = \begin{cases} 1 & \text{if } c_i^* = z_i'\gamma + \epsilon_i \geq 0, \\ 0 & \text{otherwise}, \end{cases}$$

where $c_i^*$ is a latent random variable and $\epsilon_i$ is a disturbance. The joint distribution of $u_i$ and $\epsilon_i$ is assumed to be normal with vector zero and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho \sigma \\ \rho \sigma & 1 \end{pmatrix},$$
where $\sigma^2$ is the variance of $u_t$, $\rho$ is the coefficient of correlation and the variance of $e_t$ is normalized to one for identification. From these specifications, we can see that $y_t$ is Poisson distributed with parameter $\lambda_t$, and $c_t$ is generated from the probit model. Note that correlation between $y_t$ and $c_t$ is modeled through the joint distribution of $u_t$ and $\epsilon_t$.

Suppressing dependence on the covariates and the parameters, the likelihood function of the parameters given the data $D = \{y_i, c_i\}_{i=1}^n$ can be expressed as

$$f(D) = \prod_{i=1}^{n} f(y_t | c_t, u_t) f(c_t | u_t) f(u_t) du_t,$$

$$= \prod_{i=1}^{n} f(y_t | c_t, u_t) \{1 - \Phi^*(u_t)\}^{1-c_t} \Phi^*(u_t)^{c_t} f(u_t) du_t,$$

where $f(u_t)$ is a normal density with mean zero and variance $\sigma^2$,

$$\Phi^*(u_t) = \Phi \left( \frac{z_t' \gamma + \rho u_t / \sigma}{\sqrt{1 - \rho^2}} \right),$$

and $\Phi$ denotes the cumulative distribution function of the standard normal distribution. Consequently, maximum likelihood estimation involves computation of the integral. Terza (1998) suggests that the integral can be evaluated by using the Hermite quadrature method though this was not conducted in his numerical analysis. Furthermore, he considered a NWLS estimator as well as a two-stage method of moments, which is analogous to the Heckman (1976, 1979) estimator.

There are several drawbacks in these estimation procedures. Since the Hermite quadrature method relies on the normality of $u_t$, it is difficult to apply the maximum likelihood method to non-normal cases. As shown in Terza (1998), the two-stage method of moments estimator is statistically inefficient relative to the NWLS estimator. However, the NWLS method results in estimates of $\rho$ which are not within the required interval $(-1, 1)$. Thus, we pursue a Bayesian approach using MCMC methods, which will be discussed in the next section.

3. Posterior analysis

3.1. Joint posterior distribution

The likelihood function given in (2.2) is not particularly useful for Bayesian inference because its evaluation is computationally cumbersome. In this paper, we consider an alternative approach based on the framework by Albert and Chib (1993). Their idea is to augment the model with a vector of latent variables $C^* = (c_1^*, \ldots, c_n^*)$ and consider the joint posterior distribution of the parameters and the vector $C^*$. Furthermore, an advantage of the Bayesian approach is that a vector of unobserved heterogeneity $U = (u_1, \ldots, u_n)$ can be viewed as a parameter. To implement MCMC methods, we need only to sample the parameters $\{\theta = (\alpha', \beta'), \gamma, \sigma^2, \rho\}, U$ and $C^*$, recursively. Consequently, we do not have to integrate out $u_t$ to obtain the likelihood function.
Given a prior density $\pi(\theta, \gamma, \sigma^2, \rho)$, the joint posterior distribution can be expressed as

$$
\pi(\theta, \gamma, \sigma^2, \rho, U, C^* \mid D) \propto \pi(\theta, \gamma, \sigma^2, \rho) \prod_{i=1}^{n} f(y_i \mid c_i, u_i, \theta) \pi(u_i, c_i^* \mid \gamma, \sigma^2, \rho) I(c_i^* \in J_i),
$$

where $\pi(u_i, c_i^* \mid \gamma, \sigma^2, \rho)$ is a bivariate normal density obtained from the joint distribution of $u_i$ and $c_i$, $I(A)$ denotes the indicator function of the event $A$, and $J_i$ is the interval $[0, \infty)$ if $c_i = 1$ and the interval $(-\infty, 0)$ if $c_i = 0$. The effects of $c_i$ can be seen through $f(y_i \mid c_i, u_i, \theta)$ and $J_i$.

Since we adopt a Bayesian approach, we complete the model by specifying the following prior distributions over the parameters $(\theta, \gamma, \sigma^2, \rho)$:

$$
\pi(\theta) = N(\theta_0, H_0), \quad \pi(\gamma) = N(\gamma_0, G_0),
\pi(\sigma^2) = IG(n_0/2, s_0/2), \quad \pi(\rho) \propto N(\rho_0, v_0) I(|\rho| < 1),
$$

where $IG(a, b)$ denotes an inverse gamma distribution with parameters $a$ and $b$. It should be mentioned that our prior distribution for $\rho$ is the same as the one used in Chib and Greenberg (1998) and ensures that estimates of $\rho$ are within the interval $(-1, 1)$.

### 3.2. Posterior simulation

Since the joint posterior distribution given by (3.1) is much simplified, we can now use MCMC methods. The Markov chain sampling scheme can be constructed from the complete conditional distributions of $\theta$, $\gamma$, $\sigma^2$, $\rho$, $u_i$ ($i \leq n$), and $c_i^*$ ($i \leq n$).

#### 3.2.1. Sampling $\theta$

From (3.1), the complete conditional distribution of $\theta$ is written as

$$
\pi(\theta \mid D, \gamma, \rho, U, C^*) \propto \pi(\theta) \prod_{i=1}^{n} f(y_i \mid c_i, u_i, \theta),
$$

which cannot be sampled by standard methods. Therefore, we adopt the Metropolis-Hastings (MH) algorithm (see, e.g., Tierney (1994) andGamerman (1997)).

In general, given a target density $g(\phi)$, the MH algorithm consists of the following two steps:

(i) Sample a proposal value $\phi^*$ given the current value $\phi$ from the proposal density $q(\phi^* \mid \phi)$.

(ii) Accept $\phi^*$ with probability

$$
\alpha(\phi, \phi^*) = \frac{g(\phi^*) q(\phi \mid \phi^*)}{g(\phi) q(\phi^* \mid \phi)}
$$

and return $\phi^*$. Otherwise return $\phi$. 

ENDOGENOUS SWITCHING COUNT MODEL

For the MH algorithm to work efficiently, the choice of the proposal density $q$ is critical. In order to obtain the appropriate proposal density, we consider the following adjusted linear model (see McCullagh and Nelder (1989)):

\[
\tilde{y}_i = \tilde{x}_i'\theta^* + u_i + \eta_i, \quad \eta_i \sim N(0, v_i),
\]

where

\[
\tilde{x}_i = (x'_i, c_i)', \quad \tilde{y}_i = \tilde{x}_i'\theta + (y_i - \lambda_i)/\lambda_i, \quad v_i = 1/\lambda_i.
\]

It should be noted that $\tilde{y}_i$, $v_i$, and $\lambda_i$ are evaluated at the current value of $\theta$. Then, combining the adjusted linear model (3.3) with the prior distribution $\pi(\theta)$, we have

\[
\theta^* \sim N(\tilde{\theta}, \tilde{H}),
\]

where

\[
\tilde{H}^{-1} = \sum_{i=1}^n \frac{\tilde{x}_i\tilde{x}'_i}{v_i} + H_0^{-1}, \quad \tilde{\theta} = \tilde{H} \left\{ \sum_{i=1}^n \frac{\tilde{x}_i(\tilde{y}_i - u_i)}{v_i} + H_0^{-1}\theta_0 \right\}.
\]

Although this distribution can be a good approximation to the complete conditional distribution of $\theta$, the weight function, which is the ratio of the target and proposal densities, is unbounded. Thus, as in Chib et al. (1998), we adopt the proposal density given by

\[
q(\theta^* | \theta) = \text{MVt}(\tilde{\theta}, \tilde{H}, \nu_\theta),
\]

where MVt($\mu, V, \nu$) denotes a multivariate t distribution with mean $\mu$, scale matrix $V$ and $\nu$ degrees of freedom. As one of the referees pointed out, when the dimension of $\theta$ is large, the acceptance rate of $\theta$ might become low. In that case, it would be appropriate to repeat the Newton-Raphson method to obtain the approximate mode of (3.2) and use it in (3.3). It is, however, computationally burdensome to run the Newton-Raphson method within each MCMC iteration.

### 3.2.2. Sampling the other parameters

For $\gamma$, it can be easily obtained that

\[
\pi(\gamma | D, \theta, \sigma^2, \rho, U, C^*) = N(\hat{\gamma}, \hat{G}),
\]

where

\[
\hat{G}^{-1} = \frac{1}{1 - \rho^2} \sum_{i=1}^n z_i z'_i + G_0^{-1}, \quad \hat{\gamma} = \hat{G} \left( \frac{1}{1 - \rho^2} \sum_{i=1}^n z_i \left( c_i^* - \frac{\rho}{\sigma} u_i \right) + G_0^{-1}\gamma_0 \right).
\]

The complete conditional distribution of $\sigma^2$ is given by

\[
\pi(\sigma^2 | D, \theta, \gamma, \rho, U, C^*) \propto \pi(\sigma^2) \prod_{i=1}^n \pi(u_i, c_i^* | \gamma, \sigma^2, \rho),
\]

\[
\propto \pi(\sigma^2) \prod_{i=1}^n \pi(c_i^* | u_i, \gamma, \sigma^2, \rho) \pi(u_i | \sigma^2),
\]
where

\[ \pi(c_i^* | u_i, \gamma, \sigma^2, \rho) = N \left( z_i^* \gamma + (\rho/\sigma)u_i, 1 - \rho^2 \right), \]

and \( \pi(u_i | \sigma^2) \) is a normal density with mean zero and variance \( \sigma^2 \). Since

\[ \pi(\sigma^2) \prod_{i=1}^{n} \pi(u_i | \sigma^2) = IG(\hat{n}/2, \hat{s}/2), \]

where \( \hat{n} = n + n_0 \) and \( \hat{s} = \sum_{i=1}^{n} u_i^2 + s_0 \), the independent proposal density \( q(\sigma^{2*} | \sigma^2) = IG(\hat{n}/2, \hat{s}/2) \) can be applied. Note that the proposal density \( q(\sigma^{2*} | \sigma^2) \) becomes the same as the complete conditional distribution when \( \rho = 0 \).

From (3.1) we have

\[ \pi(\rho | D, \theta, \gamma, U, C^*) \propto \pi(\rho) \prod_{i=1}^{n} \pi(u_i, c_i^* | \gamma, \sigma^2, \rho), \]

\[ \propto \pi(\rho) \prod_{i=1}^{n} \pi(u_i | c_i^*, \gamma, \sigma^2, \rho), \]

where

\[ \pi(u_i | c_i^*, \gamma, \sigma^2, \rho) = N \left( \rho \sigma (c_i^* - z_i^* \gamma), \sigma^2 (1 - \rho^2) \right). \]

Again this distribution is non-standard. Since \( \rho \) is scalar, we use the random walk proposal density and let the increment random be an univariate normal with variance \( c_\rho/n \) where \( 1/n \) is the large-sample variance of the marginal posterior of a correlation coefficient and \( c_\rho \) is a scaling factor. It should be mentioned that the proposal density of \( \rho \) is not truncated to the interval \((-1, 1)\) since the constraint is part of the target density. Thus, if the proposed value of \( \rho \) is not within the interval, the conditional posterior is zero, and the proposal value is rejected with probability one (see Chib and Greenberg (1998)).

The complete conditional distribution of \( u_i \) is written as

\[ \pi(u_i | D, \theta, \gamma, \sigma^2, \rho, \{u_j\}_{j \neq i}, C^*) \propto f(y_i | c_i, u_i, \theta) \pi(u_i | c_i^*, \gamma, \sigma^2, \rho). \]

In a similar way to the sampling of \( \theta \), we suggest to use the proposal density

\[ q(u_i^* | u_i) = t(\tilde{u}_i, \hat{v}_i, \nu_i), \]

where

\[ \hat{v}_i = \left( 1 + \frac{1}{\nu_i} \right)^{-1}, \quad \nu_i = \hat{v}_i \left( \hat{y}_i - \bar{y}_i^2 \theta \right) + \frac{\rho \sigma (c_i^* - z_i^* \gamma)}{\sigma^2 (1 - \rho^2)}, \]

and \( \hat{y}_i \) and \( \nu_i \) are evaluated at the current value of \( u_i \). As an alternative approach, it is possible to use the adaptive rejection method proposed by Gilks and
Wild (1992) because the target density is log-concave. In count data analysis, the sample size \( n \) can be large and the adaptive rejection method may be time consuming. Thus we do not recommend to use the algorithm.

Finally, the complete conditional distribution of \( c_i^* \) is given by

\[
\pi(c_i^* \mid D, \theta, \gamma, \sigma^2, \rho, U, \{c_j^*\}_{j \neq i}) \propto \pi(c_i^*, u_i \mid \gamma, \sigma^2, \rho) I(c_i^* \in J_i),
\]

\[
\propto \pi(c_i^* \mid u_i, \gamma, \sigma^2, \rho) I(c_i^* \in J_i).
\]

Thus, the distribution of \( c_i^* \) is a normal distribution truncated to the region \( J_i \).

We have three tuning constants \( (\nu_\theta, \nu_u, c_\rho) \). In the numerical examples given below, we set \( \nu_\theta = \nu_u = 10 \), and \( c_\rho \) was determined by short preliminary runs such that the acceptance rate was around 0.5.

### 3.3. Extension to t distribution

In the previous subsection, \( u_i \) and \( \epsilon_i \) are assumed to follow a bivariate normal distribution. As pointed out by Goldberger (1983), parameter estimates are sensitive to departures from normality. Apparently there are two sources of non-normality: one is from \( u_i \) and the other is from \( \epsilon_i \). Since it seems difficult to deal with these non-normalities separately, we assume that the joint distribution of \( u_i \) and \( \epsilon_i \) is non-normal. A possible extension is to apply suitable mixtures of normal distributions.

Let us introduce the additional random variables \( \tau_i \) which follow a gamma distribution \( \text{Ga}(\nu/2, \nu/2) \). Suppose that, given \( \tau_i \), \( u_i \) and \( \epsilon_i \) are jointly normal with vector zero and covariance matrix \((1/\tau_i) \Sigma\). This specification is equivalent to assuming a bivariate \( t \) distribution with \( \nu \) degrees of freedom. Albert and Chib (1993) consider the same approach in the probit model. To implement the posterior analysis for the count model with a bivariate \( t \) distribution, it is necessary to modify the MCMC algorithm discussed above. Since, conditional on \( \tau = \{\tau_i\}_{i=1}^n \), the complete conditional distributions of \( \theta, \gamma, \sigma^2, \rho, u_i \) and \( c_i^* \) collapse to the normal case with minor modification to the variances, we show only those for \( \tau_i \) and \( \nu \).

It is easily verified that the complete conditional distribution of \( \tau_i \) is given as

\[
\pi(\tau_i \mid Y, C, \theta, \gamma, \sigma^2, \rho, U, C^*, \{\tau_j\}_{j \neq i}, \nu) = \text{Ga}((\nu + 2)/2, (\nu + \omega_i)/2),
\]

where \( \omega_i = (u_i, c_i^* - z_{i\gamma}) \Sigma^{-1}(u_i, c_i^* - z_{i\gamma})' \). The complete conditional distribution of \( \nu \) is expressed as

\[
\pi(\nu \mid Y, C, \theta, \gamma, \sigma^2, \rho, U, C^*, \tau, \nu) \propto \pi(\nu) \prod_{i=1}^n \frac{(\nu/2)^{\nu/2} \Gamma(\nu/2)}{\Gamma(\nu/2)} \tau_i^{\nu/2-1} \exp(-\nu \tau_i/2),
\]

where \( \pi(\nu) \) and \( \Gamma(x) \) denote the prior distribution of \( \nu \) and the gamma function, respectively. Since we are usually interested in the posterior probabilities for \( \nu \) in a finite set, \( \nu \) is assumed to be a discrete variable. As a result, the sampling of \( \nu \) reduces to sampling from the discrete distribution, which can be easily implemented. For the sampling of \( \nu \) in a continuous case, see Watanabe (2001) who uses the MH acceptance-rejection algorithm.
4. Numerical examples

4.1. Simulated data

To illustrate the Bayesian approach discussed in the previous section, \( y_i \) and \( c_i^* \) \((i = 1, \ldots, 300)\) were generated from the Poisson distribution with parameter

\[
\lambda_i = \exp(1.0x_i + 1.0c_i + u_i), \quad u_i \sim N(0, 0.3)
\]

and from the normal distribution \( N(1.0 + 2.0z_i, 1.0) \), respectively, where \( x_i \) and \( z_i \) were standard normal variates. To see the effects of the values of \( \rho \), we considered the three cases of \( \rho \) values 0.3, 0.6, 0.9. The features of the simulated data are summarized in Table 1 and the scatter plots of \( u_i \) and \( c_i^* \) are shown in Figure 1. Although we also considered the cases of negative values of \( \rho \), the results obtained were similar to those given below. Therefore, we report only the cases of positive values of \( \rho \) to save space. For the prior distributions, we set the hyper-parameters as follows

\[
\theta_0 = 0, \quad H_0 = 100 \times I, \quad \gamma_0 = 0, \quad G_0 = 100 \times I, \quad n_0 = 3, \quad s_0 = 0.01, \quad \rho_0 = 0, \quad v_0 = 100.
\]

Since it is interesting to see the effect of ignoring correlation between \( y_i \) and \( c_i \), we estimated the model with the restriction \( \rho = 0 \) as well as the model without

<table>
<thead>
<tr>
<th>( \rho = 0.3 )</th>
<th>( \rho = 0.6 )</th>
<th>( \rho = 0.9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of ( y = 0 )</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>mean</td>
<td>6.55</td>
<td>6.56</td>
</tr>
<tr>
<td>min</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>max</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>% of ( c = 1 )</td>
<td>0.68</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Figure 1. Simulated data: The scatter plots of \( u_i \) and \( c_i^* \).
Table 2. Simulated data: Posterior means and standard deviations (in parentheses) are shown.

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.6$</th>
<th>$\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$0.857 (0.087)$</td>
<td>$0.784 (0.086)$</td>
<td>$0.563 (0.091)$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$1.116 (0.099)$</td>
<td>$1.225 (0.097)$</td>
<td>$1.596 (0.112)$</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>$1.117 (0.145)$</td>
<td>$1.031 (0.135)$</td>
<td>$0.841 (0.121)$</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>$2.069 (0.224)$</td>
<td>$1.960 (0.205)$</td>
<td>$1.723 (0.178)$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>$0.302 (0.042)$</td>
<td>$0.275 (0.040)$</td>
<td>$0.293 (0.048)$</td>
</tr>
</tbody>
</table>

Count model without the restriction

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0.3$</th>
<th>$\rho = 0.6$</th>
<th>$\rho = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$0.918 (0.099)$</td>
<td>$0.933 (0.093)$</td>
<td>$0.808 (0.098)$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$1.025 (0.118)$</td>
<td>$1.010 (0.112)$</td>
<td>$1.112 (0.117)$</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>$1.134 (0.150)$</td>
<td>$1.002 (0.130)$</td>
<td>$0.833 (0.115)$</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>$2.056 (0.243)$</td>
<td>$1.942 (0.205)$</td>
<td>$1.765 (0.180)$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>$0.304 (0.044)$</td>
<td>$0.276 (0.041)$</td>
<td>$0.354 (0.050)$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$0.228 (0.164)$</td>
<td>$0.537 (0.135)$</td>
<td>$0.811 (0.095)$</td>
</tr>
</tbody>
</table>

Figure 2. Simulated data: The approximate posterior distributions of $\rho$.

the restriction. With the simulated data, we ran the MCMC algorithm, using 12,000 iterations and discarding the first 2,000 iterations. The chain was considered to have practically converged after 2,000 iterations based on a diagnostic proposed by Geweke (1992). All results reported here were generated using Ox version 2.00 (see Doornik (1998)).

Table 2 shows the posterior estimates of the parameters. From Table 2, it can be seen that, for $\rho = 0.3$, the posterior estimates are similar in both models. However, as the value of $\rho$ increases, distinct differences appear in posterior means, especially in the estimates of $\alpha$ and $\beta$. For the restricted model, we can see serious selection biases, that is, the posterior mean of $\alpha$ decreases and that of $\beta$ increases as the value of $\rho$ increases. The unrestricted model exhibits better estimates of the parameters compared with the restricted model. Thus it
is important to take the correlation between $y_i$ and $c_i$ into account.

The approximate posterior distributions of $\rho$ are shown in Figure 2. From the figure, it can be seen that the posterior distributions exhibit skewness for large values of $\rho$. Furthermore, we can observe that the posterior modes are around the true values.

Table 3 shows the acceptance rates in the parameters $(\theta, \sigma^2, \rho, u_i)$ for the unrestricted model in order to check our sampling scheme. Since there are a large number of $u_i$, the minimum and maximum acceptance rates in $u_i$ are reported. All the acceptance rates in $\theta$ and $u_i$ are around 0.9, indicating that our proposal densities based on the adjusted linear model approximate the corresponding target densities adequately. However, as the value of $\rho$ increases, the acceptance rate in $\sigma^2$ decreases even though all the values are more than 0.5.

### 4.2. Real data

As another example of our approach, let us consider the data set examined by Terza (1998). He considers a model whose dependent count variable $y_i$ is the number of trips taken by members of the $i$-th household in the 24 hr period immediately prior to the survey interview. For the binary variable $c_i$, vehicle ownership is taken, which indicates whether ($c_i = 1$) or not ($c_i = 0$) the household owns at least one motorized vehicle. The data set is comprised of 577 households of whom 490 own at least one vehicle. Following Terza (1998), we used the same covariates $x_i$ and $z_i$ which are summarized in Table 4. A detailed discussion of the data set can be found in Terza and Wilson (1990).

Using the same hyper-parameters as in the previous subsection, we ran the MCMC algorithm using 20,000 iterations and discarding the first 5,000 iterations. As in the previous example, the convergence of the MCMC simulation was checked by the Geweke’s diagnostic.

We estimated the models with and without the restriction $\rho = 0$, and the posterior estimates are found in Table 5. From the table, we can observe that the results are similar in both models. Although the posterior means for $c$ are positive, the estimate for $c$ in the unrestricted model is larger than that in the restricted model. With the restricted model the effect of vehicle ownership on trip frequency is underestimated. Furthermore, we obtained the negative estimate of $\rho$, indicating that one’s affinity for public transit is primarily motivated by disdain for the adverse aspects of private modes of transportation. Compared with the results of Terza (1998), our results are very similar to those of Terza (1998), in
Table 4. Description of independent variables: The circle denotes which variables are included in \( x \) or \( z \).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>( x )</th>
<th>( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>WORKSCHL</td>
<td>% of total trips for work or school vs. personal business or pleasure</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>HHMEM</td>
<td>number of individuals in the household</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>DISTOCBD</td>
<td>distance to the central business district in kilometers</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>AREASIZE</td>
<td>1 if SMSA ( \geq ) 2.5 million population</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>FULLTIME</td>
<td>number of fulltime workers in household</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>DISTONOD</td>
<td>distance from home to nearest transit node, in blocks</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>REALINC</td>
<td>household income divided by median income of census tract in which household resides</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>WEEKEND</td>
<td>1 if 24 hr. survey period is either Saturday or Sunday</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>( c )</td>
<td>1 if household owns at least one motorized vehicle</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>ADULTS</td>
<td>number of adults in the household 16 years of age or older</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

Figure 3. Trip data: The approximate posterior distributions of \( \rho \).

In particular, to those by the NWLS estimation.

To see the effect of non-normality on parameter estimates, we also estimated the count model, in which unobserved heterogeneity \( u_t \) and disturbances \( \epsilon_t \) follow a bivariate \( t \) distribution. In the posterior inference, a prior was used which assigned equal probabilities to \( \nu \) in the set \( \{5, 10, 15, 20, 30, 40, 50\} \). The posterior results for the parameters are also shown in Table 5. As seen from the table, we obtained the results similar to those in the normal case. In the probit equation, all the posterior means in the \( t \) distribution case, except constant term, became larger than those in the normal case. Furthermore, the estimate of \( \rho \) in the case of the \( t \) distribution became smaller than that of the normal case.

Figure 3 shows the approximate posterior distributions of \( \rho \), which are similar in both cases. We can also observe that the posterior modes are around \(-0.7\). Our
Table 5. Trip data: Posterior means and standard deviations (in parentheses) are shown.

<table>
<thead>
<tr>
<th></th>
<th>normal distribution $\rho = 0$</th>
<th>normal distribution no restriction</th>
<th>$t$ distribution no restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Poisson equation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONST</td>
<td>-0.917 (0.169)</td>
<td>-1.348 (0.199)</td>
<td>-1.389 (0.211)</td>
</tr>
<tr>
<td>WORKSCHL</td>
<td>-0.294 (0.134)</td>
<td>-0.338 (0.135)</td>
<td>-0.353 (0.136)</td>
</tr>
<tr>
<td>HHMEM</td>
<td>0.178 (0.025)</td>
<td>0.163 (0.026)</td>
<td>0.163 (0.026)</td>
</tr>
<tr>
<td>DISTOCCBD</td>
<td>-0.001 (0.002)</td>
<td>-0.002 (0.003)</td>
<td>-0.002 (0.003)</td>
</tr>
<tr>
<td>AREASIZE</td>
<td>0.018 (0.086)</td>
<td>0.045 (0.086)</td>
<td>0.035 (0.084)</td>
</tr>
<tr>
<td>FULLTIME</td>
<td>0.330 (0.055)</td>
<td>0.247 (0.060)</td>
<td>0.242 (0.059)</td>
</tr>
<tr>
<td>DISTONOD</td>
<td>0.006 (0.003)</td>
<td>0.005 (0.003)</td>
<td>0.005 (0.002)</td>
</tr>
<tr>
<td>REALINC</td>
<td>0.015 (0.013)</td>
<td>0.007 (0.014)</td>
<td>0.005 (0.014)</td>
</tr>
<tr>
<td>WEEKEND</td>
<td>-0.064 (0.090)</td>
<td>-0.101 (0.091)</td>
<td>-0.102 (0.093)</td>
</tr>
<tr>
<td>$c$</td>
<td>1.316 (0.159)</td>
<td>2.028 (0.227)</td>
<td>2.091 (0.261)</td>
</tr>
<tr>
<td></td>
<td>Probit equation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONST</td>
<td>-0.639 (0.238)</td>
<td>-0.583 (0.223)</td>
<td>-0.789 (0.289)</td>
</tr>
<tr>
<td>WORKSCHL</td>
<td>0.142 (0.267)</td>
<td>0.309 (0.277)</td>
<td>0.390 (0.314)</td>
</tr>
<tr>
<td>HHMEM</td>
<td>0.005 (0.070)</td>
<td>0.042 (0.065)</td>
<td>0.051 (0.072)</td>
</tr>
<tr>
<td>DISTOCCBD</td>
<td>0.024 (0.013)</td>
<td>0.026 (0.013)</td>
<td>0.027 (0.014)</td>
</tr>
<tr>
<td>AREASIZE</td>
<td>-0.205 (0.166)</td>
<td>-0.240 (0.165)</td>
<td>-0.232 (0.173)</td>
</tr>
<tr>
<td>FULLTIME</td>
<td>0.902 (0.158)</td>
<td>1.020 (0.154)</td>
<td>1.094 (0.178)</td>
</tr>
<tr>
<td>ADULTS</td>
<td>0.379 (0.148)</td>
<td>0.273 (0.137)</td>
<td>0.371 (0.158)</td>
</tr>
<tr>
<td>DISTONOD</td>
<td>0.010 (0.007)</td>
<td>0.010 (0.006)</td>
<td>0.011 (0.007)</td>
</tr>
<tr>
<td>REALINC</td>
<td>0.157 (0.057)</td>
<td>0.132 (0.054)</td>
<td>0.149 (0.067)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.478 (0.052)</td>
<td>0.517 (0.060)</td>
<td>0.463 (0.060)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.633 (0.145)</td>
<td>-0.652 (0.163)</td>
<td></td>
</tr>
<tr>
<td>$\nu$</td>
<td>19.05 (10.72)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Trip data: Posterior probabilities of $\nu$.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\nu</td>
<td>D)$</td>
<td>0.000</td>
<td>0.302</td>
<td>0.307</td>
<td>0.189</td>
<td>0.092</td>
<td>0.060</td>
</tr>
</tbody>
</table>

estimates of $\rho$ were in the desired interval $(-1, 1)$ while the NWLS estimate was $-1.119$. Since the NWLS method uses the change of variable, estimates of $\rho$ are calculated from those of $\sigma^2$ and $\rho\sigma$. Thus the anomalous estimate of $\rho$ obtained in Terza (1998) may be due to the small sample biases in the estimates of $\sigma^2$ and $\rho\sigma$. He concluded from these inadequate results that the Poisson specification might be invalid, but our results do not seem to support this conclusion. Finally, Table 6 shows the posterior probabilities of $\nu$. From the table, the posterior mode is 15, and we can observe departure from normality in the trip data. This finding indicates that the trip data examined are overdispersed, which is not well explained by the normal distribution.
5. Conclusions

This paper has examined the count model with endogenous switching proposed by Terza (1998) from a Bayesian point of view. We expressed the joint posterior distribution which does not involve the integral, and proposed MCMC methods to estimate the parameters of the model. We have also made an extension of the model, in which the selection mechanism is modeled using a $t$ distribution. We have illustrated our approach using simulated and real data.

From the results for the simulated data, we found serious selection biases and the importance of correlation between the count dependent variables and the binary variables. As for the real data example, we used the trip data examined by Terza (1998). We obtained the results similar to those in Terza (1998) except the estimate of $\rho$. While the NWLS estimation resulted in an anomalous estimate of $\rho$, our estimate of $\rho$ was in the desired interval. We also found departure from normality in the trip data.

In this paper, we considered only the count model with endogenous switching. There are also other variants of a count model with selectivity and it is possible to apply our approach to these models with some modifications.

Acknowledgements

This research was partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research No.12730019. The author wishes to thank the referees for helpful comments. He is also grateful to Professor Joseph V. Terza for providing the data set used in this paper.

References


