A Bayesian Model of Sample Selection with a Discrete Outcome Variable

Maksym, Obrizan
Kyiv School of Economics, Kyiv Economics Institute

2010

A Bayesian Model of Sample Selection with a Discrete Outcome Variable

Maksym Obrizan[a,b], PhD

December 26, 2010

Affiliations:	[a]Assistant Professor, Kyiv School of Economics (KSE)
	[b]Senior Economist, Kyiv Economics Institute (KEI)
Address:	13 Yakira St., Suite 318, Kyiv, 04119, Ukraine
Phone:	(+380) 44-492-8012
Fax:	(+380) 44-492-8011
E-mail:	mobrizan@kse.org.ua

Abstract

Relatively few published studies apply Heckman's (1979) sample selection model to the case of a discrete endogenous variable and those are limited to a single outcome equation. However, there are potentially many applications for this model in health, labor and financial economics. To fill in this theoretical gap, I extend the Bayesian multivariate probit setup of Chib and Greenberg (1998) into a model of non-ignorable selection that can handle multiple selection and discrete-continuous outcome equations.

The first extension of the multivariate probit model in Chib and Greenberg (1998) allows some of the outcomes to be missing. In addition, I use Cholesky factorization of the variance matrix to avoid the Metropolis-Hastings algorithm in the Gibbs sampler. Finally, using artificial data I show that the model is capable of retrieving the parameters used in the data-generating process and also that the resulting Markov Chain passes all standard convergence tests.

Keywords: *Bayesian computing, latent variable models, Markov Chain Monte Carlo, Bayesian modeling, sample selection model, multivariate probit*

1 Introduction

The problem of *sample selection* applies whenever a dependent variable is missing as a result of a non-experimental selection process. Economists have been aware for a long time that estimating such a model by ordinary least squares leads to inconsistent estimates.[1] Gronau (1974) seems to be among the first to recognize this problem, but Heckman (1979) offers a truly pioneering work with a simple two-step estimator that has been widely used for more than three decades.

Empirical applications of the sample selection model, however, have been mostly limited to the case of a continuous endogenous variable in an outcome equation. In addition, the majority of papers deal with a single selection and a single outcome equation in the sample selection model. In practice, sample may be chosen based on more than one criterion, or more than one outcome equations may be considered.

In this paper I offer a Bayesian model of sample selection with two additional features. First of all, it allows binary dependent variable in the outcome equation as well.[2] Secondly, adding extra selection or outcome equations with dichotomous or continuous dependent variables is straightforward. This extension is crucial in the presence of multiple selection equations, as explained below. These two extensions seem to be an important contribution to the existing literature with potential applications in health, labor and related empirical economic research.

While technical issues limited the use of sample selection models with multiple binary dependent variables, their applicability is potentially very wide. Consider an example from health economics where a researcher is interested in joint estimation of two or more binary morbid health events (for example, hip fracture and stroke) in a sample of Medicare-eligible older Americans. Suppose further that she observes those outcomes only for respondents who allowed her to access their Medicare claims. Clearly, joint estimation of the two health events (outcome variables of interest) with a third equation for being in the analytic sample (selection equation) tends to be more efficient than estimating them equation-by-equation.[3] More importantly, in order to obtain consistent estimates *all* of the selection equations have

[1] There is no such problem if the disturbances in two equations have zero correlation.

[2] I review some earlier work on a discrete outcome variable in Heckman's (1979) model and explain how my model differs further on.

[3] This is a standard result in seemingly unrelated regression model, which does not apply if the explanatory variables are the same or if the correlation/covariance terms are zero.

to be included.

Consider another example from financial economics. Suppose that a credit card company studies the probability of default (outcome equation) for respondents who received a credit card offer. The first selection equation may be if they accepted the offer and applied for a card, and the second whether their application was approved by the bank. In this model, the agent can default only if she was approved for a credit card, which in turn is possible only if she has responded to such an offer.

In labor economics it might be of interest to study employment discrimination (observed for candidates that seek a job) and wage discrimination (observed for candidates that seek a job and are hired). These two outcome equations can be estimated together with selection equation (if a candidate is seeking a job or not). All these and related models can be estimated in the framework developed in this paper.

How is the problem of sample selection accounted for in the multivariate probit model? To continue with the health economics example, suppose that there exists some unobserved factor that affects both the probability of being selected into a sample and of having a morbid health event. If healthier individuals are more likely to allow access to their Medicare claims, say, because of a better cognitive function, then estimating the probability of a morbid health event only for the observed subsample is not representative of the entire population, as only its healthier part is considered.

From the discussion above, it is apparent that in order to consistently estimate a model with sample selection, it is necessary to account for an omitted variable problem. In general, the sample selection problem arises if the unobserved factors determining the inclusion in the subsample are correlated with the unobservables that affect the endogenous variable of primary interest (Vella 1998). In the current paper the specification error of omitted variable resulting from selection is dealt with by considering the unobserved omitted variable as a part of the disturbance term and then jointly estimating the system of equations accounting for the correlations in the variance-covariance matrix.

The multivariate probit model can be used to handle multiple correlated dichotomous variables along the lines of Ashford and Sowden (1970) and Amemiya (1974). It seems, however, that the potential of this model has not been fully realized despite its connection to the normal distribution, which allows for a flexible correlation structure. As noticed in Chib and Greenberg (1998), at least part of the problem in earlier applications arose from the difficulties associated with evaluating the likelihood function by classical methods, except under simplifying assumptions like equicorrelated responses, as in Ochi and Prentice (1984).

Chib and Greenberg (1998) describe how the model can be reformulated in a Bayesian context using the technique of *data augmentation* (discussed in Albert and Chib [1993], among others). The discrete dependent variable in the probit model can be viewed as the outcome of an underlying linear regression with some *latent* dependent variable (i.e. unobserved by the researcher). Consider a decision to make a large purchase, as in Greene (2003, p. 669). If the benefits outweigh the costs (benefits-costs>0) then the latent dependent variable is positive and the purchase is made (the observed discrete outcome is 1), and vice versa. If the researcher makes a further assumption that the disturbance term in the model with the latent dependent variable has a standard normal distribution, then the univariate probit model results. The extension to the multivariate case is relatively straightforward.

The latent variables are clearly not observed, but their distributions are specified to be normal. Chib and Greenberg (1998) use this fact and re-introduce the latent variable back into the multivariate probit model. In a typical Bayesian model the prior distribution of the parameters and the likelihood function are used to obtain the joint posterior distribution, which combines the information from the prior and the data. Chib and Greenberg (1998) find the joint posterior distribution of the multivariate probit model as the product of the prior distribution of the parameters and *augmented* likelihood function. The latter is obtained as the product of normal distributions for latent variables taken over all respondents in the sample. It is easy to show that, after integrating over the latent variables, the joint posterior distribution of the parameters is exactly the same as the posterior distribution obtained without introducing any latent variables (see Koop, Poirier and Tobias [2007] for related examples). The computational advantage of this method — it does not require the evaluation of the truncated multivariate normal density — is the greater the more discrete dependent variables are included into the model.

Using the full conditional posterior distributions of the coefficient vector, along with elements in the variance matrix and the latent data, it is possible to construct a Markov Chain Monte Carlo (MCMC) algorithm and simulate the parameters jointly with the latent data. In the Chib and Greenberg (1998) formulation, the conditional posterior distribution for the elements in the variance matrix has a nonstandard form and the authors use a Metropolis-Hastings algorithm to draw those elements. This paper modifies the Chib and

Greenberg (1998) procedure by using the Cholesky factorization of the variance matrix. This allows a convenient multivariate normal representation of the parameters that are used to obtain the variance matrix, which considerably facilitates estimation.

Another complication in the sample selection model follows from the fact that some of the dependent binary variables in the outcome equation are not observed given the selection rule into the sample. The posterior distribution of the latent data can be used to simulate those missing observations conditional on the covariance structure of the disturbance term. Consider first an individual t with complete data in $m \times 1$ vector of binary responses $y_{.t} = (y_{1t}, ..., y_{mt})'$ for all selection and outcome equations. The Chib and Greenberg (1998) procedure implies that at each MCMC simulation the latent vector $\widetilde{y}_{.t} = (\widetilde{y}_{1t}, ..., \widetilde{y}_{mt})'$ is drawn from the truncated multivariate normal distribution with a $m \times 1$ mean vector and $m \times m$ covariance matrix Σ.[4] The distribution is truncated for the ith element $\widetilde{y}_{it}$ to $(-\infty, 0]$ if the binary outcome $y_{it} = -1$ and to $(0, +\infty)$ if $y_{it} = 1$. Now suppose that individual t has missing binary outcome y_{it} for some i. The only difference with the case of an observed binary outcome y_{it} comes from the fact that the conditional multivariate normal distribution for $\widetilde{y}_{it}$ is no longer truncated in the ith dimension. That is, if y_{it} is missing for some i, then the latent variable $\widetilde{y}_{it}$ is unrestricted and can take any value in the interval $(-\infty, \infty)$.

Identification of the parameters is an important issue in models of discrete choice. It is well-known that the multivairate probit model is not likelihood-identified with unrestricted covariance matrix. Even though the formulation of the variance matrix in this paper uses only $m(m-1)/2$ identified parameters, this turns out not to be sufficient for identification. Meng and Schmidt (1985) offer an elegant treatment of the problem of identification in the censored bivariate probit model using the general principle in Rothenberg (1971) that the parameters in the model are (locally) identified if and only if the information matrix is nonsingular. The conclusion in Meng and Schmidt (1985), that the bivariate probit model with sample selection is in general identified, applies also with my parameterization of the model.

This paper is organized as follows. Section 2 reviews the literature on sample selection especially on extensions to models with discrete outcome equation and Bayesian treatment. Section 3 sets up the model and derives the details of the multivariate probit estimator.

[4] The mean vector for individual t is a product of $m \times k$ matrix of covariates and a $k \times 1$ vector of coefficients to be defined later.

Section 4 develops the Gibbs sampler. Section 5 considers the problem of identification in greater detail. Finally, section 6 provides an illustrative example and the last section concludes the discussion.

2 Discrete Outcome Variable in Sample Selection Model

2.1 Classical treatment of a discrete outcome variable

The model of incidental truncation, which is another name for sample selection model, has been widely used in economic applications when the variable of interest is observed only for people who are selected into a sample based on some threshold rule. Heckman's (1979) treatment of sample selection model is also a standard topic in most modern econometric textbooks (such as Greene 2003 and Wooldridge 2002).

However, there are relatively few applications of Heckman's (1979) model to discrete (and count) data and Greene (2008) reviews a handful of such models, starting with Wynand and van Praag (1981). In a recent application to teen employment, Mohanty (2002) uses the formulation in Meng and Schmidt (1985), which is very similar to the bivariate probit model with sample selection in Wynand and Praag (1981). In Mohanty (2002) the applicant i for a job can be selected ($SEL_i = 1$) or not ($SEL_i = 0$) only if she has applied for a job ($SEEK_i = 1$). Both discrete variables are modeled as the latent variables y_{1i} ($SEEK_i = 1$ if $y_{1i} > 0$ and $SEEK_i = 0$ otherwise) and y_{2i} ($SEL_i = 1$ if $y_{2i} > 0$ and $SEL_i = 0$ otherwise) that have bivariate normal distribution with correlation coefficient ρ.

Estimating the hiring equation (SEL_i) only for the subsample of teens who applied for a job ($SEEK_i = 1$) produces inconsistent estimates as long as $\rho \neq 0$. Indeed, univariate probit shows misleading evidence of employment discrimination against Black teens, which disappears when participation and hiring equations are estimated jointly (Mohanty 2002).

Another relevant example in classical econometrics is Greene (1992), who refers to an earlier paper by Boyes, Hoffman and Low (1989). The (part of the) model in Greene (1992) is bivariate probit where the decision to default or not on a credit card is observed only for cardholders (and not the applicants that were rejected by a credit card company).

Terza (1998) is another important reference in this literature. He develops a model for count data that includes an endogenous treatment variable. For example, the number of trips by a family (count variable of interest) may depend on the dummy for car ownership (potentially endogenous). In this case the dependent variable for car ownership in the first equation appears as explanatory variable in the equation for the number of trips and the two equations are estimated jointly. Terza (1998) compares three estimators for this model: full information ML, non-linear weighted least squares (NWLS) and a two-stage method of moments (TSM) similar to Heckman's (1979) estimator.[5]

The setup in Terza (1998) can be potentially used in models of discrete choice with sample selection, as in a recent paper by Kenkel and Terza (2001). Kenkel and Terza (2001) use a two-step estimator in the model of alcohol consumption (number of drinks) with an endogenous dummy for advice (from a physician to reduce alcohol consumption). The first stage is univariate probit for receiving advice and the second stage applies non-linear least squares to the demand for alcohol (number of drinks). Kenkel and Terza (2001) find that advice reduces alcohol consumption in the sample of males with hypertension, and the failure to account for the endogeneity of advice would mask this result.

Munkin and Trivedi (2003) discuss the problems with different estimators of selection models with discrete outcome equation. The first class of models, which uses moment-based procedures, results in inefficient estimates and does not allow the estimation of the full set of parameters in the presence of correlated multiple outcomes. A second possibility is a weighted nonlinear instrumental variable approach that has not been very successful because of difficulties in consistent estimation of weights (Munkin and Trivedi 2003). Finally, simulated maximum likelihood method requires a sufficient number of simulations for consistency where it is not clear what is "...the operational meaning of sufficient" (Munkin and Trivedi 2003, p. 198).

It seems that none of the models discussed so far allows multiple correlated discrete dependent variables in the presence of sample selection (except for the bivariate case). The approach that I adopt in this paper is to extend the Bayesian multivariate probit model in Chib and Greenberg (1998), allowing for some missing responses. I review existing Bayesian treatments of sample selection in the next subsection and then provide further details on Chib and Greenberg (1998).

[5]The estimators are listed in the order of decreasing efficiency and computational difficulty. NWLS estimator may result in correlation coefficient being greater than one in absolute value.

2.2 Bayesian treatment of Heckman model

Recent Bayesian treatments of sample selection model are almost exclusively based on *Markov Chain Monte Carlo* (MCMC) methods with *data augmentation*.[6] The idea of data augmentation was introduced by Tanner and Wong (1987), and used in Bayesian discrete choice models starting (at least) from Albert and Chib (1993).[7] Latent variables in these models are treated as additional parameters and are sampled from the joint posterior distribution. In these models, however, the joint posterior distribution for parameters and latent variables typically does not have a recognizable form. *Gibbs sampler* is an MCMC method used when the joint posterior distribution can be represented as a full set of (simpler) conditional distributions. It is possible then to obtain the sample from the joint posterior distribution by iteratively drawing from each conditional distribution, *given* the values obtained from the remaining distributions. The model developed herein shares the two aforementioned features (data augmentation and Gibbs sampling) and simultaneous equation structure with previous studies by Li (1998), Huang (2001) and van Hasselt (2008).

Li (1998) develops Bayesian inference in the following simultaneous equation model with limited dependent variables (SLDV):

$$\begin{aligned} y_1^* &= y_2\gamma_1 + X_1\delta_1 + u_1 \\ y_2^* &= X_2\delta_2 + u_2, \end{aligned} \tag{1}$$

where y_1^* is of Tobit type (a researcher observes $y_1 = y_1^*$ if $y_1^* > 0$ and $y_1 = 0$ otherwise) and y_2^* is of probit type (the researcher observes $y_2 = 1$ if $y_2^* > 0$ and $y_2 = 0$ otherwise).[8] The vector of disturbances $(u_1, u_2)'$ is assumed to follow bivariate normal distribution with the variance of u_2 set to 1 for model identification:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix},$$

[6] Earlier developments in Bayesian statistics model selection by means of various weight functions. For example, Bayarri and DeGroot (1987 and four other papers, as cited in Lee and Berger 2001) mostly concentrate on indicator weight function: potential observation is selected into a sample if it exceeds a certain threshold. Bayarri and Berger (1998) develop nonparametric classes of weight functions that are bounded above and below by two weight functions. Lee and Berger (2001) use the Dirichlet process as a prior on the weight function.

[7] Notice that the selection equation in a Heckman-type model is univariate probit.

[8] To avoid the confusion with my parameters later on, I use different Greek letters from those used in the original papers throughout the literature review.

where σ_{11}^2 is the variance of u_1 and σ_{12} is the covariance between u_1 and u_2. Decomposing the joint bivariate distribution of $(u_1, u_2)'$ into the product of the marginal distribution of u_2 and the conditional distribution of $u_1|u_2$ allows convenient blocking in the Gibbs sampler. This decomposition in Li (1998), together with the more convenient reparametrization of the variance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 + \sigma_{12}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix},$$

appear repeatedly in later studies. With these changes the model is now re-defined as

$$\begin{aligned} y_1^* &= y_2\gamma_1 + X_1\delta_1 + u_2\sigma_{12} + v_1 \\ y_2^* &= X_2\delta_2 + u_2, \end{aligned} \tag{2}$$

with $u_2 = y_2^* - X_2\delta_2$, $\sigma^2 = \sigma_{11}^2 - \sigma_{12}^2$, and $v_1 \sim N(0, \sigma^2)$. In the resulting Gibbs sampler with data augmentation, all conditional distributions have recognizable forms that are easy to draw from (multivariate normal, univariate truncated normal and gamma).[9]

Huang (2001) develops Bayesian seemingly unrelated regression (SUR) model, where dependent variables are of the Tobit type (researcher observes $y_{ij} = y_{ij}^*$ if $y_{ij}^* > 0$ and $y_{ij} = 0$ otherwise). The Gibbs sampler with data augmentation in Huang (2001) consists of multivariate normal, Wishart and truncated multivariate normal distributions.

In the paper by van Hasselt (2008), two sample selection models — with unidentified parameters and with identified parameters — are compared.[10] The idea behind the first model is borrowed from McCulloch and Rossi (1994), who used a similar approach in multinomial probit context. The output from the Gibbs sampler is used to approximate the posterior distribution of the identified parameters. The model with identified parameters in van Hasselt (2008) uses marginal-conditional decomposition of the disturbance terms together with

[9] Chakravarti and Li (2003) apply this model to estimate dual trade informativeness in futures markets. Probit equation estimates a trader's decision to trade on her own account and tobit equation measures her (abnormal) profit from her own account trading. Chakravarti and Li (2003) did not find significant correlation between a dual trader's private information and her abnormal profit.

[10] Another interesting paper by van Hasselt (2005) compares the performance of sample selection and two-part models (when two equations are estimated independently) in a Bayesian setup. In classical econometrics Leung and Yu (1996) provide conclusive evidence against negative results in Manning, Duan and Rogers (1987) who claim that two-part model performs better than sample selection model even when the latter is the true model. Leung and Yu (1996) show that problems with sample selection model are caused by a critical problem in the design of experiments in Manning, Duan and Rogers (1987).

more convenient parameterization of the variance matrix, as in Li (1998).[11] The major contribution of van Hasselt (2008) is relaxing the normal distribution assumption in the sample selection model via mixture of normal distributions. I do not follow that route and my model remains fully parametric.

In all the papers cited above the outcome variable is continuous and not discrete. There are two Bayesian papers with discrete outcome variable (and multiple outcome equations) that are worth mentioning: Munkin and Trivedi (2003) and Preget and Waelbroeck (2006).

Munkin and Trivedi (2003) develop a three-equation model with the first equation for count data (the number of doctor visits), the second equation for a continuous variable (the associated health expenditures) and the third equation for a dummy variable (the type of health insurance plan). The selection problem — demand for health care that potentially depends on the type of health insurance — is modeled by using an (endogenous) dummy variable for private health plan. There is no problem of missing dependent variable for respondents that are not in the sample (i.e. who did not purchase private insurance). Neither of the correlation coefficients for private health plan with two variables of interest is statistically different from zero and the type of insurance does not affect the level of health care use (Munkin and Trivedi 2003).[12]

Preget and Waelbroeck (2006) develop a three-equation model with application to timber auctions. There are two binary dependent variables (if a lot received any bids and, conditional on receiving at least one bid, if a lot received two or more bids) and one continuous variable (highest bid for a lot) with an endogenous dummy variable for the number of bids. Preget and Waelbroeck (2006) comment that in such models the likelihood function is not always well behaved, especially in the direction of the correlation coefficients.[13] While in Preget and

[11] McCulloch, Polson and Rossi (2000) show that fully identified multinomial probit model comes at a cost: higher autocorrelation in the Markov Chain.

[12] In a later work, Deb, Munkin and Trivedi (2006), perhaps dissatisfied with a sample selection model, use a two-part model with endogeneity in a similar context.

[13] Consider the following sequential probit model: the second binary outcome is missing for all respondents whose first outcome is "No." The third binary outcome, if present, is missing for all respondents who answered "No" in the second equation and so on. Waelbroeck (2005) argues that in this model the likelihood function is not globally concave and flat in some directions, which limits practical applicability of the model. Notice that in a two-equation case, sequential probit is the same model as censored probit except that the two models may have different interpretation. Keane (1992) discusses similar computational issues in multinomial probit model.

Waelbroeck (2006) the correlation coefficients are never statistically different from zero, they find that their Bayesian algorithm "...yields a remarkably stable coefficient for the binary endogenous variable and was able to deal with irregularities in the likelihood function."

Two conclusions seem to follow from my review of relevant studies. First of all, there exist serious computational difficulties when the sample selection model with multiple dichotomous dependent variables is estimated by methods of classical econometrics. For example, Munkin and Trivedi (2003) comment on difficulties associated with estimating their model in a simulated maximum likelihood framework. This provides strong motivation for a Bayesian econometric methodology and also explains why models similar to mine are typically estimated in a Bayesian and not classical tradition. Second, even in the Bayesian literature, there seem to be no published papers that can be used directly to estimate a model with three or more dichotomous dependent variables. This constitutes an important contribution of the current paper.

While my work shares the methods with previous studies (data augmentation, Gibbs sampling and simultaneous equation structure) it comes from a different area — multivariate probit model developed in Chib and Greenberg (1998). The next section introduces the multivariate probit in Chib and Greenberg (1998) and provides the extensions that make it applicable in the sample selection model.

3 Multivariate Probit and Sample Selection

Suppose that a researcher observes a set of potentially correlated binary events $i = 1, ..., m$ over an independent sample of $t = 1, ..., T$ respondents. Consider the multivariate probit model reformulated in terms of latent variables as in Chib and Greenberg (1998). For each of the events $i = 1, ..., m$ define a $T \times 1$ vector of latent variables $\widetilde{y}_{i.} = (\widetilde{y}_{i1}, ..., \widetilde{y}_{iT})'$ and a $T \times k_i$ matrix of explanatory variables Z_i where each row t represents a $1 \times k_i$ vector Z_{it}. Then each latent variable can be modeled as

$$\widetilde{y}_{i.} = Z_i \beta_i + \varepsilon_{i.}, \tag{3}$$

where $\varepsilon_{i.}$ is a vector of disturbance terms that have normal distribution. There is potential correlation in the disturbance terms for respondent t across events $i = 1, ..., m$ coming from some unobserved factor that simultaneously affects selection and outcome variables. Let

$\widetilde{y}_{.t} = (\widetilde{y}_{1t}, ..., \widetilde{y}_{mt})'$ be the vector of latent variables for respondent t such that

$$\widetilde{y}_{.t} \sim N_m(Z_t\beta, \Sigma), \tag{4}$$

where $Z_t = diag(Z_{1t}, ..., Z_{mt})$ is an $m \times k$ covariate matrix, $\beta_i \, \epsilon \, R^{k_i}$ is an unknown parameter vector in equation $i = 1, ..., m$ with $\beta = (\beta_1', ..., \beta_m')' \, \epsilon \, R^k$ and $k = \sum_{i=1}^m k_i$, and Σ is the variance matrix.

The sign of $\widetilde{y}_{it}$ for each dependent variable $i = 1, ..., m$ uniquely determines the observed binary outcome y_{it}:

$$y_{it} = I(\widetilde{y}_{it} > 0) - I(\widetilde{y}_{it} <= 0) \quad (i = 1, ..., m), \tag{5}$$

where $I(A)$ is the indicator function of an event A. Suppose it is of interest to evaluate the probability of observing a vector of binary responses $Y_{.} = (Y_1, ..., Y_m)'$ for indivial t. Chib and Greenberg (1998) show that the probability $y_{.t} = Y_{.t}$ can be expressed as

$$\int_{B_{mt}} ... \int_{B_{1t}} \phi_m(\widetilde{y}_{.t} | Z_t\beta, \Sigma) d\widetilde{y}_{.t}, \tag{6}$$

where $B_{it} \, \epsilon \, (0, \infty)$ if $y_{it} = 1$ and $B_{it} \, \epsilon \, (-\infty, 0]$ if $y_{it} = -1$. Define $B_t = B_{1t} \times ... \times B_{mt}$.

Alternatively, the probability $y_{.t} = Y_{.t}$ can be expressed without introducing latent variables as

$$pr(y_{.t} = Y_{.t} | \beta, \Sigma) = \int_{A_{mt}} ... \int_{A_{1t}} \phi_m(w|0, \Sigma) dw, \tag{7}$$

where $\phi_m(w|0, \Sigma)$ is the density of a m-variate normal distribution and A_{it} is the interval defined as

$$A_{it} = \begin{cases} (-\infty, Z_{it}\beta_i) & \text{if } y_{it} = 1, \\ (Z_{it}\beta_i, \infty) & \text{if } y_{it} = -1. \end{cases}$$

The multidimensional integral over the normal distribution in (7) is hard to evaluate by conventional methods.[14]

Instead of evaluating this integral, Chib and Greenberg (1998) use the formulation in (6) and simulate the latent variable $\widetilde{y}_{.t}$ from the conditional posterior distribution with mean $Z_t\beta$ and variance matrix Σ. This distribution is truncated for the ith element to $(-\infty, 0]$ if

[14]Quadrature method is an example of nonsimulation procedure that can be used to approximate the integral. Quadrature operates effectively only when the dimension of integral is small, typically not more than four or five (Train 2003). The GHK simulator is the most widely used simulation method after Geweke (1989), Hajivassiliou (as reported in Hajivassiliou and McFadden 1998) and Keane (1994).

the observed outcome is $y_{it} = -1$ and to $(0, +\infty)$ if $y_{it} = 1$. The current model also assumes that $y_{it} = 0$ when the response for event i is missing for t.

It is important to understand what missing binary response means in terms of the latent data representation. If respondent t has missing binary response y_{it} for some i then no restriction can be imposed on the latent normal distribution in the ith dimension. Then the vector $\widetilde{y}_{\cdot t}$ is simulated from the m-variate normal distribution with the same mean and variance as in the complete data case but the distribution is not truncated for the ith element. For the case of missing outcome i the latent variable $\widetilde{y}_{it}$ can take any value in the interval $(-\infty, \infty)$.[15]

The multivariate model of incidental truncation can not be estimated using only the observed data because the endogenous selection variables are constant and equal to 1. Now, due to simulated missing data one can estimate the variance matrix Σ, which is the focus of the procedure to account for sample selection. The covariances in Σ effectively adjust for sample selectivity in the outcome equations by controlling for unobserved heterogeneity.

The issue of sample selection arises whenever the unobserved factors determining the inclusion in the sample are correlated with the unobservables that affect the outcome variable(s) of primary interest (Vella 1998). The critical idea in the current work is to account for selection in binary outcome equation(s) by jointly estimating selection and outcome equations while controlling for possible unobserved effect through multivariate probit with correlated responses. If the covariance terms belong to the highest posterior density region, this indicates the presence of unobserved effect and, hence, sample selection bias.

The elements in the variance matrix in the Chib and Greenberg (1998) formulation do not have the conditional posterior distribution of a recognizable form, which forces them to employ a Metropolis-Hastings algorithm. This paper makes the technical advance that allows convenient multivariate normal representation of the parameters used to obtain the variance matrix. Consider the Cholesky factorization of the inverse of the variance matrix $\Sigma^{-1} = \breve{F} \cdot \breve{F}'$ where $\breve{F}$ is the lower triangular matrix. If the diagonal elements of $\breve{F}$ are arrayed in a diagonal matrix Q then $\Sigma^{-1} = \breve{F}Q^{-1}Q^2Q^{-1}\breve{F}' = FQ^2F$ (Greene 2003). In the current work the variance matrix is defined by F which is a lower triangular matrix that has

[15] This methodology allows for continuous endogenous variables as well. In this case $\widetilde{y}_{jt}$ is trivially set to the observed y_{jt} for a continuous variable j in each iteration of the MCMC algorithm introduced below.

ones on the main diagonal and $D^{-1} = Q^2$ which is a diagonal matrix. Then

$$\Sigma = (F')^{-1} D F^{-1}, \tag{8}$$

with $D = diag\{d_{11}, ..., d_{mm}\}$ and F is lower triangular

$$F = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{21} & 1 & 0 & \cdots & 0 \\ f_{31} & f_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & f_{m3} & \cdots & 1 \end{pmatrix}.$$

Finally, consider the system of m equations

$$\underbrace{\widetilde{y}}_{Tm\times 1} = \begin{pmatrix} \widetilde{y}_{1.} \\ \widetilde{y}_{2.} \\ \vdots \\ \widetilde{y}_{m.} \end{pmatrix}, \quad \underbrace{Z}_{Tm\times k} = \begin{bmatrix} Z_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & Z_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & Z_m \end{bmatrix}, \quad \underbrace{\beta}_{k\times 1} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix},$$

so that the model can be represented as

$$\widetilde{y} = Z\beta + \varepsilon, \tag{9}$$

where $k = \sum_{i=1}^{m} k_i$ and

$$\underbrace{\varepsilon}_{Tm\times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}.$$

Under the maintained assumption of the normally distributed vector ε it follows that

$$\varepsilon | (\beta, F, D, Z) \sim N(0, (F')^{-1} D F^{-1} \otimes I_T). \tag{10}$$

4 Deriving the Gibbs Sampler

Consider a sample of $m \times T$ observations $y = (y_{.1}, ..., y_{.T})$ that are independent over $t = 1, ..., T$ respondents but are potentially correlated over $i = 1, ..., m$ events. Given a prior density $p(\beta, F, D)$ on the parameters β, F and D the posterior density is equal to

$$p(\beta, F, D|y) \propto p(\beta, F, D) p(y|\beta, \Sigma), \tag{11}$$

where $p(y|\beta, \Sigma) = \prod_{t=1}^{T} p(y_{.t}|\beta, \Sigma)$ is the likelihood function. Define $y_{.t} = (y_{st}, y_{ot})$, where y_{st} and y_{ot} are selection and outcome variables with some of the $y_{.t}$'s missing. In this representation the evaluation of the likelihood function is computationally intensive from a classical perspective. Albert and Chib (1993) developed an alternative Bayesian framework that focuses on the joint posterior distribution of the parameters and the latent data $p(\beta, F, D, \widetilde{y}_1, ..., \widetilde{y}_T|y)$. It follows then that

$$\begin{aligned} p(\beta, F, D, \widetilde{y}|y) &\propto p(\beta, F, D)p(\widetilde{y}|\beta, \Sigma)p(y|\widetilde{y}, \beta, \Sigma) \\ &= p(\beta, F, D)p(\widetilde{y}|\beta, \Sigma)p(y|\widetilde{y}). \end{aligned} \tag{12}$$

It is possible now to implement a sampling approach and construct a Markov chain from the distributions $[\widetilde{y}_{.t}|y_{.t}, \beta, \Sigma]$ $(t \leq T)$, $[\beta|y, \widetilde{y}, \Sigma]$ and $[F, D|y, \widetilde{y}, \beta]$.

With unrestricted F or D matrix the multivariate probit model is not identified. The observed outcomes $y_{.t}$ for respondent t depend only on signs but not magnitudes of the latent data $\widetilde{y}_{.t}$. In a multivariate probit model with m equations only $m(m-1)/2$ parameters in the variance matrix are identified. Consider the following transformation of the model $F'\widetilde{y}_{.t} \sim N(F'Z_t\beta, D)$, where D is some unrestricted diagonal matrix. The latent regression has the form $F'\widetilde{y}_{.t} = F'Z_t\beta + D^{1/2}\varepsilon_{.t}$, where $\varepsilon_{.t}$ is m-variate normal with a zero mean vector and an $m \times m$ identity variance matrix. However, pre-multiplying this equation by $\alpha > 0$ results in $\alpha F'\widetilde{y}_{.t} = F'Z_t(\alpha\beta) + \alpha D^{1/2}\varepsilon_{.t}$ which is the same model corresponding to the same observed data $y_{.t}$. Since the parameters in $D^{1/2}$ cannot be identified, D is set to identity matrix extending the logic from the univariate probit model in Greene (2003).[16]

The posterior density kernel is the product of the priors and the augmented likelihood in equation (12).[17] The parameters in β and F are specified to be independent in the prior. Let the prior distribution for β be normal $\phi_k(\beta|\underline{\beta}, \underline{B}^{-1})$ with the location vector $\underline{\beta}$ and the precision matrix $\underline{B}$.

It is convenient to concatenate the vectors below the main diagonal in F matrix as

$$F_{vector} = \begin{pmatrix} F_{2:m,1} \\ F_{3:m,2} \\ \vdots \\ F_{m,m-1} \end{pmatrix},$$

[16] Observe that this is not sufficient for identification and later I give an example from Meng and Schmidt (1985) when the model is not identified with two equations.

[17] The term "augmented likelihood" emphasizes the fact that the likelihood includes latent variables.

where $F_{i+1:m,i}$ for $i = 1, ..., m-1$ represents elements from $i+1$ to m in column i. The prior distribution of F_{vector} is assumed to be $\left(\frac{m(m-1)}{2}\right)$-variate normal

$$F_{vector} \sim N(\underline{F}_{vector}, \underline{H}^{-1}). \tag{13}$$

In this expression $\underline{F}_{vector}$ is the prior mean of the normal distribution, and the prior variance matrix $\underline{H}^{-1}$ is block-diagonal with

$$\underline{H} = \begin{pmatrix} \underline{H}_{2:m,2:m} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \underline{H}_{3:m,3:m} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \underline{H}_{1,1} \end{pmatrix}.$$

This precision matrix has $(m-1) \times (m-1)$ matrix $\underline{H}_{2:m,2:m}$ in the upper left corner and the matrix dimension is decreasing by one in each consequent block on the main diagonal. The lower right matrix $\underline{H}_{1,1}$ is a scalar. The posterior density kernel is now

$$\begin{aligned} & |\underline{B}|^{1/2} \exp\Big\{ -\frac{1}{2}(\beta - \underline{\beta})' \underline{B} (\beta - \underline{\beta}) \Big\} \\ \cdot \quad & |\underline{H}|^{1/2} \exp\Big\{ -\frac{1}{2}(F_{vector} - \underline{F}_{vector})' \underline{H} (F_{vector} - \underline{F}_{vector}) \Big\} \\ \cdot \quad & |\Sigma|^{-T/2} \prod_{t=1}^{T} \exp\Big\{ -\frac{1}{2}(\widetilde{y}_{.t} - Z_t\beta)' \Sigma^{-1} (\widetilde{y}_{.t} - Z_t\beta) \Big\} I(\widetilde{y}_{.t} \in B_t). \end{aligned} \tag{14}$$

A Gibbs sampler is constructed by drawing from the following conditional posterior distributions: the vector of coefficients β, the F_{vector} from the variance matrix decomposition and the latent vector $\widetilde{y}_{.t}$ for each respondent $t \leq T$.[18]

In a typical iteration the Gibbs sampler initiates by drawing the vector of the coefficients β conditional on F_{vector} and $\widetilde{y}_{.t}$ obtained from the previous draw. The posterior distribution of β comes from the posterior density kernel and is normal

$$\beta|(\widetilde{y}, \Sigma) \sim N_k(\beta|\overline{\beta}, \overline{B}^{-1}), \tag{15}$$

where $\overline{B} = \underline{B} + \sum_{t=1}^{T} Z_t' \Sigma^{-1} Z_t$ and $\overline{\beta} = \overline{B}^{-1}(\underline{B}\underline{\beta} + \sum_{t=1}^{T} Z_t' \Sigma^{-1} \widetilde{y}_{.t})$. In this last expression it is understood that for each t , $\widetilde{y}_{.t} \in B_t$.

[18] Technical Appendix A.1 provides complete details of the Gibbs sampler derivation.

To obtain the conditional posterior distribution of F, an alternative expression for the density of $\widetilde{y}$ is useful:

$$\begin{aligned}
p(\widetilde{y}|y,\beta,F,D) &\propto |\Sigma|^{-T/2}\prod_{t=1}^{T}\exp\Big\{-\frac{1}{2}(\widetilde{y}_{.t}-Z_t\beta)'\Sigma^{-1}(\widetilde{y}_{.t}-Z_t\beta)\Big\}I(\widetilde{y}_{.t}\;\epsilon\;B_t)\\
&= |FD^{-1}F'|^{T/2}\prod_{t=1}^{T}\exp\Big\{-\frac{1}{2}\varepsilon_t'FD^{-1}F'\varepsilon_t\Big\}I(\widetilde{y}_{.t}\;\epsilon\;B_t)\\
&= \prod_{t=1}^{T}\prod_{i=1}^{m}\exp\Big\{-\frac{1}{2}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\}\\
&= \prod_{i=1}^{m}\exp\Big\{-\frac{1}{2}\sum_{t=1}^{T}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\},
\end{aligned} \tag{16}$$

where for each t , $\widetilde{y}_{.t}\;\epsilon\;B_t$. In this derivation the restriction $D = I_m$ is already imposed. Then the posterior conditional distribution of F_{vector} is also normal

$$F_{vector}|(y,\widetilde{y},\beta)\sim N_{\left(\frac{m(m-1)}{2}\right)}(\overline{F}_{vector},\overline{H}^{\;-1}). \tag{17}$$

The conditional posterior normal distribution has the posterior precision matrix

$$\overline{H}=\underline{H}+\begin{pmatrix}
\sum_{t=1}^{T}\varepsilon_{t,2:m}\varepsilon'_{t,2:m} & \mathbf{0} & \cdots & \mathbf{0}\\
\mathbf{0} & \sum_{t=1}^{T}\varepsilon_{t,3:m}\varepsilon'_{t,3:m} & \cdots & \mathbf{0}\\
\vdots & \vdots & \ddots & \vdots\\
\mathbf{0} & \mathbf{0} & \cdots & \sum_{t=1}^{T}\varepsilon_{t,m}\varepsilon'_{t,m}
\end{pmatrix}.$$

The posterior mean of the normal distribution is equal to

$$\overline{F}_{vector}=\overline{H}^{-1}\underline{H}\;\underline{F}_{vector}-\overline{H}^{-1}\begin{pmatrix}
\sum_{t=1}^{T}\varepsilon_{t,2:m}\varepsilon_{t,1}\\
\sum_{t=1}^{T}\varepsilon_{t,3:m}\varepsilon_{t,2}\\
\vdots\\
\sum_{t=1}^{T}\varepsilon_{t,m}\varepsilon_{t,m-1}
\end{pmatrix}.$$

Finally, the latent data $\widetilde{y}_{.t}$ are drawn independently for each respondent $t\leq T$ from the truncated multivariate normal distribution as described in Geweke (1991). The algorithm makes draws conditional on Z_t, β and F as well as $\widetilde{y}_{.t}$ obtained in the previous draw. The multivariate normal distribution is truncated to the region defined by the $m\times 2$ matrix $[a,b]$ with a typical row i equal to $(0,\infty)$, if $y_{it}=1$ and $(-\infty,0)$ if $y_{it}=-1$. If y_{it} is not observed, then row i is $(-\infty,\infty)$.

Thus, this work extends Chib and Greenberg (1998) in the following two ways: (i) it permits missing outcome variables $\widetilde{y}_{.t}$, and (ii) it re-parameterizes the variance matrix in terms of more convenient multivariate normal F_{vector} that is used to obtain Σ.

5 The Problem of Identification

Identification is an important issue in models of discrete choice. Meng and Schmidt (1985) in their elegant article offer an excellent treatment of identification in a bivariate probit model under various levels of observability. Meng and Schmidt (1985) rely on the general principle in Rothenberg (1971) that the parameters are (locally) identified if and only if the information matrix is nonsingular. In particular, their *Case Three: Censored Probit* is very similar to the following bivariate sample selection model: the binary variable of interest y_{2t} is observed for respondent t only if she is selected in the sample ($y_{1t} = 1$).[19]

Let $F^t = F(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})$ specify the bivariate normal cumulative distribution function and $\Phi(Z_{ht}\beta_h)$ specify the univariate standard normal cumulative distribution function with $h = 1, 2$ for respondent t. Recall that the sign of $\widetilde{y}_{it}$ perfectly predicts y_{it} and one can write

$$
\begin{aligned}
& p(y|\widetilde{y}) = \prod_{t=1}^{T} I(\widetilde{y}_{1t} > 0)I(\widetilde{y}_{2t} > 0)I(y_{1t} = 1)I(y_{2t} = 1) \\
& +I(\widetilde{y}_{1t} > 0)I(\widetilde{y}_{2t} \leq 0)I(y_{1t} = 1)I(y_{2t} = -1) + I(\widetilde{y}_{1t} \leq 0)I(y_{1t} = -1).
\end{aligned}
$$

The likelihood function in the bivariate model can be obtained after I integrate over $\widetilde{y}$ in the

[19] I employ different parametrization of the variance matrix and, thus, the parameters have to be scaled to be comparable with Meng and Schmidt (1985).

following way

$$\begin{aligned}
\int_B \quad & p(y, \widetilde{y}|\beta, \Sigma) d\widetilde{y} = \int_B p(y|\widetilde{y}, \beta, \Sigma) p(\widetilde{y}|\beta, \Sigma) d\widetilde{y} = \int_B p(y|\widetilde{y}) p(\widetilde{y}|\beta, \Sigma) d\widetilde{y} \\
& = \prod_{t=1}^{T} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big[I(\widetilde{y}_{1t} > 0) I(\widetilde{y}_{2t} > 0) I(y_{1t} = 1) I(y_{2t} = 1) \\
& + I(\widetilde{y}_{1t} > 0) I(\widetilde{y}_{2t} \leq 0) I(y_{1t} = 1) I(y_{2t} = -1) + I(\widetilde{y}_{1t} \leq 0) I(y_{1t} = -1) \Big] \\
& \cdot f(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21}) d\widetilde{y}_{1t} d\widetilde{y}_{2t} \\
& = \prod_{t=1}^{T} \int_{0}^{\infty} \int_{0}^{\infty} I(y_{1t} = 1) I(y_{2t} = 1) f_2 d\widetilde{y}_{1t} d\widetilde{y}_{2t} \\
& + \prod_{t=1}^{T} \int_{-\infty}^{0} \left(\int_{0}^{\infty} I(y_{1t} = 1) I(y_{2t} = -1) f_2 d\widetilde{y}_{1t} \right) d\widetilde{y}_{2t} \\
& + \prod_{t=1}^{T} \int_{-\infty}^{0} I(y_{1t} = -1) \phi(Z_{1t}\beta_1) d\widetilde{y}_{1t} \\
& = \prod_{t=1}^{T} F(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})^{I(y_{1t}=1)I(y_{2t}=1)} \\
& \cdot [\Phi(Z_{1t}\beta_1) - F(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})]^{I(y_{1t}=1)I(y_{2t}=-1)} \\
& \cdot [1 - \Phi(Z_{1t}\beta_1)]^{I(y_{1t}=-1)},
\end{aligned}$$

where $f = f(Z_{1t}\beta_1, Z_{2t}\beta_2; f_{21})$ is bivariate normal density function and $\phi(\cdot)$ is univariate normal density function. Define $q_{it} = \frac{y_{it}+1}{2}$ for $i = 1, 2$ and take the natural logarithm of this expression to obtain

$$\begin{aligned}
\ln L(Z_1\beta_1, Z_2\beta_2; f_{21}) \quad & = \\
\sum_{t=1}^{T} \Big[q_{1t} q_{2t} \ln F^t \quad & + \quad q_{1t}(1 - q_{2t}) \ln[\Phi(Z_{1t}\beta_1) - F^t] + (1 - q_{1t}) \ln[1 - \Phi(Z_{1t}\beta_1)] \Big],
\end{aligned} \tag{18}$$

which is equivalent to equation (6) in Meng and Schmidt (1985) except for the correlation coefficient ρ being replaced by the parameter f_{21} that enters F^t as defined below equation (8). The general conclusion reached by Meng and Schmidt (1985) is that the parameters in this model are identified except in certain "perverse" cases. First of all, peculiar configurations of the explanatory variables may cause nonidentification, but this problem can be addressed only given the data at hand. Second, nonidentification may be caused by certain combinations of parameters in the model. For example, the censored bivariate probit model with my parametrization is not identified when $Z_{1t}\beta_1 = \frac{-f_{21}}{\sqrt{1+f_{21}^2}} Z_{2t}\beta_2$ for all respondents t and I show this result in Technical Appendix A.2.[20] The information matrix is then singular because the row for the second intercept (i.e. for $(k_1+1)th$ term) is the last row (i.e. for the parameter f_{21}), divided by a constant. In this particular example the problem of nonidentification does not arise as long as the set of explanatory variables is not the same in two equations.

Meng and Schmidt (1985) comment that there might also be other combinations of parameters or particular configurations of explanatory variables leading to nonidentification. Since it is not possible to foresee all such problems *a priori*, it is the responsibility of the researcher to check if the parameters in the model are identified. However, it is very reassuring that the sample selection model is generally identified, except in some (not very likely) cases.

6 Experiments with Artificial Data

The purpose of the experiment with artificial data is to study if the model can retrieve the parameters and the correlation coefficient that are used to generate the data when some of the outcome variables are missing. It is also of interest to assess the convergence properties of the model. I construct the following bivariate probit model with sample selection. Let y_{2t} be the dichotomous dependent variable of interest that is observed only if the selection variable y_{1t} is equal to 1.

For this experiment I generate $t = 1, ..., 500$ independent latent variables $(\widetilde{y}_{1t}, \widetilde{y}_{2t})'$ from the bivariate normal distribution with mean $\mu_t = [Z_{1t}\beta_{1,\cdot}\sqrt{1+f_{21}^2}, Z_{2t}\beta_{2,\cdot}]'$, where a 1×3 vector Z_{it} contains intercept, one discrete and one continuous variable as described below and $\beta_{i,\cdot} = [\beta_{i,1}\ \beta_{i,2}\ \beta_{i,3}]'$ for $i = 1, 2$. Each equation contains the intercept denoted $\beta_{i,1}$, continuous variable $\beta_{i,2}$ and discrete variable $\beta_{i,3}$. Continuous variable in each equation is drawn from the normal distribution with $\mu = -0.5$ and $\sigma = 2$. Discrete variable takes values of -1 and 1 with equal probability. All continuous and discrete variables are independent from each other. The coefficients used to generate the artificial data are provided in the

[20] Another example of nonidentification given in Meng and Schmidt (1985) is when there are only intercepts included in all equations. While such a model cannot be used in a meaninful way for economic analysis, it provides an interesting limiting case when all the covariate coefficients go to zero.

second column of Table 1. The correlation coefficient is set to 0.5 with the corresponding value of $f_{21} \approx -0.5774$. Finally, the 2×2 covariance matrix is the same for all respondents and is set to

$$\Sigma = \begin{bmatrix} 1 + f_{21}^2 & -f_{21} \\ -f_{21} & 1 \end{bmatrix}.$$

Observe that the true parameters of the first equation are multiplied by $\sqrt{1 + f_{21}^2}$ and in each simulation I normalize the draws of $\beta_{1,\cdot}$ by $\sqrt{1 + f_{21}^2}$ obtained in the same draw. After I obtain the 500×2 matrix of the latent variables $\widetilde{y}$, I convert it into the matrix of "observed" dichotomous dependent variables y which is used in the simulator. The coefficients that were chosen place approximately one third in each of the three bins (yes, yes), (yes, no) and $(no, missing)$.

The implementation of the Gibbs sampler is programmed in the Matlab environment with some loops written in C language. All the codes successfully passed the joint distribution tests in Geweke (2004). The results in this section are based on 24,000 draws from the posterior (the first 6,000 draws were discarded as burn-in iterations). The prior for $i = 1, 2$ vector of coefficients $\beta_{i,\cdot}$ is mutlivariate normal with the mean vector set to zeros and the variance matrix equal to the identity matrix of dimension 3. The prior for f_{21} is standard normal distribution.

The results of the experiment are shown in Figures 1-3 and Table 1.[21] The simulator works quite well in this experiment with low autocorrelation and stable results with histograms centered almost at the values of the parameters used to generate the data. Geweke's convergence diagnostic test (Geweke 1992) does not indicate problems with the convergence of the Markov Chain. The only slight problem is that the mean of the correlation coefficient ρ in the sample obtained from the joint posterior distribution (0.23) is somewhat lower than the value of 0.5 used to obtain the artificial data but it still belongs to the 95% highest posterior density interval.

7 Concluding Remarks

This paper develops a sample selection model for discrete or mixed continuous-discrete outcomes with multiple outcome and selection equations. To facilitate the estimation of a

[21] To obtain some of the statistics I used the MATLAB program *momentg.m* by James LeSage.

resulting multivariate probit model, a Bayesian reformulation in terms of latent variables is extended from the Chib and Greenberg (1998) paper that offers a convenient simulation procedure aimed at resolving the problems of evaluating the integral of multivariate normal density. The essence of the method is to jointly simulate the parameters and the latent variables from conditional posterior distributions using a Markov Chain Monte Carlo algorithm. If there is any unobserved heterogeneity for each agent t, it is properly accounted for as a part of the disturbance terms by the covariance structure of the variance matrix resulting from a joint estimation of a system of equations.

This paper also makes two technical advances to the Chib and Greenberg (1998) setup by (i) adding some missing binary responses and (ii) simplifying the estimation of the variance matrix via a multivariate normal representation of the elements in the lower triangular matrix from the Cholesky factorization of Σ^{-1}. I also discuss how the results on identification in Meng and Schmidt (1985) apply in the bivariate probit model with sample selection.

In addition to introducing the multivariate probit model with sample selection, this paper also offers some interesting topics for further research. In particular, it might be of interest to further study the identification in the case of three and more equations, which clearly depends on the selection rule into a sample. The likelihood is different in each particular case and extensive study of this topic along the lines of Meng and Schmidt (1985) may be rewarding. Alternatively, some of the potentially interesting topics in empirical health and labor economics outlined in the introduction can be done with little (or no) modification of the model in this paper.

Acknowledgements

I would like to thank Professor Fredric Wolinsky for his help and financial support through NIH grant R01 AG-022913 in the preceding two years. This paper benefited from insightful comments of seminar participants at the Department of Health Management and Policy at the University of Iowa (Iowa City, USA) in October of 2009 and Kyiv School of Economics/Kyiv Economics Institute (Kyiv, Ukraine) in March of 2010. John Geweke suggested the extension of the Chib and Greenberg (1998) paper, provided help on some derivations and also financial support in the Spring of 2009. Michelle Nourski assisted in editing the manuscript. Any remaining errors are mine.

References

Albert, J., Chib S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88 (422), 669-679.

Amemiya, T., 1974. Bivariate probit analysis: Minimum chi-square methods. Journal of the American Statistical Association 69 (348), 940-944.

Ashford, J., Sowden R., 1970. Multi-variate probit analysis. Biometrics (26), 535-546.

Bayarri, M., Berger J., 1998. Robust Bayesian analysis of selection models. Annals of Statistics (26), 645-659.

Bayarri, M., DeGroot M., 1987. Bayesian analysis of selection models. The Statistician (36), 137-146.

Boyes, W., Hoffman, D., Low S., 1989. An econometric analysis of the bank credit scoring problem. Journal of Econometrics (40), 3-14.

Chakravarty, S., Li K., 2003. A Bayesian analysis of dual trader informativeness in future markets. Journal of Empirical Finance 10 (3), 355-371.

Chib, S., Greenberg E., 1998. Analysis of multivariate probit models. Biometrika 85 (2), 347-361.

Deb, P., K. Munkin, Trivedi P., 2006. Bayesian analysis of the two-part model with endogeneity: Application to health care expenditure. Journal of Applied Econometrics (21), 1081-1099.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica (57), 13171339.

Geweke, J., 1991. Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints, in: Keramidas, E. (Ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Fairfax: Interface Foundation of North America, Inc.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of the posterior moments, in: Berger, J., Dawid J., Smith A. (Eds.), Bayesian Statistics 4. Oxford: Clarendon Press.

Geweke, J., 2004. Getting it right: Joint distribution tests of posterior simulators. Journal of the American Statistical Association (99), 799-804.

Greene, W., 1992. A statistical model for credit scoring. WP No. EC-92-29, Department of Economics, Stern School of Business, New York University.

Greene, W., 2003. Econometric Analysis, fifth ed. Prentice Hall, Upper Saddle River.

Greene, W., 2008. Econometric Analysis, sixth ed. Pearson Education, Upper Saddle River.

Gronau, R., 1974. Wage comparisons - a selectivity bias. The Journal of Political Economy 82 (6), 1119-1143.

Hajivassiliou, V., McFadden D., 1998. The method of simulated scores for the estimation of LDV models. Econometrica (66), 863-896.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153-161.

Huang, H.-C., 2001. Bayesian analysis of the SUR tobit model. Applied Economics Letters (8), 617-622.

Keane, M., 1992. A note on identification in the multinomial probit model. Journal of Business and Economic Statistics 10 (2), 193-200.

Keane, M., 1994. A computationally practical simulation estimator for panel data. Econometrica (62), 95-116.

Kenkel, D., Terza, J., 2001. The effects of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. Journal of Applied Econometrics 16 (2), 165-184.

Koop, G., Poirier, D., Tobias J., 2007. Bayesian Econometric Methods, first ed. Cambridge University Press: New York.

Lee, J., Berger, J., 2001. Semiparametric Bayesian analysis of selection models. Journal of the American Statistical Association (96), 1269-1276.

Leung, S., Yu, S., 1996. On the choice between sample selection and two-part models. Journal of Econometrics (72), 197-229.

Li, K., 1998. Bayesian inference in a simultaneous equation model with limited dependent variables. Journal of Econometrics (85), 387-400.

Manning, W., Duan, N., Rogers W., 1987. Monte Carlo evidence on the choice between sample selection and two-part models. Journal of Econometrics (35), 59-82.

McCulloch, R., Polson, N., Rossi P., 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. Journal of Econometrics (99), 173-193.

McCulloch, R., Rossi P., 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics (64), 207-240.

Meng, C., Schmidt P., 1985. On the cost of partial observability in the bivariate probit model. International Economic Review (26), 71-86.

Mohanty, M., 2002. A bivariate probit approach to the determination of employment: a study of teen employment differentials in Los Angeles county. Applied Economics 34 (2), 143-156.

Munkin, M., Trivedi P., 2003. Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare. Journal of Econometrics (114), 197-220.

Ochi, Y., Prentice R., 1984. Likelihood inference in a correlated probit regression model. Biometrika 71 (3), 531-543.

Preget, R., Waelbroeck P., 2006. Sample selection with binary endogenous variable: A Bayesian analysis of participation to timber auctions. Working Paper ESS-06-08, Telecom Paris.

Rothenberg, T., 1971. Identification in parametric models. Econometrica (39), 577-591.

Tanner, M., Wong W., 1987. The calculation of posterior distribution by data augmentation. Journal of American Statistical Association (82), 528-540.

Terza, J., 1998. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. Journal of Econometrics 84 (1), 129-154.

Train, K., 2003. Discrete Choice Methods with Simulation, first ed. Cambridge University Press: New York.

van Hasselt, M., 2005. Bayesian sampling algorithms for the sample selection and two- part models. Working paper, Department of Economics, Brown University.

van Hasselt, M., 2008. Bayesian inference in a sample selection model. Working paper, Department of Economics, The University of Western Ontario.

Vella, F., 1998. Estimating models with sample selection bias: A survey. The Journal of Human Resources 33 (1), 127-169.

Waelbroeck, P., 2005. Computational issues in the sequential probit model: A Monte Carlo study. Computational Economics 26 (2), 141-161.

Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data, first ed. The MIT Press: Cambridge.

Wynand, P., Praag B.V., 1981. The demand for deductibles in private health insurance. Journal of Econometrics (17), 229-252.