Journal of Econometrics 165 (2011) 221–232

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian inference in a sample selection model

Martijn van Hasselt *

Behavioral Health Economics Program, RTI International, 3040 Cornwallis Road,PO Box 12194, 27709 Research Triangle Park, NC, United States

ARTICLE INFO

Article history:
Received 11 November 2008
Received in revised form
29 June 2011
Accepted 1 August 2011
Available online 24 August 2011

JEL classification:
C11
C14
C15
C34

Keywords:
Sample selection
Gibbs sampling
Mixture distributions
Dirichlet process

ABSTRACT

This paper develops methods of Bayesian inference in a sample selection model. The main feature of this model is that the outcome variable is only partially observed. We first present a Gibbs sampling algorithm for a model in which the selection and outcome errors are normally distributed. The algorithm is then extended to analyze models that are characterized by nonnormality. Specifically, we use a Dirichlet process prior and model the distribution of the unobservables as a mixture of normal distributions with a random number of components. The posterior distribution in this model can simultaneously detect the presence of selection effects and departures from normality. Our methods are illustrated using some simulated data and an abstract from the RAND health insurance experiment.

1. Introduction

In this paper we develop methods of Bayesian inference in a sample selection model. In general sample selection occurs when the data at hand is not a random sample from the population of interest. Instead, members of the population may have been selected into (or out of) the sample, based on a combination of observable characteristics and unobserved heterogeneity. In this case inference based on the selected sample alone may suffer from selection bias.

A selection model typically consists of two components. The first is an equation that determines the level of the outcome variable of interest. The second is an equation describing the selection mechanism: it determines whether we observe the outcome or not. The latter can sometimes be given a structural interpretation, in which the dependent variable in the selection equation represents an agent's latent utility. If this utility crosses a certain threshold level, the agent acts in such a way that his or her outcome is observed. If the threshold is not crossed, the agent acts differently and the outcome remains unobserved. Thus, a selection model can be viewed as a model for *potential* outcomes that are only *partially* realized and observed. This interpretation applies most directly to the context of modeling a wage offer distribution. Here the wage offer is a potential outcome that is realized only when an individual actually participates in the labor force.

The importance of selection issues in the analysis of labor markets was recognized early on by, among others, Gronau (1974) and Heckman (1974). In his seminal contribution Heckman (1979) treats sample selection as a potential specification error and proposes a two-step estimator that corrects for omitted variable bias. Both Heckman's two-step procedure and full-information maximum likelihood have since been widely used in applied work, and are readily available routines in many statistical software packages. An obvious problem, however, is that these estimation methods rely on strong parametric assumptions about the distribution of unobservables. When these assumptions are violated the estimators may become inconsistent. To overcome this problem a number of semiparametric methods have been proposed. Examples include Cosslett (1991), Ichimura and Lee (1991), Ahn and Powell (1993) and Lee (1994). An excellent survey of this literature is Vella (1998).

Despite the numerous contributions in classical econometrics, the Bayesian literature on selection models has remained relatively sparse. Bayarri and DeGroot (1987); Bayarri and Berger (1998) and Lee and Berger (2001) consider inference based on a univariate selected sample. More recently, Chib et al. (2009) develop a Bayesian method of inference in regression models that are subject to sample selection and endogeneity of some of the covariates. They consider models that are potentially nonlinear, but have normally distributed structural errors.

* Tel.: +1 919 541 6925; fax: +1 919 485 5555.
E-mail address: mvhasselt@rti.org.

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.08.003

Our paper adds to this literature by developing a Bayesian approach to inference in a type 2 Tobit model (e.g., Amemiya, 1985, Ch. 10). In this model the selection rule is binary: we only observe whether the latent selection variable crosses a threshold or not.[1] Although we do not explicitly treat alternative selection mechanisms, it is relatively easy to modify the methods presented here to cover such cases. We provide Gibbs sampling algorithms that produce an approximate sample from the posterior distribution of the model parameters. Our paper differs from Chib et al. (2009) in that we also consider a model with a flexible distribution for the unobserved heterogeneity (i.e. the residuals or 'errors' in the two model equations). The starting point for our analysis is a bivariate normal distribution. Gibbs sampling in this case is fairly straightforward. The basic model may, of course, be misspecified. We therefore extend the analysis to a semiparametric model, based on the Dirichlet process prior of Ferguson (1973, 1974). This prior implies that the unobserved heterogeneity follows a mixture distribution with a random number of components. It has become increasingly popular in Bayesian semiparametric analyses, and our contribution is to incorporate it into a sample selection framework.[2]

A Bayesian approach to inference has two attractive features. First, the output of the sampling algorithm not only provides the Bayesian analogue of confidence intervals for the model parameters, it also gives an immediate indication of the presence (or absence) of a selection effect and departures from normality. Second, if the econometrician has prior information, e.g. restrictions on the parameters, then this information can be easily incorporated through the prior distribution.

The remainder of this paper is organized as follows. Section 2 presents the selection model with bivariate normal errors and a Gibbs sampling algorithm. In Section 3 we develop the extension to a mixture model. We discuss identification issues, the Dirichlet process prior and present the corresponding algorithm to approximate the posterior. The use of the Dirichlet mixture model is illustrated in Section 4 with some simulated data, whereas Section 5 contains an application to estimating a model for health care expenditures, using an abstract of the RAND health insurance experiment. Section 6 concludes and details regarding the various Gibbs samplers are collected in Appendix. With regard to notation, $\mathcal{N}_k(\mu, \Sigma)$ denotes a k-dimensional normal distribution with mean μ and variance Σ. Unless there is ambiguity about the dimension, we will usually omit the subscript k. We use $\mathcal{TN}_{(a,b)}(\mu, \Sigma)$ to denote a $\mathcal{N}(\mu, \Sigma)$ distribution, truncated to the interval (a, b). The standard normal density and distribution functions are $\phi(\cdot)$ and $\Phi(\cdot)$, respectively. Finally, $\mathcal{G}(c_0, d_0)$ denotes the gamma distribution with parameters (c_0, d_0) and expected value c_0/d_0.

2. A sample selection model

2.1. Likelihood and prior

We use the following selection model for an individual member i of the population:

$$\begin{aligned} s_i^* &= x_{i1}'\beta_1 + u_{i1}, \\ s_i &= \mathbb{I}\{s_i^* > 0\}, \\ y_i &= \begin{cases} x_{i2}'\beta_2 + u_{i2} & \text{if } s_i = 1 \\ \text{missing} & \text{if } s_i = 0, \end{cases} \end{aligned} \tag{1}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. The row vectors x_{i1}' and x_{i2}' contain k_1 and k_2 variables, respectively. If x_i denotes the vector of distinct covariates in (x_{i1}', x_{i2}'), the econometrician observes an i.i.d. sample $\{x_i, y_i, s_i\}_{i=1}^n$ of size n from the population model.[3] Note that the outcome y_i is observed if and only if $s_i = 1$. We define $N_1 = \{i : s_i = 1\}$ and $N_0 = \{i : s_i = 0\}$ as the index sets of the observed and missing outcomes, respectively.

Letting $u_i = (u_{i1}, u_{i2})'$ be the vector of errors, a simple parametric model is obtained when we assume that $u_i|x_i \sim \mathcal{N}(0, \Sigma)$. This rules out the case where some of the covariates in x_i are endogenous. Provided valid instrumental variables are available, the selection model can be expanded with a set of reduced-form equations that relate instruments to endogenous variables. A parametric model then specifies a joint distribution (e.g. multivariate normal) for u_i and the reduced-form errors. This approach to modeling endogeneity is taken by Chib et al. (2009), and can be adapted for the models we discuss in this paper. To save space and keep the notation relatively simple, we do not present such an extension here.

Similar to Koop and Poirier (1997), we parameterize the covariance matrix of u_i as

$$\Sigma = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \xi^2 + \sigma_{12}^2 \end{bmatrix}, \tag{2}$$

where σ_{12} is the covariance and ξ^2 the conditional variance of u_{i2}, given u_{i1}. When the covariance is zero, u_{i1} is independent of y_i^* and we can conduct inference about β_2 based on the subsample indexed by N_1. This strategy would lead to selection bias when $\sigma_{12} \neq 0$. Setting the variance of u_{i1} equal to one is the typical identification constraint for a binary choice model. It should be noted that in a Bayesian treatment of this model it is not necessary to impose this constraint. We could proceed with an unrestricted covariance matrix and conduct inference in a way similar to McCulloch and Rossi (1994). The main difficulty, however, lies with selecting a prior for the unidentified parameters. This prior will induce a prior for the identified parameters, and needs to be carefully checked to ensure that it appropriately reflects a researcher's beliefs. The advantage of the current model formulation is that a prior is placed directly on the identified parameters; see Li (1998) and McCulloch et al. (2000), who proposed this strategy before.

In what follows let $\theta' = (\beta_1', \beta_2', \sigma_{12}, \xi^2)$ be the vector of model parameters. For the observed outcomes we know that $y_i|\theta \sim \mathcal{N}(x_{i2}'\beta_2, \xi^2 + \sigma_{12}^2)$.[4] It follows from the bivariate normality assumption that

$$\Pr\{s_i = 1|y_i, \theta\} = \Phi\left(x_{i1}'\beta_1\sqrt{1 + \sigma_{12}^2/\xi^2} + \frac{\sigma_{12}(y_i - x_{i2}'\beta_2)}{\xi\sqrt{\xi^2 + \sigma_{12}^2}}\right).$$

On the other hand, when the outcome is missing it does not contribute to the likelihood, and the probability that this occurs is $\Pr\{s_i = 0|\theta\} = 1 - \Phi(x_{i1}'\beta_1)$.

If y and s are the n-dimensional sample vectors of (y_i, s_i) values, the likelihood is given by

$$\begin{aligned} f(y, s|\theta) = &\prod_{i \in N_0}[1 - \Phi(x_{i1}'\beta_1)] \\ &\times \prod_{i \in N_1}(\xi^2 + \sigma_{12}^2)^{-1/2}\phi\left(\frac{y_i - x_{i2}'\beta_2}{\sqrt{\xi^2 + \sigma_{12}^2}}\right) \\ &\times \Phi\left(x_{i1}'\beta_1\sqrt{1 + \sigma_{12}^2/\xi^2} + \frac{\sigma_{12}(y_i - x_{i2}'\beta_2)}{\xi\sqrt{\xi^2 + \sigma_{12}^2}}\right). \end{aligned} \tag{3}$$

[1] In some cases the selection process contains more information. Lee (1994) uses a Tobit model for the selection equation. The outcome of interest is then observed based on a selection variable which itself is partially observed.

[2] A recent example is Conley et al. (2008) who use the Dirichlet process prior in an instrumental variable model.

[3] We allow for x_{i2} to be unobserved as well when $s_i = 0$. However, x_{i1} is observed for all sampling units.

[4] Throughout this paper we will omit conditioning on x_i for notational simplicity.