MULTIVARIATE AND MIXTURE EXTENSIONS OF THE TOBIT MODEL

J. Brůha

Czech National Bank

The original Tobit model has been proposed for dealing with observations censored at zero, i.e., it can be used to describe the relationship between a non-negative dependent variable y and covariates (regressors) x. In this paper, I propose extensions of the original model for multivariate data and for treating unobserved heterogeneity. I show how these extended models can be estimated using Bayesian techniques (Gibbs sampling) and I provide some practical hints for the estimation in Matlab. I also outline two selected applications.

The multivariate Tobit model is defined as follows: let Y be a vector of non-negative numbers, with covariates X. The assumed data generating process for observation Y is:

$$Y = \max(Y^*, \mathbf{0}), \tag{1}$$

where Y^* is a random vector with the multivariate normal (henceforth MVN) distribution with mean $X\beta$ and covariance matrix Σ, $\mathbf{0}$ is the vector of zeros, and the operator max is applied component-wise. The goal of estimation is to estimate the parameters β and Σ, which then fully characterize the conditional distribution of the latent variable Y^* and the observed variable Y (conditional on X).

The likelihood function associated with Model (1) is complicated (it requires evaluation of a nasty integral), and therefore its direct maximization is difficult and time consuming. However, the latent-data form of the model suggests the Gibbs sampler as an estimator. The idea is simple: if the latent variables Y^* are observed, then the Bayesian estimation of the parameters β and Σ is basically the estimation of the seemingly unrelated regression (SUR) model, and there are many efficient algorithms for Bayesian estimation of the SUR model. However, if the parameters β and Σ are known, then it is possible to sample Y^* conditional on observations Y using another Gibbs sampler. In the paper, I discuss details how to do that efficiently.

Nevertheless, the multivariate model (1) need not be always a satisfactory. Sometimes, data manifest unobserved heterogeneity, which cannot be sufficiently described by observed covariates X and Gaussian errors with fixed covariance matrix Σ. For such a case, I propose a mixture extension. The latent variable Y^* is given as:

$$Y^* = X\beta_s + u_s, \text{ with probability } \pi_s \tag{2}$$

where β_s is one of S possible vectors of regression coefficients, the random disturbances u_s have zero mean and the covariance matrix Σ_s, and $\sum_s \pi_s = 1$. The observe variable Y is still obtained by (1), however, there are S possible models for the latent variable Y^*. The goal is to make statistical inference about S vectors $\{\beta_s\}_{s=1}^S$, covariance matrices $\{\Sigma_s\}_{s=1}^S$, and probabilities $\{\pi_s\}_{s=1}^S$.

I propose another Gibbs sampler to estimate the mixture extension of the Tobit model (2). The experience with estimation of the model on real data suggests that the algorithm needs either a lot of data or a very informative prior distribution for parameters $\{\beta_s\}_{s=1}^S$. I discuss a way of obtaining such prior.

Finally, I briefly describe two applications in econometrics. Matlab codes for the two models are available from the author.