青森公立大学経営経済学研究
第15巻第１号（2009）

Understanding Likelihood Functions for Sample Selection Models

Hiroshi Murao ※

The users of sample selection models such as the type Ⅱ and Ⅲ Tobit modes are likely to have difficulties for understanding likelihood functions for these models. There are several reasons behind it. First, resulting formulas look quite different depending on authors. Beside this point, it is possible to derive quite different variations. Without such knowledge the users of the model wonder which one is right. Second, some likelihood functions are written in abstract form and they are difficult to write in computer codes from the view point of practitioners. And so on. We show that two theorems are important in order to understand these likelihood functions. Once we understand how to apply these theorems in the context of sample selection problem, we can easily derive these likelihood functions with some variations. With such knowledge, we can easily modify the likelihood function of a standard specification based on our needs. Such knowledge is also applied to more complicated models including simultaneous equation systems and the like.

1. Introduction

Sample selection models such as the type Ⅱ and Ⅲ Tobit modes have been widely used in many fields of quantitative statistical analysis including applied Econometrics. These models are used in the context of "sample selection problem" or "self-selectivity problem." Suppose that data are missing on one or more variables for some units in a sample. Using a subset of a random sample because of missing data faces to the sample selection problem if the sample separation between missing data and observed data is not random.

Suppose we are interested in estimating a wage offer equation for married women, and we get survey data from randomly selected married women. We get missing data on wage from non-working married women while we get observed data on wage from working married women. Using the set of observed data on wage for estimating the wage offer equation faces to the sample selection problem since the data set of working married women only is not a random sample of the underling population (i.e., married women). The usual OLS estimation provides neither unbiased estimation nor consistent estimation for the wage offer equation. Here, married women make a decision whether they work or not based on their rational choice, i.e., the division between working and non-working is based on a certain rule and it is endogenous. Thus, the set of only working married women is not based on random sampling with regard to the population of married women.

Sample selection models such as the type Ⅱ and Ⅲ Tobit modes deal with this kind of problem by modeling an endogenous selection rule for the division of data. The result of selection rule can be described by a binary variable. For example, the binary variable takes

※ Aomori Public College Associate Professor

the value of one if wage is observed while it takes the value of zero if wage is not observed. Then we can construct a binary regression model for the selection rule in order to deal with the self-selectivity problem. The type Ⅱ Tobit mode is based on such a binary selection equation. The result of selection rule might be described by a censored quantitative variable. For example, the dependent variable of the selection equation have positive values of working hours for working women while zero or missing value for the non-working women. Then we can construct a censored regression model for the selection rule in order to deal with the self-selectivity problem. The type Ⅲ Tobit mode is based on such a censored selection mechanism.

The type Ⅱ and Ⅲ Tobit modes are often estimated with the maximum likelihood method, and their log likelihood functions can be found in books including Wooldridge (2002, p. 566), Gourieroux (2000, p.191), Maddala (1983, p.266), and Dhrymes (1986, p.1612). However, the users of these models might have difficulties for understanding these authors' likelihood functions. There are several reasons behind it. First, resulting formulas look quite different depending on authors. As we will see, quite different variations are also possible. Without such knowledge the users of the model wonder which one is right. Second, some likelihood functions are written in abstract form and they are difficult to write in computer codes from the view point of practitioners. Third, the resulting formulas are often written with little explanation except some authors including Wooldridge (2002, p. 566).

Beside difficulties of understanding, there is another kind of problem. If a model specification is different from the standard model specification, then many practitioners including myself might feel difficulties of how to modify the likelihood function. Thus, it is important to understand how these likelihood functions are derived and the meanings for the parts of these likelihood functions. Without suck knowledge, it is almost impossible to modify them based on our needs.

This paper shows that there are two important theorems to understand how to construct the likelihood functions for sample selection models with focusing on the type Ⅱ Tobit model. With this kind of knowledge, it is not difficult to modify them based on our needs.

The rest of the paper is organized as follows. Section 2 describes the type Ⅱ Tobit model with the standard model specification. It introduces a popular formula of the likelihood function for the type Ⅱ Tobit model. Section 3 shows two important theorems to understand how to construct the likelihood functions for sample selection models. Derived are two kinds of likelihood functions for the type Ⅱ Tobit model. Section 4 describes the type Ⅲ Tobit model with the standard model specification. Two kinds of likelihood functions are also derived for the type Ⅲ Tobit model. Section 5 considers a few models for program evaluation since one major use of the sample selection models is in evaluating the benefits of social programs. Section 6 considers the general model of sample selection which covers all of the above models. Section 7 concludes.

2. The Type Ⅱ Tobit Model

The sample selection models deal with "sample selection problem" or "self-selectivity problem" by modeling an endogenous selection rule for the division of data. The results of the selection rule could be binary outcomes such as "working" and "not-working." The type Ⅱ Tobit mode is based on such a binary selection mechanism for the division of data.

The type Ⅱ Tobit mode has the following standard specification:

$$y_1 = x_1\beta_1 + u_1 \tag{1a}$$

$$d = 1[x_2\beta_2 + u_2 > 0] \tag{1b}$$

where $1[\bullet]$ is the indicator function, which is written as

$$y_2^* = x_2\beta_2 + u_2 \tag{2a}$$

$$d = \begin{cases} 1 & \text{if} \quad y_2^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2b}$$

where y_2^* is a latent variable. Equation (1a) is called the main regression equation, and it describes how the quantitative dependent variable y_1 is influenced by a vector of explanatory variables x_1 and its error term u_1. Equation (1b) is called the selection equation or the like, and it describes how a binary variable d, which descries the separation of values on y_1 into two regimes or groups, is influenced by a vector of explanatory variables x_2 and its error term u_2. We assume (x_2, d) are always observed while y_1 is observed only when $d=1$. Since many authors of textbooks assume x_1 is always observed, we also assume x_1 is always observed. This means that the main regression equation is a censored regression model. There is the following fact behind this assumption. Since the selection equation is not typically a structure equation, x_2 may contain all variables in x_1. Let $y = (y_1, d)'$, $x = (x_1, x_2)'$, and $u = (u_1, u_2)'$. We assume that u is independent of x, and u follows a multivariate normal distribution with mean zero and the variance-covariance matrix Σ:

$$u \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}, \quad \sigma_{12} \neq 0 \tag{3}$$

As usual, Σ is assumed to be a positive definite matrix, and the assumption $\mathrm{var}(u_2) = 1$ is made without loss of generality because d is a binary variable. If $\sigma_{12} = 0$, then there is no sample selection problem, and β_1 can be consistently estimated by OLS with using the sub data set of $d=1$.

Based on random sampling, $i = 1,2,3,\cdots,n$, a log likelihood function for the type Ⅱ Tobit mode is given by

$$\begin{aligned}&\log L(\beta,\Sigma) = \\ &\sum_{d_i=1} \log\left[1 - \Phi\left[\left(-x_{2i}\beta_2 - \sigma_{12}\sigma_1^{-1}(y_{1i} - x_{1i}\beta_1)\right)\left(1 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]\right] \\ &+ \sum_{d_i=1} \log\phi\left(\frac{y_{1i} - x_{1i}\beta_1}{\sigma_1}\right) + \sum_{d_i=1}\log(\sigma_1^{-1}) + \sum_{d_i=0}\log\left[\Phi(-x_{2i}\beta_2)\right]\end{aligned} \tag{4}$$

where $\Phi(\bullet)$ stands for the standard normal cumulative distribution function, $\phi(\bullet)$ stands for the standard normal density function, $\sum_{d_i=1}$ denotes the summation for $d_i=1$, and $\sum_{d_i=0}$ denotes the summation for $d_i=0$. Note that $(1 - \sigma_{12}^2\sigma_1^{-2})$ is positive since Σ is a positive definite matrix.

Due to the symmetry of the normal distribution, we have $1 - \Phi(-z) = \Phi(z)$ and $\Phi(-z) = 1 - \Phi(z)$. Thus, the above log likelihood function can be written as

$$\begin{aligned}&\log L(\beta,\Sigma) = \\ &\sum_{d_i=1} \log\Phi\left[\left(x_{2i}\beta_2 + \sigma_{12}\sigma_1^{-1}(y_{1i} - x_{1i}\beta_1)\right)\left(1 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right] \\ &+ \sum_{d_i=1} \log\phi\left(\frac{y_{1i} - x_{1i}\beta_1}{\sigma_1}\right) + \sum_{d_i=1}\log(\sigma_1^{-1}) + \sum_{d_i=0}\log\left[1 - \Phi(x_{2i}\beta_2)\right]\end{aligned} \tag{5}$$

This kind of log likelihood function can be found in Wooldridge (2002, p. 566), Gourieroux (2000, p.191), Maddala (1983, p.266), and Dhrymes (1986, p.1612) even though different authors use different notations.

3. Understanding the Log Likelihood Function for the Type Ⅱ Tobit Model

There are two theorems to be mentioned for understanding the likelihood function of the type Ⅱ Tobit model. The first theorem can be found in any basic textbooks in Statistics. Let $f(w_1, w_2)$ be the join density function of two random variables w_1 and w_2, $f(w_2|w_1)$ be the conditional density function of w_2 given w_1, and $f(w_1)$ be the marginal density function of w_1[2]. Then we have

$$f(w_1, w_2) = f(w_2|w_1) \times f(w_1) = f(w_1|w_2) \times f(w_2) \quad (6)$$

The Bayes' rule can be written as

$$f(w_1|w_2) = \frac{f(w_2 \mid w_1) \times f(w_1)}{f(w_2)} \quad (7)$$

The second theorem is related to the conditional normal distribution and the partition of matrix. Let w be a vector of normal random variables with the mean μ and the variance-covariance matrix Σ, w_1 be any subset of w, including a single variable, and w_2 be the remaining variables. Partition the mean vector μ and the variance-covariance matrix Σ, so that

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad (8)$$

Then, the marginal distributions are also normal. In particular,

$$w_1 \sim N(\mu_1, \Sigma_{11}) \quad \text{and} \quad w_2 \sim N(\mu_2, \Sigma_{22}) \quad (9)$$

The condition distribution of w_2 given w_1 is normal as well:

$$w_2 \mid w_1 \sim N(\mu_{2\bullet1}, \Sigma_{22\bullet1}) \quad (10)$$

where $\mu_{2\bullet1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(w_1 - \mu_1)$ and $\Sigma_{22\bullet1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

We consider the second theorem in terms of a case close to the type Ⅱ Tobit model. Let w_1 and w_2 be scalar variables, and let $\Sigma_{11} = \mathrm{var}(w_1) = \sigma_1^2$, $\Sigma_{22} = \mathrm{var}(w_2) = 1$, and $\Sigma_{12} = \Sigma_{21} = \mathrm{cov}(w_1, w_2) = \sigma_{12}$. The conditional distribution of w_2 given w_1 is

$$w_2 \mid w_1 \sim N(\mu_{2\bullet1}, \Sigma_{22\bullet1}) \quad (11)$$

where $\mu_{2\bullet1} = \mu_2 + \sigma_{12}\sigma_1^{-2}(w_1 - \mu_1)$ and $\Sigma_{22\bullet1} = 1 - \sigma_{12}^2\sigma_1^{-2} > 0$.

Since the selection equation is the probit model, we review the basic knowledge about the probit model with the following specification:

$$d = 1[x_2\beta_2 + u_2 > 0], \quad u_2 \sim N(0,1) \quad (12)$$

For $d=1$ or $x_2\beta_2 + u_2 > 0$, the probability function $f(d|x_2)$ is written as

$$f(d|x_2) = P(d = 1|x_2) = P(y_2^* > 0|x_2) = P(u_2 > -x_2\beta_2|x_2)$$
$$= \int_{-x_2\beta_2}^{\infty} \phi(u_2)du_2 = 1 - \Phi(-x_2\beta_2) = \Phi(x_2\beta_2) \quad (13)$$

For $d=0$ or $x_2\beta_2 + u_2 \le 0$, the probability function $f(d|x_2)$ is written as

$$f(d|x_2) = P(d = 0|x_2) = P(y_2^* \le 0|x_2) = P(u_2 \le -x_2\beta_2|x_2)$$
$$= \int_{-\infty}^{-x_2\beta_2} \phi(u_2)du_2 = \Phi(-x_2\beta_2) = 1 - \Phi(x_2\beta_2) \quad (14)$$

Thus, the probability function for the probit model is given by

$$f(d|x_2) = P(d|x_2) = [1 - \Phi(-x_2\beta_2)]^d [\Phi(-x_2\beta_2)]^{1-d}$$
$$= [\Phi(x_2\beta_2)]^d [1 - \Phi(x_2\beta_2)]^{1-d} \quad (15)$$

Now we go back to the type Ⅱ Tobit model and derive its likelihood function. We can use the joint density $f(y_1, d|x)$ when both

y_1 and d are observed, i.e., when $d=1$. On the other hand, we use the marginal density $f(d|x)$ when d is observed and y_1 is not observed, i.e., when $d=0$. To find the joint density $f(y_1,d|x)$ when $d=1$, we can use either

$$f(y_1,d)=f(d|y_1)\times f(y_1) \text{ or } f(y_1,d)=f(y_1|d)\times f(d)$$

We use the former case so that we derive the log likelihood function found in the books mentioned in the above. From the second theorem we know

$$y_2^*|(y_1,x)\sim N(\mu_{2\bullet 1},\Sigma_{22\bullet 1})$$

$$\mu_{2\bullet 1}=x_2\beta_2+\Sigma_{21}\Sigma_{11}^{-1}(y_1-x_1\beta_1)$$
$$=x_2\beta_2+\sigma_{12}\sigma_1^{-2}(y_1-x_1\beta_1)$$
$$\Sigma_{22\bullet 1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}=1-\sigma_{12}^2\sigma_1^{-2}>0$$

For $d=1$, the density of d given (y_1,x) is calculated as

$$P(d=1|y_1,x)=P(y_2^*>0|y_1,x)$$
$$=P\left((y_2^*-\mu_{2\bullet 1})\Sigma_{22\bullet 1}^{-\frac{1}{2}}>-\mu_{2\bullet 1}\Sigma_{22\bullet 1}^{-\frac{1}{2}}|y_1,x\right)=1-\Phi\left[-\mu_{2\bullet 1}\Sigma_{22\bullet 1}^{-\frac{1}{2}}\right]$$
$$=1-\Phi\left[\left(-x_2\beta_2-\sigma_{12}\sigma_1^{-2}(y_1-x_1\beta_1)\right)\left(1-\sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]$$

The marginal density $f(y_1|x)$ for $d=1$ is written as

$$f(y_1|x)=\frac{1}{\sigma_1}\phi\left(\frac{y_1-x_1\beta_1}{\sigma_1}\right) \tag{16}$$

Putting these pieces together yields

$$P(d=1|y_1,x)\times f(y_1|x)=\left[1-\Phi\left[\left(-x_2\beta_2-\sigma_{12}\sigma_1^{-2}(y_1-x_1\beta_1)\right)\left(1-\sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]\right]\times\phi\left(\frac{y_1-x_1\beta_1}{\sigma_1}\right)\times\frac{1}{\sigma_1} \tag{17}$$

For $d=0$, all we know is the density $f(d|x)$ and it is given by

$$P(d=0|x)=P(u_2\le -x_2\beta_2)=\Phi(-x_2\beta_2) \tag{18}$$

Putting these pieces together yields the following likelihood function.

$$L(\beta,\Sigma)=\prod_{i=1}^{n}f(y_{1i},d_i|x_i)$$
$$=\prod_{i=1}^{n}\left[P(d_i=1|y_{1i},x_i)\times f(y_{1i}|x_i)\right]^{d_i}\times\left[P(d_i=0|x_i)\right]^{1-d_i}$$
$$=\prod_{i=1}^{n}\left[\left\{1-\Phi\left[\left(-x_{2i}\beta_2-\sigma_{12}\sigma_1^{-2}(y_{1i}-x_{1i}\beta_1)\right)\left(1-\sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]\right\}\right.$$
$$\left.\times\left[\phi\left(\frac{y_{1i}-x_{1i}\beta_1}{\sigma_1}\right)\times\sigma_1^{-1}\right]^{d_i}\times\left[\Phi\left(-x_{2i}\beta_2\right)\right]^{1-d_i} \tag{19}$$

The log likelihood function is written as

$$\log L(\beta,\Sigma)=$$
$$\sum_{d_i=1}\log\left\{1-\Phi\left[\left(-x_{2i}\beta_2-\sigma_{12}\sigma_1^{-2}(y_{1i}-x_{1i}\beta_1)\right)\left(1-\sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]\right\}$$
$$+\sum_{d_i=1}\log\phi\left(\frac{y_{1i}-x_{1i}\beta_1}{\sigma_1}\right)+\sum_{d_i=1}\log(\sigma_1^{-1})+\sum_{d_i=0}\log\left[\Phi(-x_{2i}\beta_2)\right] \tag{20}$$

Let's see how the log likelihood function looks different if $f(y_1,d)=f(y_1|d)\times f(d)$ is used.

$$L(\beta,\Sigma)=\prod_{i=1}^{n}f(y_{1i},d_i|x_i)$$
$$=\prod_{i=1}^{n}\left[f(y_{1i}|d_i=1,x_i)\times P(d_i=1|x_i)\right]^{d_i}\times\left[P(d_i=0|x_i)\right]^{1-d_i}$$
$$=\prod_{i=1}^{n}\left[\phi\left[\left(y_{1i}-x_1\beta_1-\sigma_{12}(1-x_{2i}\beta_2)\right)\left(\sigma_1^2-\sigma_{12}^2\right)^{-\frac{1}{2}}\right]\right.$$
$$\left.\times\left(\sigma_1^2-\sigma_{12}^2\right)^{-\frac{1}{2}}\right]^{d_i}\times\left[1-\Phi\left(-x_{2i}\beta_2\right)\right]^{d}\times\left[\Phi\left(-x_{2i}\beta_2\right)\right]^{1-d_i} \tag{21}$$

Note that $(\sigma_1^2-\sigma_{12}^2)$ is positive since the variance-covariance matrix Σ is a positive definite matrix. The log likelihood function is written as

$$\log L(\beta,\Sigma)=\sum_{d_i=1}\log\phi\left[\left(y_{1i}-x_{1i}\beta_1-\sigma_{12}(1-x_{2i}\beta_2)\right)\left(\sigma_1^2-\sigma_{12}^2\right)^{-\frac{1}{2}}\right]$$
$$+\sum_{d_i=1}\log\left[\left(\sigma_1^2-\sigma_{12}^2\right)^{-\frac{1}{2}}\right]+\sum_{d_i=1}\log\left[1-\Phi\left(-x_{2i}\beta_2\right)\right]+\sum_{d_i=0}\log\left[\Phi(-x_{2i}\beta_2)\right] \tag{22}$$

Notice that the last two elements in the right hand side of (22) constitute the log likelihood function for the standard probit model. Using this property, the above log likelihood function can be written as

$$\log L(\beta,\Sigma)=\sum_{d_i=1}\log f(y_{1i}|d_i,x_i)+\sum_{d_i=0,1}\log f(d_i|x_i) \tag{23}$$

where $f(y_{1i}|d_i, x_i)$ is the density function of the normal distribution

$$N\left[\{x_{1i}\beta_1 + \sigma_{12}(1 - x_{2i}\beta_2)\}, (\sigma_1^2 - \sigma_{12}^2)\right] \tag{24}$$

evaluated at y_{1i}, and $f(d_i|x_i)$ is the standard probit density, and $\sum_{d_i=0,1}$ denotes the summation for both $d_i = 0$ and $d_i = 1$.

4. The Type Ⅲ Tobit Model

The type Ⅲ Tobit mode has a censured dependent variable, say y_2, rather than a binary dependent variable, for the selection equation. The type Ⅲ Tobit mode has the following standard model specification:

$$y_1 = x_1\beta_1 + u_1 \tag{25a}$$

$$y_2 = \max(0, x_2\beta_2 + u_2) \tag{25b}$$

Using a latent variable y_2^* the selection equation (25b) can be written as

$$y_2^* = x_2\beta_2 + u_2 \tag{26a}$$

$$y_2 = \begin{cases} y_2^* & \text{if} \quad y_2^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{26b}$$

As before we assume (x_2, y_2) as well as x_1 are always observed while y_1 is observed only when $y_2 > 0$. For example, wage y_1 is observed only when working hours y_2 are positive. Notice that the selection equation is the simple Tobit model. The remaining part is the same as before with $\operatorname{var}(u_2) = \sigma_2^2$. Let $y = (y_1, y_2)'$, $x = (x_1, x_2)'$, and $u = (u_1, u_2)'$. We assume that u is independent of x, and u follows a multivariate normal distribution:

$$u \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}, \quad \sigma_{12} \neq 0 \tag{27}$$

Since the selection equation is the simple Tobit model, we review the basic knowledge about the simple Tobit model with the following specification:

$$y_2 = \max(0, x_2\beta_2 + u_2), \quad u_2 \sim N(0, \sigma_2^2) \tag{28}$$

For regime switching between $y_2 > 0$ and $y_2 = 0$, it is convenient to use a binary variable d such that $d = 1$ when $y_2 > 0$ and $d = 0$ when $y_2 > 0$. For $y_2 = 0$ or equivalently $u_2 \le -x_2\beta_2$, the density $f(y_2|x_2)$ is given by

$$\begin{aligned} P(y_2 = 0|x_2) &= P(y_2^* \le 0|x_2) = P(u_2 \le -x_2\beta_2|x_2) \\ &= P\left(\frac{u_2}{\sigma_2} \le -\frac{x_2\beta_2}{\sigma_2}\Big|x_2\right) = \Phi\left(-\frac{x_2\beta_2}{\sigma_2}\right) \end{aligned} \tag{29}$$

For $y_2 > 0$, the density $f(y_2|x_2)$ is written as

$$f(y_2|x_2) = \frac{1}{\sigma_2}\phi\left(\frac{y_2 - x_2\beta_2}{\sigma_2}\right) \tag{30}$$

Putting these pieces together yields a likelihood function for the simple Tobit model.

$$\begin{aligned} L(\beta_2, \sigma_2^2) &= \prod_{i=1}^{n} f(y_{2i}|x_{2i}) = \prod_{i=1}^{n} \left[f(y_{2i}|x_{2i})\right]^{d_i} \times \left[P(y_{2i} = 0|x_{2i})\right]^{1-d_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma_2}\phi\left(\frac{y_{2i} - x_{2i}\beta_2}{\sigma_2}\right)\right]^{d_i} \times \left[\Phi\left(-\frac{x_{2i}\beta_2}{\sigma_2}\right)\right]^{1-d_i} \end{aligned} \tag{31}$$

Its log likelihood function is given by

$$\begin{aligned} \log L(\beta_2, \sigma_2^2) &= \sum_{d_i=1} \log \phi\left(\frac{y_{2i} - x_{2i}\beta_2}{\sigma_2}\right) \\ &+ \sum_{d_i=1} \log(\sigma_2^{-1}) + \sum_{d_i=0} \log\left[\Phi\left(-\frac{x_{2i}\beta_2}{\sigma_2}\right)\right] \end{aligned} \tag{32}$$

Now we go back to the type Ⅲ Tobit model and derive its likelihood function. When both y_1 and y_2 are observable, or equivalently when $y_2 > 0$, we can use the joint density $f(y_1, y_2)$. On the other hand, when y_2 is observable and y_1 is not observable, or equivalently when $y_2 = 0$, all we know is the density of y_2. To find the joint density $f(y_1, y_2)$ when $y_2 > 0$, we can use either

$$f(y_1, y_2) = f(y_2|y_1) \times f(y_1) \text{ or } f(y_1, y_2) = f(y_1|y_2) \times f(y_2)$$

First, we use the former case. We know

$f(y_1.y_2|x\ ) = f(y_2|y_1, x\ ) \times f(y_1|x\ )$

$y_1|x\ \sim N(\mu_1, \Sigma_{11})$ where $\mu_1 = x_1\beta_1$ and $\Sigma_{11} = \sigma_1^2$

$$f(y_1|x\ ) = \frac{1}{\sigma_1}\phi\left(\frac{y_1 - x_1\beta_1}{\sigma_1}\right)$$

$y_2|(y_1, x\ ) \sim N(\mu_{2\bullet1}, \Sigma_{22\bullet1})$ where

$$\mu_{2\bullet1} = x_2\beta_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - x_1\beta_1) = x_2\beta_2 + \sigma_{12}\sigma_1^{-1}(y_1 - x_1\beta_1)$$

and $\Sigma_{22\bullet1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \sigma_2^2 - \sigma_{12}^2\sigma_1^{-2} > 0$

$$\begin{aligned} f(y_2|y_1, x\ ) &= \Sigma_{22\bullet1}^{-\frac{1}{2}} \times \phi\left[\Sigma_{22\bullet1}^{-\frac{1}{2}}(y_2 - \mu_{2\bullet1})\right] \\ &= \phi\left[\left(y_2 - x_2\beta_2 - \sigma_{12}\sigma_1^{-2}(y_1 - x_1\beta_1)\right)\left(\sigma_2^2 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right] \\ &\times\left(\sigma_2^2 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}} \end{aligned}$$

For $y_2 = 0$, all we know is the marginal density of y_2. Thus we can get

$$\begin{aligned} f(y_2|x\ ) &= P(y_2^* \le 0|x\ ) = P(y_2^* - x_2\beta_2 \le -x_2\beta_2|x\ ) \\ &= P(u_2 \le -x_2\beta_2|x\ ) = P\left(\frac{u_2}{\sigma_2} \le -\frac{x_2\beta_2}{\sigma_2}|x\ \right) = \Phi\left(-\frac{x_2\beta_2}{\sigma_2}\right) \end{aligned}$$

Putting these pieces together yields the following likelihood function:

$$\begin{aligned} L(\beta,\Sigma) &= \prod_{i=}^{n} f(y_{1i}, y_{2i}|x_i) = \prod_{i=1}^{n}\left[f(y_{1i}, y_{2i}|x_i)\right]^{d_i} \times \left[P(y_{2i} = 0|x_i)\right]^{1-d_i} \\ &= \prod_{i=1}^{n}\left[f(y_{2i}|y_{1i}, x_i) \times f(y_{1i}|x_i)\right]^{d_i} \times \left[P(y_{2i}^* \le 0|x_i)\right]^{1-d_i} \\ &= \prod_{i=1}^{n}\left[\phi\left[\left(y_{2i} - x_{2i}\beta_{2i} - \sigma_{12}\sigma_1^{-2}(y_{1i} - x_{1i}\beta_1)\right)\left(\sigma_2^2 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]\right. \\ &\left.\times\left(\sigma_2^2 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right]^{d_i} \times \left[\phi\left(\frac{y_{1i} - x_{1i}\beta_1}{\sigma_1}\right) \times \sigma_1^{-1}\right]^{d_i} \times \left[\Phi\left(-\frac{x_{2i}\beta_2}{\sigma_2}\right)\right]^{1-d_i} \end{aligned} \tag{33}$$

Its log likelihood function is given by

$$\begin{aligned} &\log L(\beta,\Sigma) = \\ &\sum_{d_i=1}\log\phi\left[\left(y_{2i} - x_{2i}\beta_{2i} - \sigma_{12}\sigma_1^{-2}(y_{1i} - x_{1i}\beta_1)\right)\left(\sigma_2^2 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right] \\ &+ \sum_{d_i=1}\log\left[\left(\sigma_2^2 - \sigma_{12}^2\sigma_1^{-2}\right)^{-\frac{1}{2}}\right] + \sum_{d_i=1}\log\phi\left(\frac{y_{1i} - x_{1i}\beta_1}{\sigma_1}\right) \\ &+ \sum_{d_i=1}\log\left[\sigma_1^{-1}\right] + \sum_{d_i=0}\log\Phi\left(-\frac{x_{2i}\beta_2}{\sigma_2}\right) \end{aligned} \tag{34}$$

Let's use $f(y_1, y_2) = f(y_1|y_2) \times f(y_2)$ to see how the likelihood function looks different.

$$\begin{aligned} L(\beta,\Sigma) &= \prod_{i=1}^{n}\left[f(y_{1i}|y_{2i}, x_i) \times f(y_{2i}|x_i)\right]^{d_i} \times \left[P(y_{2i} = 0|x_i)\right]^{1-d_i} \\ &= \prod_{i=1}^{n}\left[\phi\left[\left(y_{1i} - x_{1i}\beta_1 - \sigma_{12}\sigma_2^{-2}(y_{2i} - x_{2i}\beta_2)\right)\left(\sigma_1^2 - \sigma_{12}^2\sigma_2^{-2}\right)^{-\frac{1}{2}}\right]\right. \\ &\left.\times\left(\sigma_1^2 - \sigma_{12}^2\sigma_2^{-2}\right)^{-\frac{1}{2}}\right]^{d_i} \times \left[\phi\left(\frac{y_{2i} - x_{2i}\beta_2}{\sigma_2}\right) \times \sigma_2^{-1}\right]^{d_i} \times \left[\Phi\left(-\frac{x_{2i}\beta_2}{\sigma_2}\right)\right]^{1-d_i} \end{aligned} \tag{35}$$

Its log likelihood function is given by

$$\begin{aligned} &\log L(\beta,\Sigma) = \\ &\sum_{d_i=1}\log\phi\left[\left(y_{1i} - x_{1i}\beta_1 - \sigma_{12}\sigma_2^{-2}(y_{2i} - x_{2i}\beta_2)\right)\left(\sigma_1^2 - \sigma_{12}^2\sigma_2^{-2}\right)^{-\frac{1}{2}}\right] \\ &+ \sum_{d_i=1}\left[\left(\sigma_1^2 - \sigma_{12}^2\sigma_2^{-2}\right)^{-\frac{1}{2}}\right] + \sum_{d_i=1}\log\phi\left(\frac{y_{2i} - x_{2i}\beta_2}{\sigma_2}\right) \\ &+ \sum_{d_i=1}\log\left[\sigma_2^{-1}\right] + \sum_{d_i=0}\log\left[\Phi\left(-\frac{x_{2i}\beta_2}{\sigma_2}\right)\right] \end{aligned} \tag{36}$$

Notice that the last three elements in the right hand side of (36) constitute the log likelihood function of the standard Tobit model. This property is also shown in Wooldridge (2002, p. 573). He shows the following formula of log likelihood function for observation i:

$$\log L_i(\beta,\Sigma) = d_i \times \log f(y_{1i}|y_{2i}, x_i) + \log f(y_{2i}|x_i) \tag{37}$$

where $f(y_{1i}|y_{2i}, x_i)$ is the density function of the normal distribution

$$N\left[\{x_{1i}\beta_1 + \sigma_{12}(y_{2i} - x_{2i}\beta_2)\}, (\sigma_1^2 - \sigma_{12}^2\sigma_2^{-2})\right] \tag{38}$$

evaluated at y_{1i}, $f(y_{2i}|x_i)$ is the standard censored Tobit density.

5. Models for Program Evaluation

One major use of the sample selection models is in evaluating the benefits of social programs. Keeping this in mind, we consider the following program evaluation model:

$$y_1 = x_1\beta_1 + u_1 \tag{39a}$$
$$y_2 = x_2\beta_2 + u_2 \tag{39b}$$
$$d = 1[x_3\beta_3 + u_3 > 0] \tag{39c}$$

The selection equation (39c) can be written as

$$y_3^* = x_3\beta_3 + u_3 \tag{40a}$$
$$d = \begin{cases} 1 & \text{if} \quad y_3^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{40b}$$

In the context of program evaluation, the selection equation is usually called as the participation decision function or the like. As before (x_3, d) are always observed. With regard to the main regression equation, (x_1, y_1) are observed when the i-th agent participates a social program, or equivalently when $d=1$, and (x_2, y_2) are observed when $d=0$. Thus, the observed y is defined as

$$y = \begin{cases} y_1 & \text{if} \quad d = 1 \\ y_2 & \text{if} \quad d = 0 \end{cases} \tag{41}$$

For this context it is very possible to use the same variables for x_1 and x_2. Let $y = (y_1 y_2, d)'$, $x = (x_1 x_2, x_3)'$, and $u = (u_1, u_2, u_3)'$. We assume that u is independent of x, and u follows a multivariate normal distribution:

$$u \sim N(0, \Sigma),\ \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix},\ \Sigma_{13} = \Sigma_{31} \neq 0,\ \Sigma_{23} = \Sigma_{32} \neq 0 \tag{42}$$

Here we use matrix notation such that $\Sigma_{11} = \text{var}(u_1) = \sigma_1^2$, $\Sigma_{22} = \text{var}(u_2) = \sigma_2^2$, $\Sigma_{33} = \text{var}(u_3) = \sigma_3^2 = 1$, $\Sigma_{13} = \Sigma_{31} = \text{cov}(u_1, u_3) = \sigma_{13}$, and so forth even though the elements are scalars. The use of this kind of notation gives us clear patterns for the following arguments.

For $d=1$, we can utilize the followings:

$$f(y_1 . d|x) = P(d=1|y_1, x) \times f(y_1|x) = P(y_3^* > 0|y_1, x) \times f(y_1|x)$$

$$y_1|x \sim N(\mu_1, \Sigma_{11}) \text{ where } \mu_1 = x_1\beta_1$$

$$f(y_1|x) = \Sigma_{11}^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_1 - x_1\beta_1)\right)$$

$$y_3^*|(y_1, x) \sim N(\mu_{3\bullet1}, \Sigma_{33\bullet1}) \text{ where}$$

$$\mu_{3\bullet1} = x_3\beta_3 + \Sigma_{31}\Sigma_{11}^{-1}(y_1 - x_1\beta_1) \text{ and } \Sigma_{33\bullet1} = \Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13}$$

$$P(y_3^* > 0|y_1, x) = P\left(\Sigma_{33\bullet1}^{-\frac{1}{2}}(y_3^* - \mu_{3\bullet1}) > -\Sigma_{33\bullet1}^{-\frac{1}{2}}\mu_{3\bullet1}\Big|y_1, x\right)$$
$$= 1 - \Phi\left[-\Sigma_{33\bullet1}^{-\frac{1}{2}}\mu_{3\bullet1}\right] = 1 - \Phi\Big[(\Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13})^{-\frac{1}{2}}$$
$$\left(-x_3\beta_3 - \Sigma_{31}\Sigma_{11}^{-1}(y_1 - x_1\beta_1)\right)\Big]$$

Note that $\Sigma_{11}^{-\frac{1}{2}} = (\sigma_1^2)^{-\frac{1}{2}} = \sigma_1^{-1} > 0$ and $\Sigma_{33\bullet1} = \Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13} = \sigma_3^2 - \sigma_{13}^2\sigma_1^{-2} > 0$. For $d=0$, we can utilize the followings:

$$f(y_2, d|x) = P(y_3^* < 0|y_2, x) \times f(y_2|x)$$

$$y_2|x \sim N(\mu_2, \Sigma_{22}) \text{ where } \mu_2 = x_2\beta_2$$

$$f(y_2|x) = \Sigma_{22}^{-\frac{1}{2}} \times \phi\left(\Sigma_{22}^{-\frac{1}{2}}(y_2 - x_2\beta_2)\right)$$

$$y_3^*|(y_2, x) \sim N(\mu_{3\bullet2}, \Sigma_{33\bullet2}) \text{ where}$$

$$\mu_{3\bullet2} = x_3\beta_3 + \Sigma_{32}\Sigma_{22}^{-1}(y_2 - x_2\beta_2) \text{ and}$$

$$\Sigma_{33\bullet2} = \Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23}$$

$$P(y_3^* \leq 0|y_2, x) = P\left[\Sigma_{33\bullet2}^{-\frac{1}{2}}(y_3^* - \mu_{3\bullet2}) \leq -\Sigma_{33\bullet2}^{-\frac{1}{2}}\mu_{3\bullet1}\Big|y_2, x\right]$$
$$= \Phi\left[-\Sigma_{33\bullet2}^{-\frac{1}{2}}\mu_{3\bullet1}\right]$$
$$= \Phi\left[(\Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23})^{-\frac{1}{2}}\left(-x_3\beta_3 - \Sigma_{32}\Sigma_{22}^{-1}(y_2 - x_2\beta_2)\right)\right]$$

Using these properties, it is not difficult to construct the following likelihood function for the program evaluation model:

$$L(B, \Sigma) = \prod_{i=1}^{n}\left[f(y_{1i}, d_i|x_i)\right]^{d_i} \times \left[f(y_{2i}, d_i|x_i)\right]^{1-d_i}$$
$$= \prod_{i=1}^{n}\left[P(d_i = 1|y_{1i}, x_i) \times f(y_{1i}|x_i)\right]^{d_i} \times \left[P(d_i = 0|y_{2i}, x_i) \times f(y_{2i}|x_i)\right]^{1-d_i}$$
$$= \prod_{i=1}^{n}\left[\left\{1 - \Phi\left[(\Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13})^{-\frac{1}{2}}\left(-x_{3i}\beta_3 - \Sigma_{31}\Sigma_{11}^{-1}(y_{1i} - x_{1i}\beta_1)\right)\right]\right\}\right]^{d_i}$$
$$\times\left[\Sigma_{11}^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_{1i} - x_{1i}\beta_1)\right)\right]^{d_i}$$
$$\times\left[\Phi\left[(\Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23})^{-\frac{1}{2}}\left(-x_{3i}\beta_3 - \Sigma_{32}\Sigma_{22}^{-1}(y_{2i} - x_{2i}\beta_2)\right)\right]\right]^{1-d_i}$$
$$\times\left[\Sigma_{22}^{-\frac{1}{2}} \times \phi\left(\Sigma_{22}^{-\frac{1}{2}}(y_{2i} - x_{2i}\beta_2)\right)\right]^{1-d_i} \tag{43}$$

Next, we consider the case in which the selection equation is the simple Tobit mode:

$$y_3 = \max\left(0, x_3\beta_3 + u_3\right) \tag{44a}$$

$$u_3 \sim N(0, \Sigma_{33}) \tag{44b}$$

where $\Sigma_{33} = \mathrm{var}(u_3) = \sigma_3^2$. This selection equation can be written as

$$y_3^* = x_3\beta_3 + u_3 \tag{45a}$$

$$y_3 = \begin{cases} y_3^* & \text{if} \quad y_3^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{45b}$$

For example, we can think of y_3 as the amount of loan. The observed y is defined as

$$y = \begin{cases} y_1 & \text{if} \quad y_3^* > 0 \\ y_2 & \text{otherwise} \end{cases} \tag{46}$$

For $y_3 > 0$, we can utilize the followings:

$$f(y_1, y_3|x_i) = f(y_3|y_1, x) \times f(y_1|x)$$

$$y_1|x \sim N(\mu_1, \Sigma_{11}) \text{ where } \mu_1 = x_1\beta_1$$

$$f(y_1|x) = \Sigma_{11}^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_1 - x_1\beta_1)\right)$$

$$y_3|(y_1, x) \sim N(\mu_{3\bullet1}, \Sigma_{33\bullet1}) \text{ where } \mu_{3\bullet1} = x_3\beta_3 + \Sigma_{31}\Sigma_{11}^{-1}(y_1 - x_1\beta_1)$$

$$\text{and } \Sigma_{33\bullet1} = \Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13}$$

$$f(y_3|y_1, x) = \left(\Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13}\right)^{-\frac{1}{2}}$$

$$\times \phi\left[(\Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13})^{-\frac{1}{2}}\left(y_3 - x_3\beta_3 - \Sigma_{31}\Sigma_{11}^{-1}(y_1 - x_1\beta_1)\right)\right]$$

For $y_3 > 0$, we can utilize the followings:

$$f(y_2, y_3|x) = P(y_3 = 0|y_2, x) \times f(y_2|x) = P(y_3^* \le 0|y_2, x) \times f(y_2|x)$$

$$y_2|x \sim N(\mu_2, \Sigma_{22}) \text{ where } \mu_2 = x_2\beta_2$$

$$f(y_2|x) = \Sigma_{22}^{-\frac{1}{2}} \times \phi\left(\Sigma_{22}^{-\frac{1}{2}}(y_2 - x_2\beta_2)\right)$$

$$y_3^*|(y_2, x) \sim N(\mu_{3\bullet2}, \Sigma_{33\bullet2}) \text{ where}$$

$$\mu_{3\bullet2} = x_3\beta_3 + \Sigma_{32}\Sigma_{22}^{-1}(y_2 - x_2\beta_2) \text{ and}$$

$$\Sigma_{33\bullet2} = \Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23}$$

$$P(y_3^* \le 0|y_2, x) = P\left(\Sigma_{33\bullet2}^{-\frac{1}{2}}\left(y_3^* - \mu_{3\bullet2}\right) \le -\Sigma_{33\bullet2}^{-\frac{1}{2}}\mu_{3\bullet2}\Big|y_2, x\right)$$

$$= \Phi\left[-\Sigma_{33\bullet2}^{-\frac{1}{2}}\mu_{3\bullet2}\right]$$

$$= \Phi\left[(\Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23})^{-\frac{1}{2}}\left(-x_3\beta_3 - \Sigma_{32}\Sigma_{22}^{-1}(y_2 - x_2\beta_2)\right)\right]$$

Using these properties, it is not difficult to construct the following likelihood function for the program evaluation model:

$$\begin{aligned}
L(\beta, \Sigma) &= \prod_{i=1}^{n}\left[f(y_{1i}, y_{3i}|x_i)\right]^{d_i} \times \left[f(y_{2i}, y_{3i}|x_i)\right]^{1-d_i} \\
&= \prod_{i=1}^{n}\left[f(y_{3i}|y_{1i}, x_i) \times f(y_{1i}|x_i)\right]^{d_i} \times \left[P(y_{3i} = 0|y_{2i}, x_i) \times f(y_{2i}|x_i)\right]^{1-d_i} \\
&= \prod_{i=1}^{n}\Bigg[\left(\Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13}\right)^{-\frac{1}{2}} \times \phi\Big[\left(\Sigma_{33} - \Sigma_{31}\Sigma_{11}^{-1}\Sigma_{13}\right)^{-\frac{1}{2}} \\
&\quad \left(y_{3i} - x_{3i}\beta_{3i} - \Sigma_{31}\Sigma_{11}^{-1}(y_{1i} - x_{1i}\beta_1)\right)\Big]\Bigg]^{d_i} \\
&\quad \times \left[\Sigma_{11}^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_{1i} - x_{1i}\beta_1)\right)\right]^{d_i} \\
&\quad \times \left[\Phi\left[\left(\Sigma_{33} - \Sigma_{32}\Sigma_{22}^{-1}\Sigma_{23}\right)^{-\frac{1}{2}}\left(-x_{3i}\beta_{3i} - \Sigma_{32}\Sigma_{22}^{-1}(y_{2i} - x_{2i}\beta_2)\right)\right]\right]^{1-d_i} \\
&\quad \times \left[\Sigma_{22}^{-\frac{1}{2}} \times \phi\left(\Sigma_{22}^{-\frac{1}{2}}(y_{2i} - x_{2i}\beta_2)\right)\right]^{1-d_i}
\end{aligned} \tag{47}$$

6. The General Model

The general structure of the type of model we shall consider in this section is as follows. L latent variables, $y^* = (y_1^*, y_2^*, \cdots, y_L^*)'$, are defined by a simultaneous equation system:

$$A_1 y^* - A_2 X = \varepsilon \tag{48a}$$

$$\varepsilon \sim N(0, \Sigma_\varepsilon) \tag{48b}$$

where X is the matrix of exogenous variables, and ε is a vector of L error terms. The observations on these variables are partitioned into K regimes or categories. M observable variables, $y = (y_1, y_2, \cdots, y_M)'$, are obtained for each regime as know. There is a mapping between the latent variables y^* and the observable variables y for each regime:

$$y_i = C_k y_i^* \text{ for } k = 1, 2 \cdots, K \tag{49}$$

where C_k is a mapping matrix of known constants for regime k. Notice that some latent variables may not contribute the definition of the partitions or, for that matter, they may not be observable. This general model covers a great number of cases.

For example, if $L=M=1$ and $K=2$, we recover dichotomous models such as the probit model:

$$y=\begin{cases}1 & \text{if} \quad y^*>0\\ 0 & \text{otherwise}\end{cases}$$

and the simple Tobit model:

$$y=\begin{cases}y^* & \text{if} \quad y^*>0\\ 0 & \text{otherwise}\end{cases}$$

If $L=K=2$ and $M=1$, we have the type Ⅱ Tobit mode:

$$y_1=\begin{cases}y_1^* & \text{if} \quad y_2^*>0\\ 0 & \text{otherwise}\end{cases}$$

For $L=K=2$ and $M=1$, a disequilibrium model is written as

$$y=\begin{cases}y_1^* & \text{if} \quad y_2^*>y_1^*\\ y_2^* & \text{if} \quad y_1^*\geq y_2^*\end{cases}$$

where y_1^* and y_2^* are the quantities demanded and supplied, respectively. If $L=M=K=2$, we have the type Ⅲ Tobit mode:

$$y_1=\begin{cases}y_1^* & \text{if} \quad y_2^*>0\\ 0 & \text{otherwise}\end{cases} \quad \text{and} \quad y_2=\begin{cases}y_2^* & \text{if} \quad y_2^*>0\\ 0 & \text{otherwise}\end{cases}$$

With $L=3$ and $M=K=2$, the previous program evaluation model with the probit selection equation can be written as

$$y_1=\begin{cases}y_1^* & \text{if} \quad y_3^*>0\\ y_2^* & \text{otherwise}\end{cases}$$

$$d=\begin{cases}1 & \text{if} \quad y_3^*>0\\ 0 & \text{otherwise}\end{cases}$$

The previous program evaluation model with the censored selection equation can be written as

$$y_1=\begin{cases}y_1^* & \text{if} \quad y_3^*>0\\ y_2^* & \text{otherwise}\end{cases}$$

$$y_3=\begin{cases}y_3^* & \text{if} \quad y_3^*>0\\ 0 & \text{otherwise}\end{cases}$$

Before determining the likelihood function for the general model, we must first find the distribution of the latent variables. Then we can derive the distribution of the observable endogenous variables. To find the distribution of the latent variables, we rewrite the system in the reduced form:

$$y^*=A_1^{-1}A_2X+A_1^{-1}\varepsilon=BX+u \tag{50}$$

where $B=A_1^{-1}A_2$ and $u=A_1^{-1}\varepsilon$. The distribution of u is written as

$$u\sim N(0,\Sigma) \tag{51}$$

where $\Sigma=A_1^{-1}\Sigma_\varepsilon(A_1^{-1})'$. The likelihood function for observation i in terms of the reduced-form parameters is given by

$$L_i(B,\Sigma)=(2\pi)^{-\frac{L}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(y_i^*-BX_i\right)'\Sigma^{-1}\left(y_i^*-BX_i\right)\right] \tag{52}$$

where $|\Sigma|$ is the determinant of Σ, which is positive. This can be written as

$$L_i(B,\Sigma)=|\Sigma|^{-\frac{1}{2}}\phi\left[\Sigma^{-1}\left(y_i^*-BX_i\right)\right] \tag{53}$$

where $\phi(\bullet)$ stands for the multivariate standard normal density function, and $\Sigma^{\frac{1}{2}}$ denotes the Cholesky decomposition matrix of Σ, which has the property: $(\Sigma^{\frac{1}{2}})'\Sigma^{\frac{1}{2}}=\Sigma$. Using $B=A_1^{-1}A_2$ and $\Sigma=A_1^{-1}\Sigma_\varepsilon(A_1^{-1})'$, the following elements can be rewrite as

$$\begin{aligned}&\left(y_i^*-BX_i\right)'\Sigma^{-1}\left(y_i^*-BX_i\right)\\ &=\left(y_i^*-A_1^{-1}A_2X_i\right)'\left[A_1^{-1}\Sigma_\varepsilon(A_1^{-1})'\right]^{-1}\left(y_i^*-A_1^{-1}A_2X_i\right)\\ &=\left(y_i^*-A_1^{-1}A_2X_i\right)'\left[A_1\Sigma_\varepsilon^{-1}(A_1)'\right]\left(y_i^*-A_1^{-1}A_2X_i\right)\\ &=\left(A_1y_i^*-A_2X_i\right)'\Sigma_\varepsilon^{-1}\left(A_1y_i^*-A_2X_i\right)\end{aligned} \tag{54}$$

and

$$|\Sigma| = \left|A_1^{-1}\Sigma_\varepsilon (A_1^{-1})'\right| = \left|A_1^{-1}\right|\left|\Sigma_\varepsilon\right|\left|(A_1^{-1})'\right| = \frac{|\Sigma_\varepsilon|}{|A_1|^2} \tag{55}$$

where $|A_1|$ is the absolute value of the Jacobian of the transformation from ε to y^*. Thus we can write the likelihood function for observation i as

$$\begin{aligned} &L_i(A_1, A_2, \Sigma_\varepsilon) = \left|A_1^{-1}\Sigma_\varepsilon A_1^{-1}\right|^{-\frac{1}{2}} \\ &\times \phi\left[\left(A_1^{-1}\Sigma_\varepsilon (A_1^{-1})'\right)^{-\frac{1}{2}}\left(y_i^* - A_1^{-1}A_2 X_i\right)\right] \\ &= |A_1| \times |\Sigma_\varepsilon|^{-\frac{1}{2}} \times \phi\left[\Sigma_\varepsilon^{-\frac{1}{2}}\left(A_1 y_i^* - A_2 X_i\right)\right] \end{aligned} \tag{56}$$

Based on random sampling, the log likelihood function for all observations is written as

$$\log L(A_1, A_2, \Sigma_\varepsilon^*) = \sum \log L_i(A_1, A_2, \Sigma_\varepsilon) \tag{57}$$

With regard to the log likelihood function for the structure parameters in the simultaneous equation framework, we can refer to equation (8.5.19) in Hayashi (2000, p.531), equation (12.80) in Davidson and MacKinnon (2004, p.533), the equation after (16-33) in Greene (2000, p. 694), equation (7.40) in Arellano (2003, p.138), and equation (4.24) in Hausman (1983).

Using the mapping $y_i = C_k y_i^*$ for $k = 1, 2 \cdots, K$, we can get the likelihood function in terms of the observable variables. This idea is clear, but it is cumbersome to write it down in details.

For this reason, we consider a more simple case for the system of multiple equations. Let $K=2$ and $L=M$. For $L=M$, we means that each latent variable y_j^* corresponds to a particular observable variable y_j for $j = 1, 2, \cdots, L$. We recall the reduced form:

$$y^* = BX + u, \quad u \sim N(0, \Sigma)$$

This system is partitioned as follows:

$$y^* = (y_1^*, y_2^*)', \quad X = (X_1, X_2)', \quad B = (B_1, B_2)',$$
$$u = (u_1, u_2)', \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{58}$$

The vector of the observable variables is also partitioned as $y = (y_1, y_2)'$ and y_2 is assumed to be a single variable. In particular, y_2 is assumed to be a censored variable such that

$$y_2 = \begin{cases} y_2^* & \text{if} \quad y_2^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{59}$$

The rest of the endogenous variables y_1 are assumed to be observed always for simplicity. Of course, it is possible to modify this assumption so that y_1 is observed only when $y_2 > 0$. However, such a case is already considered in the section of the type Ⅲ Tobit mode, and it is not interesting to do the same kind of excise again.

When both y_1 and y_2 are observed, i.e., when $y_2 > 0$, the joint density function $f(y_1, y_2)$ for observation i is given by (53) with $y_i^* = y_i$. For $y_2 > 0$ we can also use

$$f(y_1^*, y_2^*) = f(y_1, y_2) = f(y_2 | y_1) \times f(y_1) \tag{60}$$

For this case we know

$$y_1 | X \sim N(\mu_1, \Sigma_{11}) \text{ where } \mu_1 = B_1 X_1$$
$$f(y_1 | X) = |\Sigma_{11}|^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_1 - B_1 X_1)\right)$$
$$y_2 | (y_1, X) \sim N(\mu_{2\bullet 1}, \Sigma_{22\bullet 1}) \text{ where}$$
$$\mu_{2\bullet 1} = B_2 X_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - B_1 X_1) \text{ and}$$
$$\Sigma_{22\bullet 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$
$$\begin{aligned} f(y_2 | y_1, X) &= |\Sigma_{22\bullet 1}|^{-\frac{1}{2}} \times \phi\left[\Sigma_{22\bullet 1}^{-\frac{1}{2}}(y_2 - \mu_{2\bullet 1})\right] \\ &= \left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right|^{-\frac{1}{2}} \times \phi\left[\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-\frac{1}{2}}\right. \\ &\left.\left(y_2 - B_2 X_2 - \Sigma_{21}\Sigma_{11}^{-1}(y_1 - B_1 X_1)\right)\right] \end{aligned}$$

For $y_2 = 0$ or equivalently $y_2^* \le 0$, we can use

$$f(y_1^*, y_2^* | X\ ) = P(y_2^* \le 0 | y_1, X_i) \times f(y_1 | X_i) \quad (61)$$

where $y_1^* = y_1$. For this case we know

$$y_1 | X \sim N(\mu_1, \Sigma_{11}) \text{ where } \mu_1 = B_1 X_1$$

$$f(y_1 | X\ ) = |\Sigma_{11}|^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_1 - B_1 X_1)\right)$$

$$y_2^* | (y_1, X\ ) \sim N(\mu_{2\bullet 1}, \Sigma_{22\bullet 1}) \text{ where}$$

$$\mu_{2\bullet 1} = B_2 X_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - B_1 X_1) \text{ and}$$

$$\Sigma_{22\bullet 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

$$P(y_2^* \le 0 | y_1, X) = P\left(\Sigma_{22\bullet 1}^{-\frac{1}{2}}(y_2^* - \mu_{2\bullet 1}) \le -\Sigma_{22\bullet 1}^{-\frac{1}{2}}\mu_{2\bullet 1} \middle| y_1, X\right)$$

$$= \Phi\left[-\Sigma_{22\bullet 1}^{-\frac{1}{2}}\mu_{2\bullet 1}\right]$$

$$= \Phi\left[(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-\frac{1}{2}}(-B_2 X_2 - \Sigma_{21}\Sigma_{11}^{-1}(y_1 - B_1 X_1))\right]$$

Putting these pieces together yields the following likelihood function for the above system of multiple equations:

$$L(B,\Sigma) = \prod_{i=1}^{n}\left[f(y_{1i}, y_{2i} | X_i)\right]^{d_i} \times \left[f(y_{1i}, y_{2i} | X_i)\right]^{1-d_i}$$

$$= \prod_{i=1}^{n}\left[f(y_{2i} | y_{1i}, X_i) \times f(y_{1i} | X_i)\right]^{d_i}$$

$$\times \left[P(y_{2i} = 0 | y_{1i}, X_i) \times f(y_{1i} | X_i)\right]^{1-d_i}$$

$$= \prod_{i=1}^{n}\left[\left|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right|^{-\frac{1}{2}} \times \phi\left[\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-\frac{1}{2}}\right.\right.$$

$$\left.\left.\left(y_{2i} - B_2 X_{2i} - \Sigma_{21}\Sigma_{11}^{-1}(y_{1i} - B_1 X_{1i})\right)\right]\right]^{d_i}$$

$$\times \left[|\Sigma_{11}|^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_{1i} - B_1 X_{1i})\right)\right]^{d_i}$$

$$\times \left[\Phi\left[\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-\frac{1}{2}}\left(-B_2 X_{2i} - \Sigma_{21}\Sigma_{11}^{-1}(y_{1i} - B_1 X_{1i})\right)\right]\right]^{1-d_i}$$

$$\times \left[|\Sigma_{11}|^{-\frac{1}{2}} \times \phi\left(\Sigma_{11}^{-\frac{1}{2}}(y_{1i} - B_1 X_{1i})\right)\right]^{1-d_i} \quad (62)$$

7. Conclusions

In order to understand likelihood functions for sample selection models such as the type Ⅱ and Ⅲ Tobit modes, we show that two theorems are fundamentally important. Once we understand how to apply these theorems in the context of likelihood functions for sample selection models, we can derive likelihood functions easily with some variations. With such knowledge, we can modify the likelihood function for a standard specification easily based on our needs. Such knowledge is also applied to more complicated models including simultaneous equation systems and the like.

(Received : May 29, 2009, Accepted : June 24, 2009)

References

Amemiya, Takeshi (1985) *Advance Econometrics*, Harvard University Press.

Arellano, Manuel (2003) *Panel Data Econometrics* (Advanced Texts in Econometrics No.11), Oxford University Press.

Davidson, Russell and James G. MacKinnon (2004) *Econometric Theory and Methods*. Oxford University Press.

Dhrymes, Phoebus J. (1986) "Limited Dependent Variables," in Griliches, Z. and M. D. Intriligator (eds), *Handbook of Econometrics*, Vol. 3, Elsevier Science Publishers, pp. 1567-1631.

Gourieroux, Christian (2000) *Econometrics of Qualitative Dependent Variables*, Cambridge University Press.

Greene, William H. (2000) *Econometric Analysis* (Forth Edition), Prentice Hall International.

Hausman, Jerry A. (1983) "Specification and Estimation of Simultaneous Equation Models," in Griliches, Z. and M. D. Intriligator (eds), *Handbook of Econometrics*, Vol. 1, Elsevier Science Publishers, pp. 392-448.

Hayashi, Fumio (2000) *Econometrics*, Princeton: Princeton University Press.

Krishnakumar, Jayalakshmi (1988) *Estimation of Simultaneous Equation Models with Error Components Structure*, Berlin: Springer-Verlag.

Maddala, G. S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs.

Maddala, G. S.(1986) "Disequilibrium, Self-Selection, and Switching Models," in Griliches, Z. and M. D. Intriligator (eds), *Handbook of Econometrics*, Vol. 3, Elsevier Science Publishers, pp.1633-1688.

Wooldridge, Jeffrey M. (2002) *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Notes

1) We define x as a row vector and β as a column vector so that $x\beta$ is the inner product of x and β .

2) We may use rigorous notation such as $f_{12}(w_1, w_2)$, $f_{2\bullet 1}(w_2|w_1)$, $f_1(w_1)$, and $f_2(w_2)$. However, this kind of notation becomes cumbersome in the course of our arguments. For notational simplicity, $f_{12}(w_1, w_2)$, $f_{2\bullet 1}(w_2|w_1)$, $f_1(w_1)$, and $f_2(w_2)$ are written as $f(w_1, w_2)$, $f(w_2|w_1)$, $f(w_1)$, and $f(w_2)$, respectively.