Chapter 10

Quasi-Maximum Likelihood: Applications

10.1 Binary Choice Models

In many economic applications the dependent variables of interest may assume only finitely many integer values, each labeling a category of data. The simplest case of discrete dependent variables is the binary variable that takes on the values one and zero. For example, the ownership of a durable goods and the choice of participating a particular event may be represented by a binary dependent variable. Such variables are different from standard "continuous" variables such as GDP and income. It is then natural to take into account the data characteristics in the specification of quasi-likelihood function.

Conditional on the explanatory variables $\boldsymbol{x}_t$, the binary variable y_t is such that

$$y_t = \begin{cases} 1, & \text{with probability } \mathbb{P}(y_t = 1|\boldsymbol{x}_t), \\ 0, & \text{with probability } 1 - \mathbb{P}(y_t = 1|\boldsymbol{x}_t). \end{cases}$$

The density function of y_t given $\boldsymbol{x}_t$ is of the Bernoulli type:

$$g(y_t|\boldsymbol{x}_t) = \mathbb{P}(y_t = 1|\boldsymbol{x}_t)^{y_t}[1 - \mathbb{P}(y_t = 1|\boldsymbol{x}_t)]^{1-y_t}.$$

A standard modeling approach is to find a function F such that $F(\boldsymbol{x}_t; \boldsymbol{\theta})$ approximates the conditional probability $\mathbb{P}(y_t = 1|\boldsymbol{x}_t)$. The quasi-likelihood function is then:

$$f(y_t|\boldsymbol{x}_t; \boldsymbol{\theta}) = F(\boldsymbol{x}_t; \boldsymbol{\theta})^{y_t}[1 - F(\boldsymbol{x}_t; \boldsymbol{\theta})]^{1-y_t}.$$

The QMLE $\tilde{\boldsymbol{\theta}}_T$ is then obtained by maximizing

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\big[y_t \log F(\boldsymbol{x}_t; \boldsymbol{\theta}) + (1 - y_t)\log(1 - F(\boldsymbol{x}_t; \boldsymbol{\theta}))\big].$$

As $0 \leq \mathbb{P}(y_t = 1|\boldsymbol{x}_t) \leq 1$, it is natural to choose F as a function bounded between zero and one, such as a distribution functions. When $F(\boldsymbol{x}_t;\boldsymbol{\theta}) = \Phi(\boldsymbol{x}_t'\boldsymbol{\theta})$ with Φ the standard normal distribution function:

$$\Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\,\mathrm{d}v,$$

we have the *probit* model. For $\Phi(u) = p$, its inverse $\Phi^{-1}(p)$ is known as the *probit transformation*. When $F(\boldsymbol{x}_t;\boldsymbol{\theta}) = G(\boldsymbol{x}_t'\boldsymbol{\theta})$ with G the logistic distribution function:

$$G(u) = \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u}.$$

it is the *logit* model. The logistic distribution has mean zero and variance $\pi^2/3$, and it is more peaked around its mean and has slightly thicker tails than the standard normal distribution. For $G(u) = p$, its inverse

$$G^{-1}(p) = \log\Big(\frac{p}{1-p}\Big),$$

is known as the *logit transformation*. It is easy to verify that $G'(u) = G(u)[1-G(u)]$ which is convenient for estimating a logit model.

For the logit model,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}\left[y_t \frac{G'(\boldsymbol{x}_t'\boldsymbol{\theta})}{G(\boldsymbol{x}_t'\boldsymbol{\theta})} - (1-y_t)\frac{G'(\boldsymbol{x}_t'\boldsymbol{\theta})}{1-G(\boldsymbol{x}_t'\boldsymbol{\theta})}\right]\boldsymbol{x}_t \\ &= \frac{1}{T}\sum_{t=1}^{T}[y_t - G(\boldsymbol{x}_t'\boldsymbol{\theta})]\boldsymbol{x}_t,\end{aligned}$$

and

$$\nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} G(\boldsymbol{x}_t'\boldsymbol{\theta})[1-G(\boldsymbol{x}_t'\boldsymbol{\theta})]\boldsymbol{x}_t\boldsymbol{x}_t'.$$

Given that $G(\boldsymbol{x}_t'\boldsymbol{\theta})[1-G(\boldsymbol{x}_t'\boldsymbol{\theta})] > 0$ for all $\boldsymbol{\theta}$, this matrix is negative definite, so that the quasi-log-likelihood function is globally concave on the parameter space. When $\boldsymbol{x}_t$ contains a constant term, the first order condition implies

$$\frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T} G(\boldsymbol{x}_t'\tilde{\boldsymbol{\theta}}_T);$$

that is, the average of fitted values is the relative frequency of $y_t = 1$.

For the probit model,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}\left[y_t \frac{\phi(\boldsymbol{x}_t'\boldsymbol{\theta})}{\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})} - (1-y_t)\frac{\phi(\boldsymbol{x}_t'\boldsymbol{\theta})}{1-\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})}\right]\boldsymbol{x}_t \\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{y_t - \Phi(\boldsymbol{x}_t'\boldsymbol{\theta})}{\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})[1-\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})]}\phi(\boldsymbol{x}_t'\boldsymbol{\theta})\boldsymbol{x}_t,\end{aligned}$$

where ϕ is the standard normal density function. It can also be verified that

$$\nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}\left[y_t \frac{\phi(\boldsymbol{x}_t'\boldsymbol{\theta}) + \boldsymbol{x}_t'\boldsymbol{\theta}\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})}{\Phi^2(\boldsymbol{x}_t'\boldsymbol{\theta})} + (1-y_t)\,\frac{\phi(\boldsymbol{x}_t'\boldsymbol{\theta}) - \boldsymbol{x}_t'\boldsymbol{\theta}[1-\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})]}{[1-\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})]^2}\right]\phi(\boldsymbol{x}_t'\boldsymbol{\theta})\boldsymbol{x}_t\boldsymbol{x}_t',$$

which is also negative definite; see e.g., Amemiya (1985, pp. 273–274).

Clearly, the conditional mean of y_t is just the conditional probability $\mathbb{P}(y_t = 1 \mid \boldsymbol{x}_t)$, and its conditional variance is

$$\operatorname{var}(y_t|\boldsymbol{x}_t) = \mathbb{P}(y_t = 1|\boldsymbol{x}_t)[1 - \mathbb{P}(y_t = 1|\boldsymbol{x}_t)],$$

which changes with $\boldsymbol{x}_t$. Thus, $F(\boldsymbol{x}_t;\boldsymbol{\theta})$ is also an approximation to the conditional mean function; $F(\boldsymbol{x}_t;\boldsymbol{\theta})[1 - F(\boldsymbol{x}_t;\boldsymbol{\theta})]$ is an approximation to the conditional variance function. Writing

$$y_t = F(\boldsymbol{x}_t;\boldsymbol{\theta}) + e_t,$$

we can see that the probit and logit models are in effect different nonlinear mean specifications with conditional heteroskedasticity. Although $\boldsymbol{\theta}$ may be estimated using the NLS method, the resulting estimator cannot be efficient because it ignores conditional heteroskedasticity. Even a weighted NLS estimator that takes into account the conditional variance is still inefficient because it does not consider the Bernoulli feature of y_t. Note that the linear probability model in which $F(\boldsymbol{x}_t;\boldsymbol{\theta}) = \boldsymbol{x}_t'\boldsymbol{\theta}$ (Section 4.4) is not appropriate because the fitted values may be outside the range of $[0, 1]$.

It should be noted that the marginal response of the choice probability to the change of a particular variable x_{tj} in a probit model is

$$\frac{\partial\Phi(\boldsymbol{x}_t'\boldsymbol{\theta})}{\partial x_{tj}} = \phi(\boldsymbol{x}_t'\boldsymbol{\theta})\theta_j,$$

and the marginal response in a logit model is

$$\frac{\partial G(\boldsymbol{x}_t'\boldsymbol{\theta})}{\partial x_{tj}} = G(\boldsymbol{x}_t'\boldsymbol{\theta})[1 - G(\boldsymbol{x}_t'\boldsymbol{\theta})]\theta_j,$$

which change with $\boldsymbol{x}_t$. By contrast, the marginal response to the change of a regressor in a linear regression model is the associated coefficient which is invariant with respect to t. Thus, it is typical to evaluate the marginal response based on a particular value of $\boldsymbol{x}_t$, such as $\boldsymbol{x}_t = \mathbf{0}$ or $\boldsymbol{x}_t = \bar{\boldsymbol{x}}$, the sample average of $\boldsymbol{x}_t$. Note that when $\boldsymbol{x}_t = \mathbf{0}$, $\varphi(0) \approx 0.4$ and $G(0)[1 - G(0)] = 0.25$. This suggests that the QMLE for the logit model is approximately 1.6 times the QMLE for the probit model when $\boldsymbol{x}_t$ are close to zero.

An alternative view of the binary choice model is to assume that the observed variable y_t is determined by the latent (index) variable y_t^*:

$$y_t = \begin{cases} 1, & y_t^* > 0, \\ 0, & y_t^* \leq 0, \end{cases}$$

where $y_t^* = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$. Thus,

$$\mathbb{P}(y_t = 1|\boldsymbol{x}_t) = \mathbb{P}(y_t^* > 0|\boldsymbol{x}_t) = \mathbb{P}(e_t > -\boldsymbol{x}_t\boldsymbol{\beta}|\boldsymbol{x}_t).$$

The probability on the right-hand side is also $\mathbb{P}(e_t < \boldsymbol{x}_t\boldsymbol{\beta}|\boldsymbol{x}_t)$ provided that e_t is symmetric about zero. The probit specification Φ or the logit specification G can then be viewed as specifications of the conditional distribution of e_t. This suggests that a possible modification of these two models is to consider an asymmetric distribution function for e_t.

10.2 Models for Multiple Choices

A more general discrete choice model would be needed when one faces multiple choices of job, event or transportation alternatives. In this case, the dependent variable y_t takes on $J+1$ integer values $(0, 1, \ldots, J)$ which correspond to different categories that are not overlapping and do not have a natural ordering. We have

$$y_t = \begin{cases} 0, & \text{with probability } \mathbb{P}(y_t = 0|\boldsymbol{x}_t), \\ 1, & \text{with probability } \mathbb{P}(y_t = 1|\boldsymbol{x}_t), \\ \vdots & \\ J, & \text{with probability } \mathbb{P}(y_t = J|\boldsymbol{x}_t). \end{cases}$$

The number of choices J may depend on t because, for example, an individual who does not own a car can not have the choice of driving to work. We shall not consider this complication here, however.

10.2.1 Multinomial Logit Models

Define the new binary variable $d_{t,j}$ for $j = 0, 1, \ldots, J$ as

$$d_{t,j} = \begin{cases} 1, & \text{if } y_t = j, \\ 0, & \text{otherwise .} \end{cases}$$

Note that $\sum_{j=0}^{J} d_{t,j} = 1$. The density function of $d_{t,0}, \ldots, d_{t,J}$ given $\boldsymbol{x}_t$ is then

$$g(d_{t,0}, \ldots, d_{t,J}|\boldsymbol{x}_t) = \prod_{j=0}^{J} \mathbb{P}(y_t = j|\boldsymbol{x}_t)^{d_{t,j}}.$$

The conditional probabilities can be approximated by some functions $F(\boldsymbol{x}_t;\boldsymbol{\theta}_j)$, where the regressors are t-specific (individual specific) but not choice specific, yet their responses to the choices j may be different. Note that only J conditional probabilities need to be specified; otherwise, there would be an identification problem because the sum of $J+1$ probabilities must be one.

Common specifications of the conditional probabilities are

$$F(\boldsymbol{x}_t;\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J) = G_{t,0} = \frac{1}{1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_t'\boldsymbol{\theta}_k)},$$

$$F(\boldsymbol{x}_t;\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J) = G_{t,j} = \frac{\exp(\boldsymbol{x}_t'\boldsymbol{\theta}_j)}{1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_t'\boldsymbol{\theta}_k)},$$

which lead to the so-called *multinomial logit* model. The quasi-log-likelihood function of this model reads

$$L_T(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} d_{t,j}\boldsymbol{x}_t'\boldsymbol{\theta}_j - \frac{1}{T}\sum_{t=1}^{T}\log\left(1+\sum_{k=1}^{J}\exp(\boldsymbol{x}_t'\boldsymbol{\theta}_k)\right).$$

The gradient vector with respect to $\boldsymbol{\theta}_j$ is

$$\nabla_{\boldsymbol{\theta}_j} L_T(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J) = \frac{1}{T}\sum_{t=1}^{T}(d_{t,j}-G_{t,j})\boldsymbol{x}_t, \qquad j=1,\ldots,J.$$

Setting these equations to zero we can solve for the QMLE of $\boldsymbol{\theta}_j$. The first order condition again implies that, when $\boldsymbol{x}_t$ contains a constant term, the predicted relative frequency of the j^{th} category equals its actual relative frequency.

It can also be verified that the $(i,j)^{\text{th}}$ off-diagonal block of the Hessian matrix is

$$\nabla_{\boldsymbol{\theta}_j\boldsymbol{\theta}_i'} L_T(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J) = \frac{1}{T}\sum_{t=1}^{T}(G_{t,j}G_{t,i})\boldsymbol{x}_t\boldsymbol{x}_t', \qquad i\neq j,\ i,j=1,\ldots,J,$$

and the j^{th} diagonal block is

$$\nabla_{\boldsymbol{\theta}_j\boldsymbol{\theta}_j'} L_T(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J) = -\frac{1}{T}\sum_{t=1}^{T} G_{t,j}(1-G_{t,j})\boldsymbol{x}_t\boldsymbol{x}_t', \qquad j=1,\ldots,J.$$

We may then use these information to compute the asymptotic covariance matrix for the QMLEs.

As in the binomial logit model, one should be careful about the the marginal response of the choice probability, $G_{t,j}$, to the change of the variables $\boldsymbol{x}_t$. It is easy to verify that

$$\nabla_{\boldsymbol{x}_t} G_{t,0} = -G_{t,0}\sum_{i=1}^{J} G_{t,i}\boldsymbol{\theta}_i,$$

$$\nabla_{\boldsymbol{x}_t} G_{t,j} = G_{t,j}\left(\boldsymbol{\theta}_j - \sum_{i=1}^{J} G_{t,i}\boldsymbol{\theta}_i\right), \qquad j=1,\ldots,J.$$

Thus, when $\boldsymbol{x}_t$ changes, all coefficient vectors $\boldsymbol{\theta}_i$, $i = 1, \ldots, J$, enter the marginal response of $G_{t,j}$.

10.2.2 Conditional Logit Model

A model closely related to the multinomial logit model is McFadden's *conditional logit model*, in which the regressors are choice-specific. For example, when an individual makes choice among several commuting modes, the regressors may include the in-vehicle time and waiting time that vary with the vehicle he/she chooses. As such, we shall consider $\boldsymbol{x}_t = (\boldsymbol{x}_{t,0}' \ \boldsymbol{x}_{t,1}' \ \ldots \ \boldsymbol{x}_{t,J}')'$, where $\boldsymbol{x}_{t,j}$ is the vector of regressors characterizing the t^{th} individual's attributes with respect to the j^{th} choice.

As in Section 10.2.1, we define the binary variable $d_{t,j}$ for $j = 0, 1, \ldots, J$ as

$$d_{t,j} = \begin{cases} 1, & \text{if } y_t = j, \\ 0, & \text{otherwise}, \end{cases}$$

and the density function of $d_{t,0}, \ldots, d_{t,J}$ given $\boldsymbol{x}_t$ is

$$g(d_{t,0}, \ldots, d_{t,J} | \boldsymbol{x}_t) = \prod_{j=0}^{J} \mathbb{P}(y_t = j | \boldsymbol{x}_t)^{d_{t,j}}.$$

In the conditional logit model, the conditional probabilities are approximated by

$$G_{t,j}^{\dagger} = \frac{\exp(\boldsymbol{x}_{t,j}'\boldsymbol{\theta})}{\sum_{k=0}^{J} \exp(\boldsymbol{x}_{t,k}'\boldsymbol{\theta})}, \qquad j = 0, 1, \ldots, J.$$

Note that $\boldsymbol{\theta}$ is now common for all j, so that we may have $J + 1$ specifications for the conditional probabilities without causing an identification problem. This model would be convenient if one wants to predict the probability of a new choice.

Similar to the preceding subsection, the quasi-log-likelihood function is

$$L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=0}^{J} d_{t,j} \boldsymbol{x}_{t,j}'\boldsymbol{\theta} - \frac{1}{T} \sum_{t=1}^{T} \log \left(\sum_{k=0}^{J} \exp(\boldsymbol{x}_{t,k}'\boldsymbol{\theta}) \right).$$

The first order condition is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=0}^{J} (d_{t,j} - G_{t,j}^{\dagger}) \boldsymbol{x}_{t,j} = \mathbf{0},$$

from which the QMLE for $\boldsymbol{\theta}$ can be easily calculated. The Hessian matrix of the quasi-log-likelihood function is

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}^2 L_T(\boldsymbol{\theta}) &= -\frac{1}{T} \sum_{t=1}^{T} \left[\sum_{j=0}^{J} G_{t,j}^{\dagger} \boldsymbol{x}_{t,j} \boldsymbol{x}_{t,j}' - \left(\sum_{j=0}^{J} G_{t,j}^{\dagger} \boldsymbol{x}_{t,j} \right) \left(\sum_{j=0}^{J} G_{t,j}^{\dagger} \boldsymbol{x}_{t,j}' \right) \right] \\ &= -\frac{1}{T} \sum_{t=1}^{T} \sum_{j=0}^{J} G_{t,j}^{\dagger} (\boldsymbol{x}_{t,j} - \bar{\boldsymbol{x}}_t)(\boldsymbol{x}_{t,j} - \bar{\boldsymbol{x}}_t)', \end{aligned}$$

where $\bar{\boldsymbol{x}}_t = \sum_{j=0}^{J} G_{t,j}^{\dagger} \boldsymbol{x}_{t,j}$ is the weighted average of $\boldsymbol{x}_{t,j}$. It is then clear that the Hessian matrix is negative definite so that the quasi-log-likelihood function is globally concave. The marginal response of $G_{t,j}^{\dagger}$ to $\boldsymbol{x}_{t,i}$, is

$$\nabla_{\boldsymbol{x}_{t,i}} G_{t,j}^{\dagger} = -G_{t,j}^{\dagger} G_{t,i}^{\dagger} \boldsymbol{\theta}, \qquad i \neq j,\ i = 0, \ldots, J,$$

and the marginal response of $G_{t,j}^{\dagger}$ to $\boldsymbol{x}_{t,j}$ is

$$\nabla_{\boldsymbol{x}_{t,j}} G_{t,j}^{\dagger} = G_{t,j}^{\dagger}(1 - G_{t,j}^{\dagger}) \boldsymbol{\theta}, \qquad j = 0, \ldots, J.$$

This shows that each choice probability $G_{t,j}^{\dagger}$ is affected not only by $\boldsymbol{x}_{t,j}$ but also the regressors for other choices, $\boldsymbol{x}_{t,i}$.

The conditional logit model may be understood under a random utility framework (McFadden, 1974). Given the random utility of the choice j:

$$U_{t,j} = \boldsymbol{x}_{t,j}' \boldsymbol{\theta} + \varepsilon_{t,j}, \qquad j = 0, 1, \ldots, J,$$

the alternative i would be chosen if

$$\mathbb{P}(y_t = i | \boldsymbol{x}_t) = \mathbb{P}(U_{t,i} > U_{t,j}, \text{ for all } j \neq i | \boldsymbol{x}_t).$$

Suppose that $\varepsilon_{t,j}$ are independent random variables across j such that they have the *type I extreme value distribution*: $\exp[-\exp(-\varepsilon_{t,j})]$. To see how this setup leads to a conditional logit model, consider the simple case that there are 3 choices ($J = 2$). Letting $\mu_{t,j} = \boldsymbol{x}_{t,j}' \boldsymbol{\theta}$ we have

$$\begin{aligned}
\mathbb{P}(y_t = 2 | \boldsymbol{x}_t) &= \mathbb{P}(\varepsilon_{t,2} + \mu_{t,2} - \mu_{t,1} > \varepsilon_{t,1} \text{ and } \varepsilon_{t,2} + \mu_{t,2} - \mu_{t,0} > \varepsilon_{t,0}) \\
&= \int_{-\infty}^{\infty} f(\varepsilon_{t,2}) \left(\int_{-\infty}^{\varepsilon_{t,2} + \mu_{t,2} - \mu_{t,1}} f(\varepsilon_{t,1}) \, \mathrm{d}\varepsilon_{t,1} \right) \\
&\qquad \left(\int_{-\infty}^{\varepsilon_{t,2} + \mu_{t,2} - \mu_{t,0}} f(\varepsilon_{t,0}) \, \mathrm{d}\varepsilon_{t,0} \right) \mathrm{d}\varepsilon_{t,2},
\end{aligned}$$

where $f(\varepsilon_{t,j})$ is the type I extreme value density function: $\exp(-\varepsilon_{t,j}) \exp[-\exp(-\varepsilon_{t,j})]$. Then,

$$\begin{aligned}
\mathbb{P}(y_t = 2 | \boldsymbol{x}_t) &= \int_{-\infty}^{\infty} \exp(-\varepsilon_{t,2}) \exp[-\exp(-\varepsilon_{t,2})] \exp[-\exp(-\varepsilon_{t,2} - \mu_{t,2} + \mu_{t,1})] \\
&\qquad \times \exp[-\exp(-\varepsilon_{t,2} - \mu_{t,2} + \mu_{t,0})] \, \mathrm{d}\varepsilon_{t,2} \\
&= \frac{\exp(\mu_{t,2})}{\exp(\mu_{t,0}) + \exp(\mu_{t,1}) + \exp(\mu_{t,2})}.
\end{aligned}$$

This is precisely the conditional logit specification of $\mathbb{P}(y_t = 2 | \boldsymbol{x}_t)$. The specifications for other conditional probabilities can be obtained similarly. Thus, the conditional logit

model hinges on the condition that the choices must be quite different such that they are independent of each other. This is also known as the condition of *independence of irrelevant alternatives* (IIA). The IIA condition may not hold in applications and therefore must be tested empirically.

10.2.3 Ordered Data

In some applications, the multiple categories of interest may have a natural ordering, such as the levels of education, opinion, and price. In particular, y_t is said to be *ordered* if

$$y_t = \begin{cases} 0, & y_t^* \le c_1 \\ 1, & c_1 < y_t^* \le c_2 \\ \vdots & \\ J, & c_J < y_t^*; \end{cases}$$

otherwise, it is *unordered*. The variable y_t in Section 10.2.1 is unordered.

Setting $y_t^* = \boldsymbol{x}_t'\boldsymbol{\theta} + e_t$, the conditional probabilities are:

$$\begin{aligned} \mathbb{P}(y_t = j|\boldsymbol{x}_t) &= \mathbb{P}(c_j < y_t^* < c_{j+1} \mid \boldsymbol{x}_t) \\ &= F(c_{j+1} - \boldsymbol{x}_t'\boldsymbol{\theta}) - F(c_j - \boldsymbol{x}_t'\boldsymbol{\theta}), \qquad j = 0, 1, \ldots, J, \end{aligned}$$

where $c_0 = -\infty$, $c_{J+1} = \infty$, and F is the distribution of e_t. When F is the logistic distribution function G, this is the *ordered logit model*; when F is the standard normal distribution function Φ, it is the *ordered probit model*. The quasi-log-likelihood function of the ordered model is

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=0}^{J} \log[F(c_{j+1} - \boldsymbol{x}_t'\boldsymbol{\theta}) - F(c_j - \boldsymbol{x}_t'\boldsymbol{\theta})],$$

from which we can easily compute its gradient and Hessian matrix. Pratt (1981) showed that the Hessian matrix is negative definite so that the quasi-log-likelihood function is globally concave. The marginal responses of the choice probabilities to the change of the variables can be easily calculated. For the ordered probit model, these responses are

$$\nabla_{\boldsymbol{x}_t}\, \mathbb{P}(y_t = j|\boldsymbol{x}_t) = \begin{cases} -\phi(c_1 - \boldsymbol{x}_t'\boldsymbol{\theta})\boldsymbol{\theta}, & j = 0 \\ \big[\phi(c_j - \boldsymbol{x}_t'\boldsymbol{\theta}) - \phi(c_{j+1} - \boldsymbol{x}_t'\boldsymbol{\theta})\big]\boldsymbol{\theta}, & j = 1, \ldots, J-1, \\ \phi(c_J - \boldsymbol{x}_t'\boldsymbol{\theta})\boldsymbol{\theta}, & j = J. \end{cases}$$

See Exercise 10.5 for the marginal responses of the ordered logit model.

10.3 Models of Count Data

In practice, there are integer-valued dependent variables that are not categorical, such as the number of accidents of a plant or the number of patents filed by a company. Such data, also known as *count data*, differ from categorical data in that they represent the intensity of the occurrence of an event.

It is typical to postulate a Poisson distribution for the conditional probability of the count data $y_t = 0, 1, 2, \ldots$:

$$\mathbb{P}(y_t|\boldsymbol{x}_t) = \exp(-\lambda_t)\frac{\lambda_t^{y_t}}{y_t!},$$

where the parameter λ_t, which is also the conditional mean, has a log-linear form:

$$\log \lambda_t = \boldsymbol{x}_t'\boldsymbol{\theta}.$$

The quasi-log-likelihood function is then

$$\begin{aligned} L_T(\boldsymbol{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}\big[-\lambda_t + y_t \log(\lambda_t) - \log(y_t!)\big] \\ &= \frac{1}{T}\sum_{t=1}^{T}\big[-\exp(\boldsymbol{x}_t'\boldsymbol{\theta}) + y_t(\boldsymbol{x}_t'\boldsymbol{\theta}) - \log(y_t!)\big]. \end{aligned}$$

The gradient vector is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}[y_t - \exp(\boldsymbol{x}_t'\boldsymbol{\theta})]\boldsymbol{x}_t,$$

and the Hessian matrix is

$$\nabla^2_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}\exp(\boldsymbol{x}_t'\boldsymbol{\theta})\boldsymbol{x}_t\boldsymbol{x}_t'.$$

Since $\exp(\boldsymbol{x}_t'\boldsymbol{\theta})$ are non-negative, the Hessian matrix is also negative definite on the parameter space.

A drawback of the Poisson specification is that it in effect restricts the conditional variance to be the same as the conditional mean λ_t. Writing

$$y_t = \exp(\boldsymbol{x}_t'\boldsymbol{\theta}) + e_t,$$

we simply have a nonlinear specification of the conditional mean function without restricting the conditional variance. Such a specification does not consider the feature that y_t are discrete-valued count data, however.

10.4 Models of Limited Dependent Variables

There are also numerous economic applications in which a dependent variable can only be observed within a limited range. Such a variable is known as a *limited dependent variable*. For example, in the study of the household expenditure on durable goods, Tobin (1958) noticed that positive expenditures can be observed only when households spend more than the cheapest available price; potential expenditures below the "minimum" price are not observable but appear as zeros. In this case, the expenditure data are said to be *censored*. On the other hand, data may be *truncated* if they are completely lost outside a given range; for example, income data may be truncated when they are below certain level. A proper specification for a limited dependent variable must take data censoring or truncation into account, as will be apparent in subsequent sub-sections.

10.4.1 Truncated Regression Models

First consider the dependent variable y_t which is truncated from below at c, i.e., $y_t > c$. The conditional density of truncated y_t is

$$\tilde{g}(y_t|y_t > c, \boldsymbol{x}_t) = g(y_t|\boldsymbol{x}_t)/\,\mathbb{P}(y_t > c|\boldsymbol{x}_t),$$

and we may approximate $g(y_t|\boldsymbol{x}_t)$ using

$$f(y_t|\boldsymbol{x}_t;\boldsymbol{\beta},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\Big(\frac{y_t-\boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\Big).$$

Setting $u_t = (y_t - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma$, we have

$$\begin{aligned}\mathbb{P}(y_t > c|\boldsymbol{x}_t) &= \frac{1}{\sigma}\int_c^\infty \phi\Big(\frac{y_t-\boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\Big)\,\mathrm{d}y_t \\ &= \int_{(c-\boldsymbol{x}_t'\boldsymbol{\beta})/\sigma}^\infty \phi(u_t)\,\mathrm{d}u_t \\ &= 1-\Phi\Big(\frac{c-\boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\Big).\end{aligned}$$

It follows that the truncated density function $\tilde{g}(y_t|y_t > c, \boldsymbol{x}_t)$ can be approximated by

$$f(y_t|y_t > c, \boldsymbol{x}_t;\boldsymbol{\beta},\sigma^2) = \frac{\phi[(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})/\sigma]}{\sigma\left[1-\Phi\big((c-\boldsymbol{x}_t'\boldsymbol{\beta})/\sigma\big)\right]},$$

which, of course, depends on the truncation parameter c.

The quasi-log-likelihood function is therefore

$$-\frac{1}{2}[\log(2\pi)+\log(\sigma^2)] - \frac{1}{2T\sigma^2}\sum_{t=1}^T (y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2 - \frac{1}{T}\sum_{t=1}^T \log\left[1-\Phi\Big(\frac{c-\boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\Big)\right].$$

Reparameterizing by $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$ and $\gamma = \sigma^{-1}$, we have

$$L_T(\boldsymbol{\theta}) = -\frac{\log(2\pi)}{2} + \log(\gamma) - \frac{1}{2T}\sum_{t=1}^{T}(\gamma y_t - \boldsymbol{x}_t'\boldsymbol{\alpha})^2 - \frac{1}{T}\sum_{t=1}^{T}\log\big[1 - \Phi(\gamma c - \boldsymbol{x}_t'\alpha)\big],$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}'\ \gamma)'$. The first-order condition is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\left[(\gamma y_t - \boldsymbol{x}_t'\boldsymbol{\alpha}) - \frac{\phi(\gamma c - \boldsymbol{x}_t'\boldsymbol{\alpha})}{1-\Phi(\gamma c - \boldsymbol{x}_t'\boldsymbol{\alpha})}\right]\boldsymbol{x}_t \\ \frac{1}{\gamma} - \frac{1}{T}\sum_{t=1}^{T}\left[(\gamma y_t - \boldsymbol{x}_t'\boldsymbol{\alpha})y_t - \frac{\phi(\gamma c - \boldsymbol{x}_t'\boldsymbol{\alpha})}{1-\Phi(\gamma c - \boldsymbol{x}_t'\boldsymbol{\alpha})}c\right] \end{bmatrix} = \mathbf{0},$$

from which we can solve for the QMLE of $\boldsymbol{\theta}$. It can be verified that $L_T(\boldsymbol{\theta})$ is globally concave in $\boldsymbol{\theta}$.

When $f(y_t|y_t > c, \boldsymbol{x}_t; \boldsymbol{\theta}_o)$ is the correct specification for $\tilde{g}(y_t|y_t > c, \boldsymbol{x}_t)$, it is not too difficult to derive the conditional mean of truncated y_t as

$$\begin{aligned} \mathbb{E}(y_t|y_t > c, \boldsymbol{x}_t) &= \int_c^\infty y_t f(y_t|y_t > c, \boldsymbol{x}_t; \boldsymbol{\theta}_o)\,\mathrm{d}y_t \\ &= \boldsymbol{x}_t'\boldsymbol{\beta}_o + \sigma_o \frac{\phi\big((c - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)}{1 - \Phi\big((c - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)}; \end{aligned} \tag{10.1}$$

which differs from $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ by a nonlinear term; see Exercise 10.8. Thus, the OLS estimator of regressing y_t on $\boldsymbol{x}_t$ would be inconsistent for $\boldsymbol{\beta}_o$ when y_t are truncated. This also shows why a proper model must consider data truncation.

Define the *hazard* function λ as

$$\lambda(u) = \phi(u)/[1 - \Phi(u)],$$

which is also known as the *inverse Mill's ratio*. Setting $u_t = (c - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o$, the truncated mean can be expressed as

$$\mathbb{E}(y_t|y_t > c, \boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \sigma_o\lambda(u_t).$$

It can also be shown that the truncated variance is

$$\mathrm{var}(y_t|y_t > c, \boldsymbol{x}_t) = \sigma_o^2\big[1 - \lambda^2(u_t) + \lambda(u_t)u_t\big],$$

instead of σ_o^2; see Exercise 10.9. The marginal response of $\mathbb{E}(y_t|y_t > c, \boldsymbol{x}_t)$ to a change of $\boldsymbol{x}_t$ is

$$\boldsymbol{\beta}_o\big[1 - \lambda^2(u_t) + \lambda(u_t)u_t\big] = \frac{\boldsymbol{\beta}_o}{\sigma_o^2}\ \mathrm{var}(y_t|y_t > c, \boldsymbol{x}_t),$$

which depends on the truncated variance.

Consider now the case that the dependent variable y_t is truncated from above at c, i.e., $y_t < c$. The truncated density function $\tilde{g}(y_t|y_t < c, \boldsymbol{x}_t)$ can be approximated by

$$f(y_t|y_t < c, \boldsymbol{x}_t; \boldsymbol{\beta}, \sigma^2) = \frac{\phi[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma]}{\sigma\,\Phi\big((c - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma\big)}.$$

Then, analogous to (10.1), we have

$$\mathbb{E}(y_t|y_t < c, \boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o - \sigma_o \frac{\phi\big((c - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)}{\Phi\big((c - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o\big)},$$

with the inverse Mill's ratio $\lambda(u) = -\phi(u)/\Phi(u)$.

10.4.2 Censored Regression Models

Consider the following censored dependent variable:

$$y_t = \begin{cases} y_t^*, & y_t^* > 0, \\ 0, & y_t^* \le 0, \end{cases}$$

where $y_t^* = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$ is an index variable. It does not matter whether the threshold value of y_t^* is zero or a non-zero constant c (e.g., the cheapest available price) because c can always be absorbed into the constant term in the specification of y_t^*.

Let g and g^* denote, respectively, the densities of y_t and y_t^* conditional on $\boldsymbol{x}_t$. When $y_t^* > 0$, $g(y_t|\boldsymbol{x}_t) = g^*(y_t^*|\boldsymbol{x}_t)$, and when $y_t^* \le 0$, censoring yields the conditional probability

$$\mathbb{P}(y_t = 0|\boldsymbol{x}_t) = \int_{-\infty}^{0} g^*(y_t^*|\boldsymbol{x}_t)\,\mathrm{d}y_t^*.$$

In the standard *Tobit* model (Tobin, 1958), $g^*(y_t^*|\boldsymbol{x}_t)$ is approximated by

$$f(y_t^*|\boldsymbol{x}_t; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t^* - \boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\Big(\frac{y_t^* - \boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\Big).$$

For $y_t^* \le 0$, $\mathbb{P}\{y_t = 0|\boldsymbol{x}_t\}$ is approximated by

$$\frac{1}{\sigma}\int_{-\infty}^{0} \phi((y_t^* - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma)\,\mathrm{d}y_t^* = \int_{-\infty}^{-\boldsymbol{x}_t'\boldsymbol{\beta}/\sigma} \phi(v_t)\,\mathrm{d}v_t = 1 - \Phi\Big(\frac{\boldsymbol{x}_t'\boldsymbol{\beta}}{\sigma}\Big).$$

Letting $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma$ and $\gamma = \sigma^{-1}$, the model would be correctly specified for $\{y_t^*|\boldsymbol{x}_t\}$ if there exists $\boldsymbol{\theta}_o = (\boldsymbol{\alpha}_o'\ \gamma_o)'$ such that $f(y_t^*|\boldsymbol{x}_t; \boldsymbol{\theta}_o) = g^*(y_t^*|\boldsymbol{x}_t)$.

Given the specification above, the quasi-log-likelihood function is

$$-\frac{T_1}{2T}[\log(2\pi) + \log(\sigma^2)] + \frac{1}{T}\sum_{\{t:y_t=0\}} \log(1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}/\sigma)) - \frac{1}{2T}\sum_{\{t:y_t>0\}} [(y_t - \boldsymbol{x}_t'\boldsymbol{\beta})/\sigma]^2,$$

where T_1 is the number of t such that $y_t > 0$. The QMLE of $\boldsymbol{\theta} = (\boldsymbol{\alpha}' \ \gamma)'$ can then be obtained by maximizing

$$L_T(\boldsymbol{\theta}) = \frac{1}{T}\left(\sum_{\{t:y_t=0\}} \log(1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\alpha})) + T_1 \log\gamma - \frac{1}{2}\sum_{\{t:y_t>0\}} (\gamma y_t - \boldsymbol{x}_t'\boldsymbol{\alpha})^2 \right).$$

The first order condition is

$$\nabla_{\boldsymbol{\theta}} L_T(\boldsymbol{\theta}) = \frac{1}{T}\begin{bmatrix} -\sum_{\{t:y_t=0\}} \frac{\phi(\boldsymbol{x}_t'\boldsymbol{\alpha})}{1-\Phi(\boldsymbol{x}_t'\boldsymbol{\alpha})}\boldsymbol{x}_t + \sum_{\{t:y_t>0\}}(\gamma y_t - \boldsymbol{x}_t'\boldsymbol{\alpha})\boldsymbol{x}_t \\ \frac{T_1}{\gamma} - \sum_{\{t:y_t>0\}}(\gamma y_t - \boldsymbol{x}_t'\boldsymbol{\alpha}) y_t. \end{bmatrix} = \mathbf{0},$$

from which the QMLE can be computed. Again, $L_T(\boldsymbol{\theta})$ is globally concave in $\boldsymbol{\alpha}$ and γ.

The conditional mean of censored y_t is

$$\begin{aligned} \mathbb{E}(y_t|\boldsymbol{x}_t) &= \mathbb{E}(y_t|y_t > 0, \boldsymbol{x}_t)\ \mathbb{P}(y_t > 0|\boldsymbol{x}_t) + \mathbb{E}(y_t|y_t = 0, \boldsymbol{x}_t)\ \mathbb{P}(y_t = 0|\boldsymbol{x}_t) \\ &= \mathbb{E}(y_t^*|y_t^* > 0, \boldsymbol{x}_t)\ \mathbb{P}(y_t^* > 0|\boldsymbol{x}_t). \end{aligned}$$

When $f(y_t^*|\boldsymbol{x}_t; \boldsymbol{\beta}, \sigma^2)$ is correctly specified for $g^*(y_t^*|\boldsymbol{x}_t)$,

$$\mathbb{P}(y_t^* > 0|\boldsymbol{x}_t) = \mathbb{P}\left(\frac{y_t^* - \boldsymbol{x}_t'\boldsymbol{\beta}_o}{\sigma_o} > \frac{-\boldsymbol{x}_t'\boldsymbol{\beta}_o}{\sigma_o} \middle| \boldsymbol{x}_t \right) = \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o).$$

This leads to the following conditional density:

$$\tilde{g}(y_t^*|y_t^* > 0, \boldsymbol{x}_t) = \frac{g^*(y_t^*|\boldsymbol{x}_t)}{\mathbb{P}(y_t^* > 0|\boldsymbol{x}_t)} = \frac{\phi[(y_t^* - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o]}{\sigma_o\, \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)}.$$

Then, analogous to (10.1), we have

$$\begin{aligned} \mathbb{E}(y_t^*|y_t^* > 0, \boldsymbol{x}_t) &= \int_0^\infty y_t^* \tilde{g}(y_t^*|y_t^* > 0, \boldsymbol{x}_t)\, \mathrm{d}\, y_t^* \\ &= \boldsymbol{x}_t'\boldsymbol{\beta}_o + \sigma_o \frac{\phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)}{\Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)}; \end{aligned} \tag{10.2}$$

see Exercise 10.10. It follows that

$$\mathbb{E}(y_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o\Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o) + \sigma_o\phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o).$$

This shows that $\boldsymbol{x}_t'\boldsymbol{\beta}$ can not be the correct specification of the conditional mean of censored y_t. Similar to truncated regression models, regressing y_t on $\boldsymbol{x}_t$ results in an inconsistent estimator for $\boldsymbol{\beta}_o$. It is also easy to calculate that the marginal response of $\mathbb{E}(y_t|\boldsymbol{x}_t)$ to a change of $\boldsymbol{x}_t$ is $\boldsymbol{\beta}_o\Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)$.

Consider the case that y_t is censored from above:

$$y_t = \begin{cases} y_t^*, & y_t^* < 0, \\ 0, & y_t^* \geq 0, \end{cases}$$

where $y_t^* = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$. When $f(y_t^*|\boldsymbol{x}_t;\boldsymbol{\beta},\sigma^2)$ is correctly specified for $g^*(y_t^*|\boldsymbol{x}_t)$,

$$\mathbb{P}(y_t^* < 0|\boldsymbol{x}_t) = 1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o).$$

Then,

$$\tilde{g}(y_t^*|y_t^* < 0, \boldsymbol{x}_t) = \frac{g^*(y_t^*|\boldsymbol{x}_t)}{\mathbb{P}(y_t^* < 0|\boldsymbol{x}_t)} = \frac{\phi[(y_t^* - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o]}{\sigma_o\,[1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)]},$$

and hence, analogous to (10.2),

$$\mathbb{E}(y_t^*|y_t^* < 0, \boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o - \sigma_o\,\frac{\phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)}{1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)}.$$

Consequently,

$$\mathbb{E}(y_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o[1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)] - \sigma_o\phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o).$$

and its marginal response to a change of $\boldsymbol{x}_t$ is $\boldsymbol{\beta}_o[1 - \Phi(\boldsymbol{x}_t'\boldsymbol{\beta}_o/\sigma_o)]$.

10.4.3 Models of Sample Selection

It is also common to encounter the problem of *sample selection* or *incidental truncation* in empirical studies. For example, in the study on how wages and individual characteristics determine working hours, the data of working hours are not resulted from random sampling but from those who select to work. In other words, only those whose market wages are above their reservation wages (the minimum wages for which they are willing to work) may be observed. Thus, the decision of working must be related to working hours.

Consider two variables y_1 and y_t, where y_1 is a binary choice variable. The selection problem arises when the former affects the latter. These two variables are

$$y_{1,t} = \begin{cases} 1, & y_{1,t}^* > 0, \\ 0, & y_{1,t}^* \le 0. \end{cases}$$

$$y_{2,t} = y_{2,t}^*, \quad \text{if } y_{1,t} = 1,$$

where $y_{1,t}^* = \boldsymbol{x}_{1,t}'\boldsymbol{\beta} + e_{1,t}$ and $y_{2,t}^* = \boldsymbol{x}_{2,t}'\boldsymbol{\gamma} + e_{2,t}$ are index variables. We observe $y_{1,t}$ but not the index variable $y_{1,t}^*$. Also, we observe $y_{2,t}$ when $y_{1,t} = 1$ so that $y_{2,t}$ are *incidentally truncated* when $y_{1,t} = 0$. It is typical to model $(y_{1,t}^*\ y_{2,t}^*)'$ based on a bivariate normal distribution:

$$\mathcal{N}\left(\begin{pmatrix} \boldsymbol{x}_{1,t}'\boldsymbol{\beta} \\ \boldsymbol{x}_{2,t}'\boldsymbol{\gamma} \end{pmatrix},\ \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right).$$

When this is a correct specification for the conditional distribution of $(y_{1,t}^*\ y_{2,t}^*)'$, the corresponding true parameters are $\boldsymbol{\beta}_o$, $\boldsymbol{\gamma}_o$, $\sigma_{1,o}^2$, $\sigma_{2,o}^2$ and $\sigma_{12,o}$.

Computing the QMLE based on this bivariate specification is quite complicated. Heckman (1976) suggest a simpler, two-step estimator for the sample-selection model. For notation simplicity, we shall, in what follows, omit the conditioning variables $\boldsymbol{x}_{1,t}$ and $\boldsymbol{x}_{2,t}$. First recall that the conditional mean of $y_{2,t}$ given $y_{1,t}$ is

$$\mathbb{E}(y_{2,t}^*|y_{1,t}^*) = \boldsymbol{x}_{2,t}'\boldsymbol{\gamma}_o + \sigma_{12,o}(y_{1,t}^* - \boldsymbol{x}_{1,t}'\boldsymbol{\beta}_o)/\sigma_{1,o}^2.$$

The result of the truncated regression model also shows that, when the truncation parameter $c = 0$,

$$\mathbb{E}(y_{1,t}^*|y_{1,t}^* > 0) = \boldsymbol{x}_{1,t}'\boldsymbol{\beta}_o + \sigma_{1,o}\lambda(\boldsymbol{x}_{1,t}'\boldsymbol{\beta}_o/\sigma_{1,o}),$$

with $\lambda(u) = \phi(u)/\Phi(u)$. Putting these results together we have

$$\mathbb{E}(y_{2,t}^*|y_{1,t}^* > 0) = \boldsymbol{x}_{2,t}'\boldsymbol{\gamma}_o + \frac{\sigma_{12,o}}{\sigma_{1,o}}\lambda\Big(\frac{\boldsymbol{x}_{1,t}'\boldsymbol{\beta}_o}{\sigma_{1,o}}\Big).$$

This motivates the following specification:

$$y_{2,t} = \boldsymbol{x}_{2,t}'\boldsymbol{\gamma} + \frac{\sigma_{12}}{\sigma_1}\lambda(\boldsymbol{x}_{1,t}'\boldsymbol{\beta}/\sigma_1) + e_t,$$

which is a complicated nonlinear regression with heteroskedastic errors.

Heckman's two-step estimator is computed as follows.

1. Estimate $\boldsymbol{\alpha} = \boldsymbol{\beta}/\sigma_1$ based on the probit model of $y_{1,t}$ and denote the estimator as $\tilde{\boldsymbol{\alpha}}_T$.

2. Regress $y_{2,t}$ on $\boldsymbol{x}_{2,t}$ and $\lambda(\boldsymbol{x}_{1,t}'\tilde{\boldsymbol{\alpha}}_T)$.

As the QMLE $\tilde{\boldsymbol{\alpha}}_T$ in the first step is consistent for $\boldsymbol{\alpha}_o$, the second step is equivalent to regressing $y_{2,t}$ on $\boldsymbol{x}_{2,t}$ and $\lambda(\boldsymbol{x}_{1,t}'\boldsymbol{\alpha}_o)$. Thus, the OLS estimator of regressing $y_{2,t}$ on $\boldsymbol{x}_{2,t}$ alone, which ignores the λ term, is inconsistent and may have severe sample-selection bias in finite samples.

Exercises

10.1 In the logit model with the $k \times 1$ parameter vector $\boldsymbol{\theta}$, consider the null hypothesis that an $s \times 1$ subvector $\boldsymbol{\theta}$ is zero, where $s < k$. Write down the Wald statistic with a heteroskedasticity-consistent covariance matrix estimator.

10.2 In the logit model with the $k \times 1$ parameter vector $\boldsymbol{\theta}$, consider the null hypothesis that an $s \times 1$ subvector $\boldsymbol{\theta}$ is zero, where $s < k$. Write down the LM statistic with a heteroskedasticity-consistent covariance matrix estimator.

10.3 Construct a PSE test for testing the null hypothesis of a logit model with $F(\boldsymbol{x}_t;\boldsymbol{\theta}) = G(\boldsymbol{x}_t'\boldsymbol{\theta})$ against the alternative hypothesis of a probit model with $F(\boldsymbol{x}_t;\boldsymbol{\theta}) = \Phi(\boldsymbol{x}_t'\boldsymbol{\theta})$.

10.4 In the conditional logit model with the $k \times 1$ parameter vector $\boldsymbol{\theta}$, consider the null hypothesis that an $s \times 1$ subvector $\boldsymbol{\theta}$ is zero, where $s < k$. Write down the Wald statistic with a heteroskedasticity-consistent covariance matrix estimator.

10.5 For the ordered logit model, find the marginal responses of the choice probabilities to the change of the variables.

10.6 In the Poisson regression model with the parameter vector $\boldsymbol{\theta}$, consider the null hypothesis $\boldsymbol{R\theta} = \boldsymbol{r}$, with $\boldsymbol{R}$ a $q \times k$ given matrix. Write down the Wald statistic with a heteroskedasticity-consistent covariance matrix estimator.

10.7 In the Poisson regression model with the parameter vector $\boldsymbol{\theta}$, consider the null hypothesis $\boldsymbol{R\theta} = \boldsymbol{r}$, with $\boldsymbol{R}$ a $q \times k$ given matrix. Write down the LM statistic with a heteroskedasticity-consistent covariance matrix estimator.

10.8 In the truncated regression model with the truncation parameter c, show that $\mathbb{E}(y_t|y_t > c, \boldsymbol{x}_t)$ is given by (10.1).

10.9 In the truncated regression model with the truncation parameter c, show that $\operatorname{var}(y_t|y_t > c, \boldsymbol{x}_t) = \sigma_o^2\big[1 - \lambda^2(u_t) + \lambda(u_t)u_t\big]$, where $\lambda(u_t) = \phi(u_t)/[1 - \Phi(u_t)]$ and $u_t = (c - \boldsymbol{x}_t'\boldsymbol{\beta}_o)/\sigma_o$.

10.10 In the Tobit model, show that $\mathbb{E}(y_t^*|y_t^* > 0, \boldsymbol{x}_t)$ is given by (10.2).

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Heckman, James J. (1976). Common structure of statistical medels of truncation, sample selection, and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement*, **5**, 475–492.

McFadden, Daniel L. (1973). Conditional logit analysis of qualitative choice behavior, in P. Zarembka (ed.), *Frontier of Econometrics*, New York: Academic Press.

Tobin, James (1958). Estimation of relationships for limited dependent variables, *Econometrica*, **26**, 24–36.