from
Maddala, G. S. (1983), Limited-Dependent and Qualitative Variables in Economics, New York: Cambridge University Press, pp. 257-91.

CHAPTER 9

# Models with self-selectivity

## 9.1 Introduction

There are many problems in which the data we have are generated by individuals making choices of belonging to one group or another (i.e., by individual self-selection). An early discussion of this problem of self-selectivity was that of Roy (1951), who discussed the problem of individuals choosing between two professions, hunting and fishing, based on their productivity in each. The observed distribution of incomes of hunters and fishermen was determined by these choices.

Suppose $Y_{1i}$ is the output of the $i$th individual in hunting and $Y_{2i}$ the output in fishing. Individual $i$ will choose to be a hunter if $Y_{1i} > Y_{2i}$. Output here is defined in dollar terms. Assume that $(Y_1, Y_2)$ have a joint normal distribution, with means $(\mu_1, \mu_2)$ and covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Define

$$u_1 = Y_1 - \mu_1, \quad u_2 = Y_2 - \mu_2, \quad \sigma^2 = \text{Var}(u_1 - u_2),$$

$$Z = \frac{\mu_1 - \mu_2}{\sigma} \quad \text{and} \quad u = \frac{u_2 - u_1}{\sigma}$$

The condition $Y_1 > Y_2$ implies $u < Z$. The mean income of hunters is given by

$$E(Y_1 \mid u < Z) = \mu_1 - \sigma_{1u} \frac{\phi(Z)}{\Phi(Z)} \tag{9.1}$$

where $\sigma_{1u}=\text{Cov}(u_1, u)$ and $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the density function and the distribution function of the standard normal. The mean income of fishermen is given by

$$E(Y_2 \mid u > Z) = \mu_2 + \sigma_{2u}\frac{\phi(Z)}{1-\Phi(Z)} \tag{9.2}$$

where $\sigma_{2u}=\text{Cov}(u_2, u)$. Because

$$\sigma_{1u} = \frac{\sigma_{12}-\sigma_1^2}{\sigma} \quad \text{and} \quad \sigma_{2u} = \frac{\sigma_2^2-\sigma_{12}}{\sigma}$$

we have $\sigma_{2u}-\sigma_{1u}>0$. We can now consider different cases.

*Case 1.* $\sigma_{1u}<0$, $\sigma_{2u}>0$. In this case the mean income of hunters is greater than $\mu_1$ and the mean income of fishermen is greater than $\mu_2$; that is, those who chose hunting are better than average hunters, and those who chose fishing are better than average fishermen.

Case 2. $\sigma_{1u}<0$, $\sigma_{2u}<0$. In this case the mean income of hunters is greater than $\mu_1$, and the mean income of fishermen is less than $\mu_2$. In this case those who chose hunting are better than average in both hunting and fishing, but they are better in hunting than in fishing. Those who chose fishing are below average in both hunting and fishing, but they are better in fishing than in hunting.

*Case 3.* $\sigma_{1u}>0$, $\sigma_{2u}>0$. This is the reverse of case 2.

*Case 4.* $\sigma_{1u}>0$, $\sigma_{2u}<0$. This is not possible, given the definitions of $\sigma_{1u}$ and $\sigma_{2u}$.

Note that case 2 typically occurs if $\sigma_1$ is very large compared with $\sigma_2$. Thus, the individuals with better skills go into the profession with higher variance in earnings.

The more detailed analysis of this model can be found in the work of Roy (1951). The important thing to note here is the importance of the covariance terms $\sigma_{1u}$ and $\sigma_{2u}$ in the interpretation of the results. We shall see later how they play an important role in discussions of selectivity bias.

The econometric discussion of the consequences of self-selectivity began with the studies by Gronau (1974), Lewis (1974), and Heckman (1974). In this case the problem is about women choosing to be in the labor force or not. The observed distribution of wages is a truncated distribution. It is the distribution of wage offers truncated by reservation wages. The Gronau–Lewis model consisted of two equations:

$$W_o = X\beta_1 + u_1$$
$$W_r = X\beta_2 + u_2 \tag{9.3}$$

We observe $W = W_o$ if and only if $W_o \geq W_r$. Otherwise, $W = 0$. We discussed the estimation of this model in Chapter 8, and we shall not repeat it here. The term *selectivity bias* refers to the fact that if we estimate equation (9.3) by OLS, based on the observations for which we have wages $W$, we get inconsistent estimates of the parameters. Note that

$$E(u_1 \mid W_o \geq W_r) = -\sigma_{1u}\frac{\phi(Z)}{\Phi(Z)}$$

where $Z = (X\beta_1 - X\beta_2)/\sigma$ and the other terms are as defined earlier. Hence, we can write (9.3) as

$$W = X\beta_1 - \sigma_{1u}\frac{\phi(Z)}{\Phi(Z)} + V \tag{9.4}$$

where $E(V) = 0$. A test for selectivity bias is a test for $\sigma_{1u} = 0$. Heckman (1976*b*) suggested a two-stage estimation method for such models. First, get consistent estimates for the parameters in $Z$ by the probit method applied to the dichotomous variable (in the labor force or not). Then estimate equation (9.4) by OLS, using the estimated values $\hat{Z}$ for $Z$. This two-stage method has been discussed in detail in Chapter 8.

The self-selectivity problem has more recently been analyzed in different contexts by several people. Lee and Trost (1978) applied it to the problem of housing demand, with choices of owning and renting. Willis and Rosen (1979) applied the model to the problem of education and self-selection. These are all switching regression models. Griliches et al. (1978) and Kenny et al. (1979) considered models with both selectivity and simultaneity. These models are switching simultaneous-equations models. As for methods of estimation, both two-stage and maximum-likelihood methods have been used. For two-stage methods, the study by Lee et al. (1980) gave the asymptotic covariance matrices when the selectivity criterion was of the probit and tobit types (see Chapter 8).

In the literature on self-selectivity, a major concern has been with testing for selectivity bias. These are tests for $\sigma_{1u} = 0$ and $\sigma_{2u} = 0$ in equations of the form (9.1) and (9.2). However, a more important issue concerns the signs and magnitudes of these covariances, and often not much attention is devoted to this. In actual practice, we ought to have $\sigma_{2u} - \sigma_{1u} > 0$, but $\sigma_{1u}$ and $\sigma_{2u}$ can have any signs.[1] It is also important to

[1] Trost (1981) discussed this point in reference to returns from college education.

estimate the mean values of the dependent variables for the alternative choice. for instance, in the case of college education and income, we should estimate the mean income of college graduates had they chosen not to go to college and the mean income of non-college-graduates had they chosen to go to college. In the example of hunting and fishing, we should compute the mean income of hunters had they chosen to be fishermen and the mean income of fishermen had they chosen to be hunters. Such computations throw light on the effects of self-selection and also reveal deficiencies in the model that are not revealed by simple tests for the existence of selectivity bias. In the example concerning hunting and fishing, the mean income of hunters, had they chosen fishing, would be

$$E(Y_2 \mid Y_1 > Y_2) = E(Y_2 \mid u_2 > Z)$$
$$= \mu_2 - \sigma_{2u}\frac{\phi(Z)}{\Phi(Z)}$$

and the mean income of fishermen, had they chosen hunting, would be

$$E(Y_1 \mid Y_1 < Y_2) = \mu_1 + \sigma_{1u}\frac{\phi(Z)}{1-\Phi(Z)}$$

Also, if we denote by $\overline{Y_1}$ and $\overline{Y_2}$ the actual mean incomes of hunters and fishermen, then from (9.1) and (9.2) we have

$$E(\overline{Y_1} - \overline{Y_2}) = \mu_1 - \mu_2 - \sigma_{1u}\frac{\phi(Z)}{\Phi(Z)} - \sigma_{2u}\frac{\phi(Z)}{1-\Phi(Z)}$$

If $\sigma_{1u}$ and $\sigma_{2u}$ are both negative, then $\overline{Y_1}-\overline{Y_2}$ is an upward-biased estimate of $\mu_1-\mu_2$. If $\sigma_{1u}$ and $\sigma_{2u}$ are both positive, then $\overline{Y_1}-\overline{Y_2}$ is a downward-biased estimate of $\mu_1-\mu_2$. If $\sigma_{1u}<0$ and $\sigma_{2u}>0$, the direction of bias is not unambiguous.[2]

The foregoing discussion generalizes easily to models with explanatory variables. All we do is substitute $\mu_1=\beta_1'X_1$ and $\mu_2=\beta_2'X_2$ in all the expressions.

## 9.2 Self-selection and evaluation of programs

One major use of the self-selection models is in evaluating the benefits of social programs. To evaluate the benefit from a program, a model commonly employed is the following:

$$Y = X\beta + \alpha I + u \tag{9.5}$$

[2] Some illustrative examples are given by Maddala (1977*a*). Because the examples can be worked out easily, they will not be repeated here.

where $Y$ is the outcome (test score, earnings, etc.), $X$ is a vector of exogenous personal characteristics, and $I$ is a dummy variable ($I=1$ if the individual participates in the program; $I=0$ otherwise). For this model, the effect of the program is measured by the estimate of $\alpha$. However, the dummy variable $I$ cannot be treated as exogenous if the decision of an individual to participate or not participate in the program is based on individual self-selection. If the variable $I$ is endogenous, equation (9.5) must be estimated by instrumental-variable techniques.

The foregoing model is very restrictive, because the program may create interaction effects with observed or unobserved personal characteristics; a more general model is the following:

$$y_{1i} = X_i\beta_1 + u_{1i} \quad \text{(for participants)}$$
$$y_{2i} = X_i\beta_2 + u_{2i} \quad \text{(for nonparticipants)}$$
$$I_i^* = Z_i\gamma = \epsilon_i \quad \text{(participation decision function)}$$
$$I_i = 1 \quad \text{iff } I_i^* > 0$$
$$I_i = 0 \quad \text{iff } I_i^* \leqslant 0$$

The observed $y_i$ is defined as

$$y_i = y_{1i} \quad \text{iff } I_i = 1$$
$$y_i = y_{2i} \quad \text{iff } I_i = 0$$
$$\text{Cov}(u_{1i}, u_{2i}, \epsilon_i) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1\epsilon} \\ \sigma_{12} & \sigma_{22} & \sigma_{2\epsilon} \\ \sigma_{1\epsilon} & \sigma_{2\epsilon} & 1 \end{bmatrix}$$

To evaluate the benefit of the program that has already been created, we need to consider the total gross benefit for all the participants. For each participant with characteristics $X_i$ and $Z_i$, we can compare the outcome $y_{1i}$ in the program and the expected potential outcome without the program, that is, $E(y_{2i} \mid I_i=1)$. Under the normality assumption, the gross benefit for participants $i$ is

$$y_{1i} - E(y_{2i} \mid I_i = 1) = y_{1i} - X_i\beta_2 + \sigma_{2\epsilon}\frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \tag{9.6}$$

The total benefit is the summation of (9.6) over all the participants. Thus, to evaluate the success of a program from the cost-benefit point of view, the conditional expectation of $u_{2i}$ for the participants needs to be evaluated. Note that, under self-selection, those individuals who have a comparative advantage with the program will be joining the program and thus will benefit more from it than would a randomly selected individual with the same characteristics. The expected gross benefit for participant $i$ is

$$E(y_{1i} \mid I_i = 1) - E(y_{2i} \mid I_i = 1) = X_i(\beta_1 - \beta_2) + (\sigma_{2\epsilon} - \sigma_{1\epsilon}) \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \quad (9.7)$$

If self-selection is based on comparative advantage, as in Roy's example on hunting and fishing in the preceding section, $\sigma_{2\epsilon} - \sigma_{1\epsilon}$ is greater than zero.[3] Thus, the program will produce greater benefit under self-selection than under a random assignment. The difference is measured by the summation of the last term in (9.7) over all participants.

The preceding discussion assumes that there are only two groups of individuals, one participating in the program (treatment group) and the other not participating (control group), and the assignment of individuals to the two groups is by self-selection rather than by random assignment. There can be other types of self-selection. Suppose that there is a social experiment (say a time-of-day pricing experiment) for which we draw a random sample. Some of the individuals in the sample may not wish to participate. Among those who participate, the assignment to the control or treatment group could be a random assignment. In this case the self-selection is at the stage of entering an experiment. What one will do is estimate an equation of the form (9.5) using the method of the censored or truncated regression models described in Chapter 6. What we have is a model of the form

$$I^* = Z\gamma - \epsilon \quad (9.8)$$

The individual is in the experiment, and $I = 1$ if and only if $I^* > 0$. Otherwise, the individual is not in the experiment. Also,

$$Y = X\beta + \alpha D + u \quad (9.9)$$

where $D = 1$ if the individual is in the treatment group and $D = 0$ if in the control group. Because the assignment to the treatment and control groups is random, $D$ is an exogenous dummy variable. However, there is censoring or truncation produced by (9.8). If data on $Z$ are available on all individuals, we shall estimate (9.9) as a censored regression model. If data on $Z$ are available only for the participants in the experiment, we shall estimate equation (9.9) as a truncated regression model. If the residuals $\epsilon$ and $u$ in equations (9.8) and (9.9) are independent, of course we can estimate (9.9) by the OLS method. The important thing, however, is that $D$ is exogenous, because individuals are randomly assigned to the control and treatment groups.

What if there is self-selection at the stage of choosing whether or not to participate in the experiment and also at the stage of choosing

[3] See also the work of Lee (1979*b*) on self-selection and comparative advantage.

between the treatment and control groups. There is the question whether we want to treat this as a trichotomous-choice model or a sequential self-selection model. In the trichotomous-choice model, the individual has to choose among three altenatives: to belong to the treatment group, to belong to the control group, not to participate in the experiment at all. In the sequential self-selection model, the individual first chooses whether or not to participate in the experiment, and those who decide to participate then decide whether to go into the treatment group or the control group. Such selectivity models with polychotomous choices and sequential choices will be discussed later in this chapter (section 9.4).

A third alternative is where the assignment of individuals to the control and treatment groups is made by the program administrator on the basis of a screening variable that is itself correlated with $X$ in equation (9.9). Goldberger (1972) analyzed this problem and pointed out that there are some misconceptions about the biases in the estimates of treatment effects in such cases. For example, suppose the selection procedure is to put lower-ability students into the treatment group and higher-ability students into the control group, as in the Head Start compensatory educational programs. Because ability is not measurable, the program administrator uses the pretest score, say $Z$ (measured as deviations from the mean). So the assignment is

$$\begin{aligned} D &= 1 \quad \text{if } Z < 0 \\ D &= 0 \quad \text{otherwise} \end{aligned} \tag{9.10}$$

After completion of the program, one looks at the posttest score $Y$. If $Y$ is then regressed on $Z$ and $D$, that is

$$Y = \beta Z + \alpha D + \epsilon \tag{9.11}$$

and the estimate of $\alpha$ is not significantly different from zero, it is often argued that this is not really proof that the program is not working, because the students assigned to the treatment group are students with lower ability. What Goldberger pointed out is that the estimation of equation (9.11) nevertheless produces an unbiased estimate of $\alpha$, the coefficient of $D$, or the treatment effect. The reasoning behind this fact is that controlling for $Z$ eliminates any correlation of $D$ with the other variables.

The formal argument runs as follows: Let us denote the unobserved ability variable by $X$. Because it affects the posttest score $Y$, we have

$$Y = \gamma_1 X + \alpha D + \epsilon_1 \tag{9.12}$$

The pretest score $Z$ also depends on ability. Hence,

$$Z = \gamma_2 X + \epsilon_2 \tag{9.13}$$

Suppose ability $X$ is indeed measurable. Then, of course, estimation of (9.12) by OLS will give inconsistent estimates of the parameters $\gamma_1$ and $\alpha$ so long as $\epsilon_1$ and $\epsilon_2$ are correlated. But we also know the methods of obtaining consistent estimates of the parameters in this case by correcting for the selection bias. Consider now the case in which $X$ is not observed, but we know its determinants $W$, so that

$$X = \theta W + v$$

Substituting this in equations (9.12) and (9.13), we get equations of the form

$$Y = \theta_1 W + \alpha D + u_1 \tag{9.14}$$

$$Z = \theta_2 W + u_2 \tag{9.15}$$

Again, estimation of (9.14) by OLS gives inconsistent estimates of the parameters if $u_1$ and $u_2$ are correlated, but again we know how to get consistent estimates. The model is again a model with sample selectivity that has been considered earlier. This is the case considered by Barnow et al. (1981).

Consider, finally, the case in which all we have are equations (9.12) and (9.13) and $X$ is not observable; that is, we have pretest score, posttest score, and the dummy variable $D$, which is itself determined by the pretest score. Eliminating $X$, we get

$$\begin{aligned} Y &= \frac{\gamma_1}{\gamma_2}(Z - \epsilon_2) + \alpha D + \epsilon_1 \\ &= \gamma Z + \alpha D + (\epsilon_1 - \gamma\epsilon_2) \end{aligned} \tag{9.16}$$

where $\gamma = \gamma_1/\gamma_2$. Now the question is what we can say about the estimate of $\alpha$ when (9.16) is estimated by OLS. The answer, as shown by Goldberger (1972), is that $\text{Plim}\,\hat{\alpha} = \alpha$.

The preceding discussion referred to the program administrator's assignment of individuals to the treatment and control groups. In practice, in many programs with eligibility requirements and so on, we can have the twin problems of individual decision whether or not to participate and the program administrator's decision whether or not to choose. This is a sequential-decision model with partial observability, and we shall discuss it in a later section. There are two decision variables, $I_1$ and $I_2$, and we observe the variable $Y$ if and only if $I_1 > 0$ and $I_2 > 0$. In such problems there is the further complication that the pool of applicants may be only a self-selected subsample of all those who wish to participate, because many may not apply if they know that there is a long waiting list. However, there is no easy way to deal with this problem of the discouraged applicants.

Yet another complication is that of attrition of dropout of people from the experiment. Some participants inevitably drop out of the experiment before the treatment response is measured. One way of modeling this phenomenon is as follows: Define

$$\begin{aligned} I_i^* &= Z_i\gamma - \epsilon_1 \\ A_i^* &= Z_i\delta - \epsilon_2 \\ Y_{1i} &= X_{1i}\beta_1 + \epsilon_3 \\ Y_{2i} &= X_{2i}\beta_2 + \alpha T + \epsilon_4 \end{aligned} \tag{9.17}$$

where

$I_i = 1$ and the individual participates in the experiment iff $I_i^* > 0$

$I_i = 0$ otherwise

$A_i = 1$ and the individual continues in the experiment iff $A_i^* > 0$

$A_i = 0$ otherwise (the individual drops out)

If $I_i = 0$, neither $Y_{1i}$ nor $Y_{2i}$ is observed. If $I_i = 1$, $A_i = 0$, we observe only $Y_{1i}$. If $I_i = 1$, $A_i = 1$, we observe both $Y_{1i}$ and $Y_{2i}$.

An example of the estimation of this model is that of Venti and Wise (1980). We shall discuss some limitations of such models later in the section on multiple criteria of selectivity.

In summary, in evaluating the effects of several social programs, we must consider the selection and truncation that can occur at different levels. We can depict the situation by a decision tree (Figure 9.1). In practical situations, one must assume randomness at certain levels, or else the model can get too unwieldy to be of any use. As to the level at which selection and truncation bias needs to be introduced, this is a question that depends on the nature of the problem. Further, in Figure 9.1 the individual's decision to participate preceded the administrator's decision to select. This situation can be reversed, or the decisions can be simultaneous. Problems of sequential versus joint selection will be discussed in section 9.6. Another problem is that caused by the existence of multiple categories, such as no participation, partial participation, or full participation, or different types of treatment. These cases fall in the class of models with polychotomous choice and selectivity that will be discussed in section 9.5.

Finally, there is the problem of truncated samples. Very often we do not have data on all the individuals, participants and nonparticipants. If the data involve only participants in a program, but we know nevertheless that there is self-selection and we have data on the variables determining the participation decision function, then we can still correct for selectivity bias, although the two-stage methods described in the previous chapters are not applicable. What we have is the model

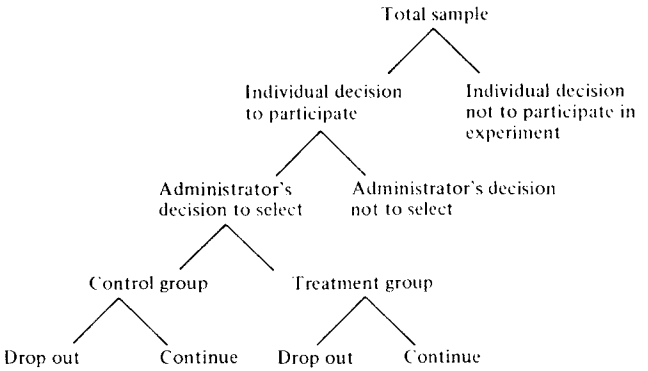


Figure 9.1. Decision tree for evaluation of social experiments

$$y_{1i} = X_i\beta_1 + u_{1i}$$
$$I_i^* = Z_i\gamma - \epsilon_i$$

As before,

$$I_i = 1 \quad \text{iff } I_i^* > 0$$
$$I_i = 0 \quad \text{otherwise}$$

We are given only those observations for which $I_i=1$, and for these we observe $y_{1i}$, $X_i$, and $Z_i$. The probit estimates of $\gamma$ cannot be obtained because we do not have the observations corresponding to $I_i=0$. Thus, we cannot use the two-stage methods. But we can use the ML method to correct for the selectivity bias. The model is different from the truncated regression model considered in section 6.9 in that the truncation is now based on an unobserved indicator $I_i^*$ rather than the variable $y_{1i}$.

The likelihood function for the model is

$$L = \prod_i \frac{\int_{-\infty}^{Z_i\gamma} f(u_{1i}, \epsilon_i)\, d\epsilon_i}{\text{Prob}(I_i=1)}$$

where $f(u_1, \epsilon)$ is the joint density function of $u_1$ and $\epsilon$. If we assume that $u_1$ and $\epsilon$ are jointly normally distributed, with mean vector zero and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{bmatrix}$$

then by writing $f(u_i, \epsilon)$ as $f_1(u_1)\cdot f_2(\epsilon \mid u_1)$, we can simplify the likelihood function as

$$L = \prod_i [\Phi(Z_i\gamma)]^{-1} \frac{1}{\sigma_1} \exp\left[-\frac{1}{2\sigma_1^2}(y_{1i} - X_i\beta_1)^2\right]$$

$$\times \Phi\left(\frac{[Z_i\gamma-\rho(y_{1i}-X_i\beta_1)]/\sigma_1}{(1-\rho^2)^{1/2}}\right)$$

The first expression is $\text{Prob}(I_i=1)$. The second is $f_1(u_1)$. The third is $\text{Prob}(\epsilon<Z\gamma)$, derived from the conditional density $f_2(\epsilon \mid u_1)$. Note that $\text{Var}(\epsilon)=1$, by normalization.

The problem of truncated samples can be handled in a similar manner with the other problems of self-selection and hence will not be elaborated here. The important thing to note is that although, theoretically, truncation does not change the identifiability of the parameters, there is nevertheless a loss of information. It is usually the case that even though we are able to correct for the selectivity bias in the OLS estimates $\beta_1$, the estimates of the parameters $\gamma$ in the selectivity criterion are not reliable. Muthén and Jöreskog (1981) reported the results of some simulated experiments they did with a simple selectivity model with truncated data and censored data. They conducted two studies, one with a sample size of 1,000 and the other with a sample size of 4,000. For each study they considered two cases. In case 1 the proportion of observations with $I_i=1$ was roughly 50%. In case 2 the proportion of observations with $I_i=1$ was roughly 75%. The difference between the censored and truncated samples was that in the case of the censored sample, data were assumed to be available on the variable $Z$ for all observations, whereas in the truncated case, data were assumed to be available on only the subsample for which $I_i=1$. In both cases, data on $X_i$ were assumed to be available for only the subsample for which $I_i=1$. Their finding was that even with such large samples, in the truncated case it was not possible to get good estimates of the parameters $\gamma$ in the selectivity criterion, although it was possible to correct for selectivity bias in the $\beta$ coefficients. Table 9.1 presents the results for the case in which $N=1{,}000$ and $I_i=1$ for about 50% of the observations.

In summary, even if the available data are for subgroups and are thus truncated samples, one can try to correct for the selectivity bias in the OLS estimates by using the ML method described here, provided one has a clear notion of what variables affect the selectivity criterion. However, one cannot expect to have good estimates for the parameters in the selectivity criterion itself.

## 9.3 Selectivity bias with nonnormal distributions

In the preceding sections, and in the several examples in Chapter 8, we discussed the selectivity-bias problem under the assumption that the disturbances are normally distributed. We shall now consider methods of relaxing this assumption.

Table 9.1. *Estimation of selectivity models from truncated and censored samples (simulated data)*

| Parameter | Population value | OLS estimates | Probit estimates | ML (truncated sample) | ML (censored sample) |
|---|---|---|---|---|---|
| $\beta_0$ | 0.0 | −0.373<br>(0.054) | | −0.209<br>(0.119) | 0.074<br>(0.179) |
| $\beta_1$ | 1.0 | 0.788<br>(0.052) | | 0.931<br>(0.095) | 1.033<br>(0.114) |
| $\sigma_1^2$ | 1.0 | 0.985<br>(0.065) | | 0.982<br>(0.076) | 1.126<br>(0.131) |
| $\gamma_0$ | 0.0 | | 0.011<br>(0.046) | 0.991<br>(1.599) | 0.013<br>(0.046) |
| $\gamma_1$ | −1.0 | | −1.033<br>(0.067) | −3.448<br>(4.542) | −1.040<br>(0.068) |
| $\rho$ | −0.5 | | | −0.248<br>(0.413) | −0.522<br>(0.164) |

*Note:* Standard errors in parentheses. $N = 1{,}000$ and $I_i = 1$ for 503 observations.
*Source:* Muthén and Jöreskog (1981, Table 3).

Consider the simple two-equation model

$$Y_1 = X\beta + u \tag{9.18}$$

$$Y^* = Z\gamma - \epsilon \tag{9.19}$$

where $X$ and $Z$ are exogenous variables. Equation (9.19) is the selectivity criterion. The dependent variable $Y^*$ is never observable, but it has a dichotomous realization $I$ that is related to $Y^*$ as follows:

$$I = 1 \quad \text{iff} \quad Y^* \geqslant 0$$
$$I = 0 \quad \text{otherwise}$$

The dependent variable $Y_1$ conditional on $X$ and $Z$ has a well-defined marginal distribution, but $Y_1$ is not observed unless $Y^* > 0$. Thus, the observed distribution of $Y_1$ is truncated.

Regarding the disturbances, we assume that

$$E(u \mid X, Z) = 0, \quad V(u \mid X, Z) = \sigma_u^2, \quad \text{Cov}(u, \epsilon \mid X, Z) = \rho\sigma_u\sigma_\epsilon$$
$$E(\epsilon \mid X, Z) = \mu_\epsilon, \quad \text{and} \quad V(\epsilon \mid X, Z) = \sigma_\epsilon^2$$

In all the examples in the preceding section, as well as in Chapter 8, we assumed $\mu_\epsilon = 0$ and $\sigma_\epsilon = 1$. But here we shall not make that assumption as yet. Following Olsen (1980*a*), we also assume that the conditional expectation of $u$, given $\epsilon$, is linear, so that

$$u = \frac{\rho\sigma_u}{\sigma_\epsilon}(\epsilon - \mu_\epsilon) + v \tag{9.20}$$

and $E(v \mid \epsilon) = 0$ and $\text{Var}(v \mid \epsilon) = \sigma_u^2(1-\rho^2)$.

Consider now the censored sample of $Y_1$. From (9.20) we get

$$E(Y_1 \mid I=1) = X\beta + E(u \mid \epsilon < Z\gamma) = X\beta + \frac{\rho\sigma_u}{\sigma_\epsilon}[g(Z\gamma) - \mu_\epsilon] \tag{9.21}$$

where $g(Z\gamma)$ is the truncated mean $E(\epsilon \mid \epsilon < Z\gamma)$. Also,

$$V(Y_1 \mid I=1) = \frac{\rho^2\sigma_u^2}{\sigma_\epsilon^2} V(\epsilon \mid \epsilon < Z\gamma) + \sigma_u^2(1-\rho^2) \tag{9.22}$$

In the usual selectivity model we have been discussing, $\mu_\epsilon = 0$ and $\sigma_\epsilon = 1$. Further, $\epsilon$ is assumed to be normal, so that

$$g(Z\gamma) = -\frac{\phi(Z\gamma)}{\Phi(Z\gamma)} = \lambda \quad \text{(say)} \tag{9.23}$$

and $V(\epsilon \mid \epsilon < Z\gamma) = 1 - \lambda(Z\gamma - \lambda)$. Hence, equations (9.21) and (9.22) become

$$E(Y_1 \mid I=1) = X\beta + \rho\sigma_u\lambda \tag{9.24}$$

$$V(Y_1 \mid I=1) = \sigma_u^2[1 - \rho^2\lambda(Z\gamma - \lambda)] \tag{9.25}$$

Note that in the derivation of (9.24) and (9.25) we have not made any assumption about the distribution of $u$. The only assumptions made are that $\epsilon$ is normal and that the conditional expectation of $u$, given $\epsilon$, is linear, as given in equation (9.20). If $u$ and $\epsilon$ are bivariate normal, this condition follows automatically.

One question we might ask is how well the expression $\lambda$, defined in (9.23) under the assumption of normality, approximates the true truncated mean $g(Z\gamma)$ if $\epsilon$ is not normal. Goldberger (1980*a*) made some calculations with alternative error distributions and showed that the normal selection-bias adjustment is quite sensitive to departures from normality. This suggests that one should use a more general functional form for the truncated mean function in practice.

Before we move on to the generalized functional forms, we should note two points about the selectivity-bias adjustment:

1. One problem that has often been pointed out is that if $Z$ includes some variables in $X$ (or variables highly correlated with those in $X$), then given that the function $g(Z\gamma)$ is a nonlinear function of $Z$, it is likely to pick up any nonlinear terms omitted in equation (9.18), and the variable $g(Z\gamma)$ could be significant, thus indicating the presence of selection bias, even when there is no selection bias. The solution to this problem is to include nonlinear terms in (9.18), if that does indeed make

economic sense, and then examine whether or not the variable $g(Z\gamma)$ is significant.

2. Another problem occurs when $(Z\gamma)$ is constant. In this case the two-stage method described in Chapter 8 breaks down if there is a constant term in (9.18), because $g(Z\gamma)$ is constant for all observations.[4] However, the ML estimator does exist if we make the assumption that $u$ and $\epsilon$ are jointly normally distributed, and thus one can test for the presence of selectivity bias.

Returning to the question of nonnormal distributions, Olsen (1982*a*) suggested that the distribution of $\epsilon$ in (9.19) be assumed to be generated from a bivariate normal $(\epsilon, v)$, with $v$ truncated so that $v \leqslant K$, a given constant. (Here $v$ is another variate introduced to produce nonnormality.) By varying $K$, we get a variety of skewed distributions. Specifically, what he suggested is to consider the distribution of $(u, \epsilon, v)$ to be trivariate normal, with correlation matrix

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1\rho_2 \\ \rho_2 & \rho_1\rho_2 & 1 \end{bmatrix}$$

The extent and direction of selection are governed by $\rho_1$. The parameters $\rho_2$ and $K$ allow for nonnormality. The model falls in the category of multiple criteria of selectivity, because what it implies is that $Y_1$ is observed if and only if $\epsilon \leqslant Z\gamma$ and $v \leqslant K$. We shall be discussing these models in the next section. Because $K$ is a constant, we have to use the ML method of estimation, and thus will involve evaluation of double integrals.

In another study, Olsen (1980*a*) suggested the use of uniform distribution for $\epsilon$ in the range (0, 1). In this case, $E(\epsilon) = 1/2$, and $v(\epsilon) = 1/12$. Hence, using equation (9.20), we get

$$E(u \mid \epsilon < Z\gamma) = E\left[\frac{\rho\sigma_u}{\rho_\epsilon}\left(\epsilon - \frac{1}{2}\right)\middle| \epsilon < Z\gamma\right]$$

$$= \frac{\rho\sigma_u}{\sigma_\epsilon}\left(\frac{Z\gamma}{2} - \frac{1}{2}\right) = \rho\sigma_u(3)^{1/2}(Z\gamma - 1)$$

Thus,

$$E(Y_1 \mid I = 1) = X\beta + \rho\sigma_u(3)^{1/2}(Z\gamma - 1) \tag{9.26}$$

[4] There are some cases that the two-stage method would break down when $Z\gamma$ is a combination of dichotomous variables (not just the case in which $Z\gamma$ is a constant).

Also,

$$V(Y_1 \mid I=1) = \sigma_u^2(1-\rho^2) + \rho^2\sigma_u^2(Z\gamma)^2 \tag{9.27}$$

Given this specification of $\epsilon$, the preliminary consistent estimates of $Z\gamma$ can be obtained from the linear probability model, and then the model derived from equation (9.26) that is estimated by ordinary least squares after substituting $Z\hat{\gamma}$ for $Z\gamma$ is

$$Y_1 = X\beta + \alpha(Z\hat{\gamma} - 1) + v$$

where

$$\alpha = \rho\sigma_u(3)^{1/2}$$

and

$$v = -\alpha Z(\hat{\gamma} - \gamma) + \eta$$

Olsen derived the asymptotic covariance matrix for these two-stage estimates of $\beta$ and $\alpha$.

One important distinction between the corrections for selectivity bias based on the linear probability model and the probit model is that in the probit model the function $g(Z\gamma)$ in (9.23) is a nonlinear function of $Z$, whereas in the linear probability model it is a linear function of $Z$. Hence, for the selectivity-bias adjustment when the probit model is used, we can have $Z$ contain the same variables as in $X$ and not cause any problems of identification, but when we use the linear probability model we cannot have the same variables as in $X$. If we believe that the same variables are important in both equations (9.18) and (9.19), we have to use nonlinear functions of these variables in $Z$. On the other hand, if $X$ contains nonlinear functions of these variables, we shall have some problems of identification even if the probit method is used. Thus, in a practical sense, problems of identification will arise even if the probit model is used for the selectivity-bias adjustment. In the empirical illustration that Olsen (1980*a*) used, the two models (probit and linear probability) gave similar results.

Finally, there are the implications about the distributions of $u$ and $\epsilon$ in equations (9.18) and (9.19). Assuming $\epsilon$ to be uniform, and assuming that the conditional expectation of $u$, given $\epsilon$, is linear in $\epsilon$, implies some outrageous assumptions about the distribution of $u$. If $u$ is assumed to be a convolution of a uniform density and a normal density, the distribution of $u$ will be symmetric, but with a broader peak and narrower tails. If $|\rho| < 0.5$, this distribution does not differ much from the normal, and in the extreme case $|\rho| = 1$ the distribution is uniform, which is a very unlikely distribution for a regression model.

## 9.4 Some general transformations to normality

In the preceding section we considered some particular alternatives to the probit method of correcting for selectivity bias. We shall now discuss some general transformations to normality suggested by Lee (1982*c*, in press).

Consider the model given in equations (9.18) and (9.19). Let $G(u)$ and $F(\epsilon)$ be the distribution functions of $u$ and $\epsilon$. Let $\Phi(\cdot)$ be the distribution function of the standard normal, and let $B(\cdot,\cdot;\rho)$ be the bivariate normal distribution, with zero means, unit variances, and correlation coefficient $\rho$. Because the distributions of $\epsilon$ and $u$ are specified, each of them can be transformed to a standard normal random variable $N(0,1)$. Let

$$\epsilon^* = J_1(\epsilon) \equiv \Phi^{-1}[F(\epsilon)] \tag{9.28}$$

$$u^* = J_2(u) \equiv \Phi^{-1}[G(u)] \tag{9.29}$$

Then $\epsilon^*$ and $u^*$ have $N(0,1)$ distributions. The transformations $J_1$ and $J_2$ involve the inverse of the standard normal distribution function. Computationally simple and accurate methods involving the use of approximation functions for this can be found in Appendix IIC of Bock and Jones (1968) and in the work of Hildebrand (1956). Errors of approximation for these methods are less than $3\times10^{-4}$.

A bivariate distribution having the marginal distributions $F(\epsilon)$ and $G(u)$ can be specified as

$$H(\epsilon,u;\rho) = B[J_1(\epsilon),J_2(u);\rho] \tag{9.30}$$

If $f(\epsilon)$ and $g(u)$ are the marginal density functions of $\epsilon$ and $u$, respectively, then the joint density function of $\epsilon$ and $u$ corresponding to the distribution function (9.30) is

$$h(\epsilon,u;\rho) = (1-\rho^2)^{-1/2}f(\epsilon)g(u) \times \exp(-\rho[2(1-\rho^2)]^{-1}\{\rho[J_1^2(\epsilon)+J_2^2(u)]-2J_1(\epsilon)J_2(u)\}) \tag{9.31}$$

When the marginal distributions of $u$ and $\epsilon$ are normally distributed, the foregoing bivariate distribution will be a bivariate normal distribution.

With this specification, one can easily derive the likelihood function for the censored regression model in (9.18) and (9.19). Let us denote the observations on $Y_1$ for $I=1$ by $Y$. Then, for this group, we have $\epsilon<Z\gamma$ and $u=Y-X\beta$. But

$$\int_{-\infty}^{Z\gamma} h(\epsilon,Y-X\beta) = \frac{\partial}{\partial u}H(\epsilon,u;\rho)\Big|_{\substack{u=Y-X\beta\\ \epsilon=Z\gamma}}$$

$$= \frac{\partial B[J_1(\epsilon), J_2(u); \rho]}{\partial J_2(u)} \cdot \frac{g(u)}{\phi[J_2(u)]}\Bigg|_{\substack{u = Y - X\beta \\ \epsilon = Z\gamma}} \tag{9.32}$$

where $\phi(\cdot)$ is the standard normal density function. Because

$$\frac{\partial B(t, s; \rho)}{\partial s} \frac{1}{\phi(s)} = \Phi\left(\frac{t - \rho s}{(1-\rho^2)^{1/2}}\right)$$

the expression in (9.32) simplifies to

$$\Phi\left(\frac{J_1 Z(\gamma) - \rho J_2(Y - X\beta)}{(1-\rho^2)^{1/2}}\right) \cdot g(Y - X\beta) \tag{9.33}$$

The log-likelihood function can therefore be written as

$$\log L(\beta, \gamma, \theta_1, \theta_2, \rho)$$
$$= \sum_{i=1}^{N} \left\{ I_i \log g(Y_i - X_i\beta) + I_i \log \Phi\left(\frac{J_1(Z_i\gamma) - \rho J_2(Y_{1i} - X_i\beta)}{(1-\rho^2)^{1/2}}\right) \right.$$
$$\left. + (1 - I_i) \log[1 - F(Z_i\gamma)] \right\} \tag{9.34}$$

where $\theta_1$ and $\theta_2$ are the unknown parameters in $F(\epsilon)$ and $G(u)$, respectively. The maximum-likelihood method can be applied to this likelihood function.

One can also use two-stage estimation methods to obtain initial consistent estimates. Specifically, assume that $u$ are $N(0, \sigma^2)$, whereas the distribution of $\epsilon$ is arbitrary. In this case, $J_2(u) = u/\sigma$, and $g(u) = (1/\sigma)\phi(u/\sigma)$. For the ML estimation, we make these substitutions in (9.34). For two-stage estimation, note that

$$I = 1 \Leftrightarrow \epsilon < Z\gamma$$
$$\Leftrightarrow J_1(\epsilon) < J_1(Z\gamma) \tag{9.35}$$

Also, $\text{Prob}(I=1) = \Phi[J_1(Z\gamma)] = F(Z\gamma)$. The usual methods of two-stage estimation (discussed in Chapter 8) apply by substituting $J_1(Z\gamma)$ for $Z\gamma$. Thus, conditional on $I=1$, we can write

$$Y = X\beta = \sigma\rho\phi[J_1(Z\gamma)]/F(Z\gamma) + \eta \tag{9.36}$$

where $E(\eta \mid I=1, X, Z) = 0$ and

$$\text{Var}(\eta \mid I=1, X, Z) = \sigma^2 - (\sigma\rho)^2\{J_1(Z\gamma) + \phi[J_1(Z\gamma)]/F(Z\gamma)\}$$
$$\times \phi[J_1(Z\gamma)]/F(Z\gamma)$$

We have substituted $F(Z\gamma)$ for $\Phi[J_1(Z\gamma)]$. In the two-stage method, we first estimate $\gamma$ by maximizing the likelihood function

$$\log L_1(\gamma) = \sum_{i=1}^{N} \{I_i \log F(Z_i\gamma) + (1 - I_i) \log[1 - F(Z_i\gamma)]\}$$

Let $\hat{\gamma}$ be the ML estimate of $\gamma$. We then substitute $\hat{\gamma}$ for $\gamma$ in (9.36) and estimate the equation by ordinary least squares. The asymptotic covariance matrix of the two-stage estimates are the same as those presented in Chapter 8, with $J_1(Z\gamma)$ replacing $Z\gamma$ throughout. Because this can be derived easily, it will not be presented here.

If $\epsilon$ follows the normal distribution, then we have the usual probit method of correction for selectivity bias. If $F(\epsilon)$ is the logistic distribution, then we have the logit method of correction for selectivity bias. In this case the estimation of (9.36) gives us the logit two-stage estimates. The methods that can be used to estimate (9.36) are thus very general. However, as explained in the preceding section, the two-stage estimation method does not depend on the assumption of normality of $u$. The only assumption needed is linearity of the conditional-expectation function, as in equation (9.20). A general class of dependence models suggested by Lee (1982*c*) that can be used in connection with equations (9.21) and (9.22) is the following: Let $J$ be a specified strictly increasing transformation, so that

$$\epsilon < Z\gamma \Leftrightarrow J(\epsilon) < J(Z\gamma)$$

Also let

$$\mu_J = D[J(\epsilon)]$$
$$\sigma_J^2 = V[J(\epsilon)]$$

A general specification for the distribution of $u$ is

$$u = \lambda[J(\epsilon) - \mu_J] + v$$

where $v$ and $J(\epsilon)$ are independent. If $\lambda = 0$, then $\epsilon$ and $u$ will be uncorrelated. If we write

$$M_1 = E[J(\epsilon) \mid J(\epsilon) < J(Z\gamma)]$$
$$M_2 = E[J^2(\epsilon) \mid J(\epsilon) < J(Z\gamma)]$$

then equations (9.21) and (9.22) can be written as

$$Y = X\beta + \frac{\rho\sigma_u}{\sigma_J}\left(\frac{M_1}{F(Z\gamma)} - \mu_J\right) + \eta \tag{9.37}$$

where

$$E(\eta \mid X, Z, I = 1) = 0$$

$$V(\eta \mid X, Z, I = 1) = \frac{\rho^2\sigma_u^2}{\sigma_J^2}\left[\frac{M_2}{F(Z\gamma)} - \left(\frac{M_1}{F(Z\gamma)}\right)^2\right] + \sigma_u^2(1 - \rho^2)$$

The transformation to normality given by equation (9.28) is one convenient candidate for $J$. The equation (9.37) is more general and can be used if some other transformations are considered.

The foregoing discussion shows that correction for selectivity bias and estimation of models with selectivity can be done with very general error distributions. The additional computational burden involved is that of computing the transformations in (9.28) and (9.29).

## 9.5 Polychotomous-choice models and selectivity bias

Throughout the preceding discussion we considered the case of two choices and a potential regression equation in each category. We shall now consider generalizations of this to multiple choices. An illustration of the multiple-choice problem was provided by Hay (1980). The example considered by Hay involved simultaneous estimation of specialty choice and specialty income for physicians. He considered a model with three alternatives: GP (general or family practice), IM (internal medicine), and OT (all other specialties).

There are two approaches to the analysis of polychotomous-choice models with mixed continuous and discrete data.[5] The first approach is to formulate them as models with multiple binary-choice rules and partial observations. The second approach depends on order statistics for polychotomous-choice models. We shall now elaborate both these approaches.

Consider the following polychotomous-choice model, with $M$ categories and one potential regression outcome in each category:

$$y_{si} = x_{si}\beta_s + u_{si} \quad (s = 1, 2, \ldots, M)$$
$$I_{si}^* = z_{si}\gamma + \eta_{si} \quad (i = 1, 2, \ldots, N)$$

The subscript $i$ refers to the $i$th observation; $x_s$ and $z_s$ are exogenous variables; $E(u_s \mid x_s, z_s) = 0$; $y_s$ is observed only if the $s$th category is chosen. Let $I$ be a polychotomous variable with values 1 to $M$ and $I = s$ if the $s$th category is chosen.

$$I = s \quad \text{iff} \quad z_s\gamma - z_j\gamma > \eta_j - \eta_s \quad \text{for all } j = 1, 2, \ldots, M \quad (j \neq s) \tag{9.38}$$

This formulation relates the polychotomous-choice model as a model with $M-1$ binary-decision rules (9.38) with partial observations. This is the approach followed by Hay (1980) and Dubin and McFadden (1980).

In the second formulation we write

$$I = s \quad \text{iff} \quad I_s^* > \operatorname{Max} I_j^* \quad (j = 1, 2, \ldots, M, \ j \neq s) \tag{9.39}$$

Let

$$\epsilon_s = \operatorname{Max} I_j^* - \eta_s \quad (j = 1, 2, \ldots, M, \ j \neq s) \tag{9.40}$$

It follows that

[5] The subsequent discussion here is based on the work of Lee (1982*c*).

$$I = s \quad \text{iff} \quad \epsilon_s < z_s\gamma \tag{9.41}$$

This second approach leads to tractable results if the distribution function of $\epsilon_s$ can be specified. For example, suppose that $\eta_j$ $(j=1,2,\ldots,M)$ are independently and identically distributed, with the type I extreme-value distribution with cumulative distribution function.

$$F(\eta_i < c) = \exp[-\exp(-c)]$$

Then, as shown in section 3.1, or as shown by Domencich and McFadden (1975),

$$\text{Prob}(\epsilon_s < z_s\gamma) = \text{Prob}(I = s) = \frac{\exp(z_s\gamma)}{\sum_j \exp(z_j\gamma)}$$

Thus, the distribution function of $\epsilon_s$ is given by

$$F_s(\epsilon) = \text{Prob}(\epsilon_s < \epsilon) = \frac{\exp(\epsilon)}{\exp(\epsilon) + \sum_{\substack{j=1,2,\ldots,M \\ j \neq s}} \exp(z_j\gamma)} \tag{9.42}$$

Thus, what we have is that for each choice $s$ we now have the model

$$y_s = x_s\beta_s + u_s$$

where $y_s$ is observed if and only if $\epsilon_s < z_s\gamma$. The distribution function of $\epsilon_s$ is given by (9.42). The estimation is now exactly the same as in the binary-choice model discussed in the preceding section. We consider a transformation as in (9.28):

$$\epsilon_s^* = J_s(\epsilon_s) = \Phi^{-1}[F_s(\epsilon)]$$

The condition $\epsilon_s < z_s\gamma \Leftrightarrow \epsilon_s^* < J_s(z_s\gamma)$, and we estimate a model like (9.36) by the two-stage method. We estimate the equation

$$y_s = x_s\beta_s - \sigma_s\rho_s\phi[J_s(z_s\gamma)]/F_s(z_s\gamma) + v_s \tag{9.43}$$

by ordinary least squares after substituting $\hat{\gamma}$ for $\gamma$; $\sigma_s^2 = \text{Var}(u_s)$, and $\rho_s$ is the correlation coefficient between $u_s$ and $\epsilon_s^*$.

The only difference between this estimation of the polychotomous-choice model and the estimation of the binary-choice model considered in the preceding section is that the preliminary estimate of $\gamma$ is obtained by the conditional logit model (described in Chapter 3).

Returning to the first approach, as followed by Hay (1980) and given by equation (9.38), let us define

$$\omega_{sj} = \eta_j - \eta_s \quad \text{and} \quad t_{sj} = z_j\gamma - z_s\gamma \tag{9.44}$$

so that the condition (9.38) becomes $\omega_{sj} < t_{sj}$. If we assume, as before, that $\eta_j$ are independently and identically distributed, with the type I

extreme-value distribution, then the $M-1$ random variables $\omega_{sj}$ will have the multivariate logistic distribution of Gumbel (1961). The joint distribution is[6]

$$F(\omega_{s1}, \omega_{s2}, \ldots, \omega_{s,s-1}, \omega_{s,s+1}, \ldots, \omega_{sm}) = \left[1 + \sum_{\substack{j=1,2,\ldots,m \\ j \neq s}} \exp(-\omega_{sj})\right]^{-1} \tag{9.45}$$

If $l' = (1, 1, \ldots, 1)$ is an $M-1$ vector with all 1's, the covariance matrix of the $M-1$ variables $\omega_{sj}$ is

$$\Sigma_\omega = \frac{\pi^2}{6}(I + ll')$$

Consider now the two-stage estimation method. For simplicity of notation, let us consider choice 1. We have the equations

$$y_1 = x_1\beta_1 + u_1$$

and $y_1$ is observed if and only if $\omega_{1j} \leq t_{1j}$ $(j = 2, 3, \ldots, M)$, where $\omega_{1j}$ follow the distribution (9.45). Denoting the vector $(\omega_{12}, \omega_{13}, \ldots, \omega_{1M})'$ by $\omega_1$ we next write, as in section 9.3,

$$\begin{aligned} u_1 &= \text{Cov}(u_1, \omega_1)[\text{Var}(\omega_1)]^{-1}[\omega_1 - E(\omega_1)] + v_1 \\ &= \sum_{j=2}^{M} \lambda_j \omega_{ij} + v_1 \end{aligned} \tag{9.46}$$

where $E(v_1 \mid \omega_1) = 0$ and $E(\omega_1) = 0$. Hence,

$$E(y_1 \mid \omega_{1j} < t_{1j}) = x_1\beta_1 + \sum_{j=2}^{M} \lambda_j E(\omega_{1j} \mid \omega_{1k} < t_{1k}) \qquad (j = 2, 3, \ldots, M,\ k = 2, 3, \ldots, M) \tag{9.47}$$

Once we evaluate the conditional expectation in (9.47), we can use the equation for two-stage estimation. For this, we use the following result:[7] If $v_1, v_2, \ldots, v_J$ have a multivariate logistic distribution

$$F(v_1, v_2, \ldots, v_J) = \left(1 + \sum_{j=1}^{J} e^{-v_j}\right)^{-1} \tag{9.48}$$

then

[6] Further details of this distribution can be found in Chapter 42 of Johnson and Kotz (1972).

[7] See the study by Lee (1982*c*), where $E(v_i)$, $E(v_i^2)$, and $E(v_i v_j)$ are presented. The variance and covariance terms are needed for deriving the correct asymptotic covariance matrices of the two-stage estimates.

$$E(v_1 \mid v_1 < x_1, v_2 < x_2, \ldots, v_J < x_J)$$
$$= [1 - e^{-x_1} F(x_1, x_2, \ldots, x_J)]^{-1}$$
$$\times [\log F(x_1, x_2, \ldots, x_J) - x_1 e^{-x_1} F(x_1, x_2, \ldots, x_J)] \tag{9.49}$$

Thus, (9.47) can be written as

$$E(y_1 \mid \omega_{1j} < t_{1j}) = x_1 \beta_1 + \sum_{j=2}^{M} \lambda_j [1 - e^{-t_{1j}} F(t_1)]^{-1} [\log F(t_1) - t_{1j} e^{-t_{1j}} F(t_1)]$$
$$(j = 2, 3, \ldots, M) \tag{9.50}$$

where $F(t_1) = F(t_{12}, t_{13}, \ldots, t_{1M})$ and $F(\cdot)$ is the multivariate logistic distribution (9.48). Note that $t_{ij}$ are functions of $z_j \gamma$. Thus, we first estimate $\gamma$ using the conditional logit model. We substitute the estimate $\hat{\gamma}$ of $\gamma$ in $t_{1j}$, calculate the values of the variables with coefficients $\lambda_j$ in (9.50), and estimate this equation by ordinary least squares to get estimates of $\beta_1$ and $\lambda_j$ $(j = 2, 3, \ldots, N)$. This procedure is repeated for each of the variables $y_s$ $(s = 1, 2, \ldots, M)$.

The expressions for the asymptotic covariance matrices of the two-stage estimates are very complicated and will not be presented here.[8] Clearly, this approach is more cumbersome than the alternative approach based on equations (9.41) through (9.43).

## 9.6 Multiple criteria for selectivity

There are several practical instances in which selectivity can be due to several sources, rather than just one, as considered in the examples in the preceding section. Griliches et al. (1978) cited several problems with the NLS data on young men that could lead to selectivity bias. Prominent among these are attrition and other missing-data problems. In such cases we need to formulate the model as a switching regression model or a switching simultaneous-equations model, where the switch depends on more than one criterion function. One such example is that by Abowd and Farber (1982), who considered the union-and-wages example of Lee (1978). The model consists of a union-wage equation ($Y_1$) and a nonunion-wage equation ($Y_2$). There are two decision functions: the decision of individuals to join a queue for union jobs ($I_1^*$) and the deci-

[8] Dubin and McFadden (1980) derived the covariance matrix for the case of $M$ alternatives and corrected some slips in Hay's (1980) calculation. Lee (1982*c*) gave expressions for the second moments of the truncated multivariate logistic distribution.

sion of employers to draw individuals from the queue ($I_2^*$). The specification of the model is

$$Y_1 = X_1\beta_1 + u_1 \tag{9.51}$$

$$Y_2 = X_2\beta_2 + u_2 \tag{9.52}$$

$$I_1^* = Z_1\gamma_1 - \epsilon_1 \tag{9.53}$$

$$I_2^* = Z_2\gamma_2 - \epsilon_2 \tag{9.54}$$

If $I_1^* > 0$, the individual decides to join the queue for union jobs. If $I_2^* > 0$, the individual is chosen from the queue for a union job. Here we observe $Y_1$ only if $I_1^* > 0$ and $I_2^* > 0$. In this example, the set $I_1^* < 0$ and $I_2^* > 0$ will be empty.

When we talk of multiple criteria for selectivity, we should distinguish two cases: the joint case and the sequential case. In the joint-decision model, (9.53) and (9.54) are defined over the entire set of observations. In the sequential-decision model, (9.54) is defined over only the subset of observations for which $I_1^* > 0$. In this example, the choice of drawing from the queue arises only for those who are in the queue.

We also have to consider whether the choices are completely observed or partially observed. Define the indicator variables

$$I_1 = 1 \quad \text{iff } I_1^* > 0$$

$$I_1 = 0 \quad \text{otherwise}$$

$$I_2 = 1 \quad \text{iff } I_2^* > 0$$

$$I_2 = 0 \quad \text{otherwise}$$

The question is whether we observe $I_1$ and $I_2$ separately or only as a single indicator variable $I = I_1 I_2$. The latter is the case with the example of Abowd and Farber. Poirier (1980) also considered a bivariate probit model with partial observability, but his model was a joint model, not a sequential model as in the example of Abowd and Farber. An example of a joint model is that of estimating the probability that an on-the-job trainee will be retained by the sponsoring agency after training. In this situation the employer must decide whether or not to make a job offer, and the applicant must decide whether or not to seek a job offer. We do not observe these individual decisions. What we observe is whether or not the trainee continues to work after training. If either the employer or the employee makes his decision first, then the model will be a sequential model.

Tunali et al. (1980) also considered a sequential-decision model, given by (9.51), (9.53), and (9.54). Here, $y_1$ is observed only if $I_1 = 1$ and $I_2 = 1$. However, in their model, both $I_1$ and $I_2$ are observed. Their example was

one of labor-force participation by women in Managua, Nicaragua. Of the 1,247 women in the sample, only 579 were labor-force participants. Of these, only 525 reported earnings. The first decision is whether or not to participate in the labor force, and the second decision is whether or not to report earnings.

In the joint-decision model with partial observability (i.e., where we observe $I = I_1 \cdot I_2$ only, not $I_1$ and $I_2$ individually), the parameters $\gamma_1$ and $\gamma_2$ in equations (9.53) and (9.54) are estimable only if there is at least one nonoverlapping variable in either one of $Z_1$ and $Z_2$. Because $V(\epsilon_1) = V(\epsilon_2) = 1$, by normalization, let us define $\text{Cov}(\epsilon_1, \epsilon_2) = \rho$. Also, write

$$\begin{aligned}\text{Prob}(I_1^* > 0,\ I_2^* > 0) &= \text{Prob}(\epsilon_1 < Z_1\gamma_1, \epsilon_2 < Z_2\gamma_2) \\ &= F(Z_1\gamma_1, Z_2\gamma_2, \rho)\end{aligned}$$

Then the ML estimates of $\gamma_1$, $\gamma_2$ and $\rho$ are obtained by maximizing the likelihood function

$$L_1 = \prod_{I=1} F(Z_1\gamma_1, Z_2\gamma_2, \rho) \cdot \prod_{I=0} [1 - F(Z_1\gamma_1, Z_2\gamma_2, \rho)] \qquad (9.55)$$

With the assumption of bivariate normality of $\epsilon_1$ and $\epsilon_2$, this involves the use of bivariate probit analysis.

In the sequential-decision model with partial observability, if we assume that the function (9.54) is defined only on the subpopulation $I_1 = 1$, then, because the distribution of $\epsilon_2$ that is assumed is considered on $\epsilon_1 < Z_1\gamma_1$, the likelihood function to be maximized will be

$$L_2 = \prod_{I=1} [\Phi(Z_1\gamma_1)\Phi(Z_2\gamma_2)] \cdot \prod_{I=0} [1 - \Phi(Z_1\gamma_1)\Phi(Z_2\gamma_2)] \qquad (9.56)$$

Again, the parameters $\gamma_1$ and $\gamma_2$ are estimable only if there is at least one nonoverlapping variable in either one of $Z_1$ and $Z_2$ (otherwise, we would not know which estimates refer to $\gamma_1$ and which refer to $\gamma_2$). In their example on job queues and union status of workers, Abowd and Farber (1982) obtained their parameter estimates using the likelihood function (9.56). One can, perhaps, argue that even in the sequential model the appropriate likelihood function is still (9.55), not (9.56). It is possible that there are persons who do not join the queue ($I_1 = 0$) but to whom employers would want to offer union jobs. The reason we do not observe these individuals in union jobs is because they decided not to join the queue. But we also do not observe in the union jobs all those with $I_2 = 0$. Thus, we can argue that $I_2^*$ exists and is, in principle, defined even for the observations $I_1 = 0$. If the purpose of the analysis is to examine what factors influence an employer's choice of employees for union jobs, then

possibly the parameter estimates should be obtained from (9.55). The difference between the two models is in the definition of the distribution of $\epsilon_2$. In the case of (9.55), the distribution of $\epsilon_2$ is defined over the whole population. In the case of (9.56), it is defined over the subpopulation $I_1 = 1$. The latter allows us to make only conditional inferences.[9] The former allows us to make both conditional and marginal inferences. To make marginal inferences, we need estimates of $\gamma_2$. To make conditional inferences, we consider the conditional distribution $f(\epsilon_2 \mid \epsilon_1 < Z_1\gamma_1)$, which involves $\gamma_1$, $\gamma_2$, and $\rho$. We shall discuss this issue of marginal versus conditional inferences in the next section.

Yet another type of partial observability arises in the case of truncated samples discussed earlier in section 9.2. An example is that of measuring discrimination in loan markets. Let $I_1^*$ refer to the decision of an individual whether or not to apply for a loan, and let $I_2^*$ refer to the decision of the bank whether or not to grant the loan.

$I_1 = 1$ if the individual applies for a loan
$I_1 = 0$ otherwise
$I_2 = 1$ if the applicant is given a loan
$I_2 = 0$ otherwise

Rarely do we have data on the individuals for whom $I_1 = 0$. Thus, what we have is a truncated sample. We can, of course, specify the distribution of $I_2^*$ only for the subset of observations $I_1 = 1$ and estimate the parameters $\gamma_2$ by, say, the probit ML method and then examine the significance of the coefficients of race, sex, age, and so forth to see if there is discrimination by any of these variables. This does not, however, allow for self-selection at the application stage, say for some individuals not applying because they feel they will be discriminated against. For this purpose, we define $I_2^*$ over the whole population and analyze the model from the truncated sample. The argument is that, in principle, $I_2^*$ exists even for the nonapplicants. The parameters $\gamma_1$, $\gamma_2$, and $\rho$ can be estimated by maximizing the likelihood function

$$L_3 = \prod_{I_2=1} \frac{F(Z_1\gamma_1, Z_2\gamma_2, \rho)}{\Phi(Z_1\gamma_1)} \cdot \prod_{I_2=0} \frac{\Phi(Z_1\gamma_1) - F(Z_1\gamma_1, Z_2\gamma_2, \rho)}{\Phi(Z_1\gamma_1)} \tag{9.57}$$

In this model, the parameters $\gamma_1$, $\gamma_2$, and $\rho$ are, in principle, estimable

[9] The conditional model does not permit us to allow for the fact that changes in $Z_2$ also might affect the probability of being in the queue. Also, the decision whether or not to join the queue can be influenced by the perception of the probability of being drawn from the queue.

even if $Z_1$ and $Z_2$ are the same variables. In practice, however, the estimates are not likely to be very good.[10]

Fishe et al. (1981) considered a two-decision model, but it is a model of joint decisions, and with both $I_1$ and $I_2$ observed. The model is one that determines wages of young women, some of whom have college education and some of whom do not. The two decision equations (9.53) and (9.54) refer to the decisions whether or not to go to college and whether or not to join the labor force.

The analysis of the model in equations (9.51) and (9.54) will depend crucially on whether the two decisions are independent or correlated, that is, whether or not $\text{Cov}(\epsilon_1, \epsilon_2) = 0$. In the case $\text{Cov}(\epsilon_1, \epsilon_2) = 0$, we can easily extend the Heckman–Lee two-stage estimation methods to this model. We define

$$\lambda_{ij} = \text{Cov}(u_i, \epsilon_j) \quad (i=1,2,\ j=1,2)$$

Then,

$$E(u_1 \mid I_1^* > 0,\ I_2^* > 0) = -\lambda_{11} \frac{\phi(Z_1\gamma_1)}{\Phi(Z_1\gamma_1)} - \lambda_{12} \frac{\phi(Z_2\gamma_2)}{\Phi(Z_2\gamma_2)} \tag{9.58}$$

Thus, we get preliminary consistent estimates of $\gamma_1$ and $\gamma_2$ by estimating equations (9.53) and (9.54) by the probit method. Next, we regress $Y_1$ on $X_1$ and the constructed variables

$$\frac{\phi(Z_1\hat{\gamma}_1)}{\Phi(Z_1\hat{\gamma}_1)} \quad \text{and} \quad \frac{\phi(Z_2\hat{\gamma}_2)}{\Phi(Z_2\hat{\gamma}_2)}$$

In case $\epsilon_1$ and $\epsilon_2$ are correlated, so that $\text{Cov}(\epsilon_1, \epsilon_2) = \sigma_{12}$, the expressions get very messy. In this case we have to use bivariate probit methods to estimate $\gamma_1$, $\gamma_2$, and $\sigma_{12}$. Further,

$$E(u_1 \mid I_1^* > 0,\ I_2^* > 0) = \lambda_{11} M_{12} + \lambda_{12} M_{21}$$

where

$$M_{ij} = (1 - \sigma_{12}^2)^{-1}(P_i - \sigma_{12} P_j)$$

$$P_j = \frac{\int_{-\infty}^{Z_1\gamma_1} \int_{-\infty}^{Z_1\gamma_1} \epsilon_j f(\epsilon_1 \epsilon_2)\, d\epsilon_2\, d\epsilon_1}{F(Z_1\gamma_1, Z_2\gamma_2)} \tag{9.59}$$

[10] See the evidence presented in section 9.2 on estimation of the parameters in the selectivity criterion from truncated samples. See also the work of Bloom et al. (1981), who reported that attempts at estimating this model did not produce good parameter estimates.

These expressions can still be evaluated numerically.[11]

Fishe et al. estimated the parameters in equatiions (9.53) and (9.54) by the bivariate probit method and evaluated expressions of the form (9.59) by numerical methods. They then used the extension of the Heckman–Lee two-stage method.

## 9.7 Endogenous switching models and mixture-distribution models

The models of self-selection discussed in this chapter (as well as the disequilibrium models discussed in the next chapter) fall in the general class of switching models with endogenous switching (Maddala and Nelson, 1975). In a recent study, Poirier and Rudd (1981) argued that there has been substantial confusion in the econometric literature over switching regression models with endogenous switching and that this confusion can cause serious interpretation problems when the model is employed in applied work. They argued that the problems of interpretation arise because there is an observational equivalence between two fundamentally different specifications: the mixture model of conditional densities and the switching regression model with endogenous switching. Because their study can convey misleading impressions about the practical usefulness of the models discussed in this chapter, we shall discuss the two models here.

The switching regression model with endogenous switching is defined as follows:

$$y_{1i} = X_{1i}\beta_1 + u_{1i} \tag{9.60}$$

$$y_{2i} = X_{2i}\beta_2 + u_{2i} \tag{9.61}$$

$$I_i^* = Z_i\gamma - \epsilon_i \tag{9.62}$$

$$\begin{aligned} I_i &= 1 \quad \text{iff } I_i^* > 0 \\ I_i &= 0 \quad \text{iff } I_i^* \leqslant 0 \end{aligned} \tag{9.63}$$

The observed $y_i$ is defined as

$$\begin{aligned} y_i &= y_{1i} \quad \text{iff } I_i = 1 \\ y_i &= y_{2i} \quad \text{iff } I_i = 0 \\ (u_1, u_2, \epsilon)' &\sim N(0, \Sigma) \end{aligned} \tag{9.64}$$

[11] See the work of Rosenbaum (1961) for moments of a truncated bivariate normal distribution. These are also reported in the Appendix at the end of the book.

with

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1\epsilon} \\ \sigma_{12} & \sigma_{22} & \sigma_{2\epsilon} \\ \sigma_{1\epsilon} & \sigma_{2\epsilon} & 1 \end{bmatrix}$$

If $\sigma_{1\epsilon} = \sigma_{2\epsilon} = 0$, we have the switching regression model with exogenous switching. Otherwise, we have endogenous switching.

Equations (9.60) and (9.61) define the marginal distributions of $y_{1i}$ and $y_{2i}$.[12] From the specification of the model we can derive the conditional distributions $f(y_{1i} \mid I_i = 1)$ and $f(y_{2i} \mid I_i = 0)$. For instance,

$$f(u_1 \mid I = 1) = \int_{-\infty}^{Z\gamma} f(u_1 \epsilon)\, d\epsilon / \Phi(Z\gamma)$$

and writing $f(u_1 \epsilon) = f(u) \cdot f(\epsilon \mid u)$, we can write

$$f(y_1 \mid I = 1) = [\Phi(Z\gamma)]^{-1} \sigma_{11}^{-1/2} \phi[\sigma_{11}^{-1/2}(y_1 - X_1\beta_1)] \times \Phi\left\{\left(1 - \frac{\sigma_{1\epsilon}^2}{\sigma_{11}}\right)^{-1/2} \left[Z\gamma - \frac{\sigma_{1\epsilon}}{\sigma_{11}}(y_1 - X_1\beta_1)\right]\right\}$$

Similarly,

$$f(y_2 \mid I = 0) = [1 - \Phi(Z\gamma)]^{-1} \sigma_{22}^{-1/2} \phi[\sigma_{22}^{-1/2}(y_2 - X_2\beta_2)] \times \left(1 - \Phi\left\{\left(1 - \frac{\sigma_{2\epsilon}^2}{\sigma_{22}}\right)^{-1/2} \left[Z\gamma + \frac{\sigma_{2\epsilon}}{\sigma_{22}}(y_2 - X_2\beta_2)\right]\right\}\right) \tag{9.65}$$

We can decompose the error terms $u_{1i}$, $u_{2i}$, and $\epsilon_i$ into a set of correlated and noncorrelated components. We can write

$$\begin{aligned} y_{1i} &= X_{1i}\beta_1 + \alpha_{1i} + w_{1i} \\ y_{2i} &= X_{2i}\beta_2 + \alpha_{2i} + w_{2i} \\ I_i^* &= Z_i\gamma + \alpha_{3i} + w_{3i} \end{aligned} \tag{9.66}$$

where $w_i' = (w_{1i}, w_{2i}, w_{3i}) \sim N(0, \Omega)$ and $\alpha_i' = (\alpha_{1i}, \alpha_{2i}, \alpha_{3i}) \sim N(0, \Lambda)$ and $w_i'$ are independent of $\alpha_i'$. $\Omega$ is a diagonal matrix

[12] Poirier and Rudd claimed that some studies have defined (9.60) and (9.61) as conditional on $I_i = 1$ and $I_i = 0$, respectively. However, we need not go into this issue in detail, because those who read the studies carefully can see that equations (9.60) and (9.61) were always meant to be marginal distributions, with the observed $y_i$ being defined by (9.64).

$$\Omega = \begin{bmatrix} \omega_{11} & 0 & 0 \\ 0 & \omega_{22} & 0 \\ 0 & 0 & \omega_{33} \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{12} & \lambda_{22} & \lambda_{23} \\ \lambda_{13} & \lambda_{23} & \lambda_{33} \end{bmatrix}$$

This method is identical with that for the model in equations (9.60) through (9.62), with

$$\lambda_{11}+\omega_{11} = \sigma_{11}, \quad \lambda_{22}+\omega_{22} = \sigma_{22}, \quad \lambda_{33}+\omega_{33} = 1,$$
$$\lambda_{12} = \sigma_{12}, \quad \lambda_{13} = \sigma_{1\epsilon}, \quad \text{and} \quad \lambda_{23} = \sigma_{2\epsilon} \tag{9.67}$$

Note that the models given by (9.66) and (9.60) through (9.62) are not different models.[13] The $\alpha_i$ now capture the correlations between the residuals, and conditional on $\alpha_i$, the variables $y_{1i}$, $y_{2i}$, and $I_i^*$ are independent.

The mixture-distribution model that Poirier and Rudd considered is

$$\begin{aligned} (y_{1i} \mid I_i = 1, \alpha_i) &\sim N(X_{1i}\beta_1 + \alpha_{1i}, \omega_{11}) \\ (y_{2i} \mid I_i = 0, \alpha_i) &\sim N(X_{2i}\beta + \alpha_{2i}, \omega_{22}) \\ (I_i^* \mid \alpha_i) &\sim N(Z_i\gamma + \alpha_{3i}, \omega_{33}) \end{aligned} \tag{9.68}$$

where $\alpha_i \sim N(0, \Lambda)$, as before. They showed that, unconditional on $\alpha_i$, we have

$$I_i^* \sim N(Z_i\gamma, 1)$$

and that $f(y_{1i} \mid I_i = 1)$ and $f(y_{2i} \mid I_i = 0)$ are the conditional densities given by (9.65), with the parameter equivalence (9.67). From this, they argued that it is possible to construct two different observationally equivalent models, which produces an ambiguity in inferences. They argued that "although the interpretations of the parameters in each formulation are entirely different, the observed data cannot distinguish between these two different interpretations" (Poirier and Rudd, 1981, p. 255).

However, as can be seen from the equivalence of models (9.66) and (9.60) through (9.63), there are not two different models. *They are just two different ways of writing the same model. Thus, there is no ambi-*

[13] The likelihood function for this model is presented in equation (8.8) in Chapter 8. As noted there, the parameter $\sigma_{12}$ is not estimable.

*guity of inferences.* Note that although it appears from (9.68) as though Poirier and Rudd defined only the conditional distributions, this is not so. Because, as mentioned earlier, conditional on $\alpha_i$, the variables $y_{1i}$, $y_{2i}$, and $I_i^*$ are independent, equations (9.66) and (9.68) are exactly equivalent. That is,

$$\begin{aligned} f(y_{1i} \mid \alpha_i, I_i = 1) &= f(y_{1i} \mid \alpha_i, I_i^* > 0) \\ &= f(y_{1i} \mid \alpha_i) \end{aligned}$$

Thus, it is not true that Poirier and Rudd constructed a different model based on a specification of conditional distributions that gives the same likelihood function.

Another argument Poirier and Rudd made (1981, p. 250) was that "the contrast between a trivariate model and bivariate data suggests a major limitation of the model." Note, however, that the data do refer to three variables, each of which is partially observed. $I^*$ is observed as a dichotomous variable; $y_i$ is observed only when $I^* > 0$, and $y_2$ is observed only when $I^* < 0$. The only problem that arises is that $y_1$ and $y_2$ are not observed simultaneously. As a consequence, $\sigma_{12}$ is not estimable. But apart from this, there are no identifiability problems.[14] Thus, the problem is not one of a trivariate model and bivariate data, but one of partial observability, and there are many such models that are of practical use.

Poirier and Rudd also seem to have argued that because $y_1$ is observed only if $I=1$, and $y_2$ only if $I=0$, we should model only these conditional distributions. As shown earlier, the conditional specification they suggested is not indeed a specification of the distributions over the subpopulations. In the next section we shall discuss such a specification. However, just because $y_1$ is observed only if $I=1$ does not mean that we should specify the distribution of $y_1$ only on this subset. To push the analogy further, consider the tobit model: Just because $y_i$ is observed if and only if $y_i > 0$ does not mean that the distribution of the disturbance term need be specified for this subpopulation alone! In the case of sequential-decision models, of course, we might specify distributions of residuals on subpopulations only.

Of more practical importance is the issue raised by Poirier and Rudd concerning the interpretation of coefficient estimates. The selectivity

[14] Some other identification problems arise in the model in which only $y_1$ or $y_2$ is observed (see section 8.4). The fact that $\sigma_{12}$ is not estimable means that there can be any number of models with different values of correlations between $u_1$ and $u_2$ in equations (9.60) and (9.61) that are observationally equivalent. But the problem is not one of conditional versus marginal distributions, but one of covariance between the errors of marginal distributions.

model permits two types of inferences: conditional and marginal. For instance, with respect to the parameters in equation (9.60), we can consider $\partial E(y_{1i})/\partial X_{1i}$ for inferences from the marginal distribution and $\partial E(y_{1i} \mid I_i=1)/\partial X_{1i}$ for inferences from the conditional distributions. The former are given by the estimates of the parameter $\beta_1$ in (9.57). For the latter, we note that under the assumptions of normality of the residuals,

$$E(y_{1i} \mid I_i=1) = X_{1i}\beta_1 - \sigma_{1\epsilon}\frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)}$$

where $\sigma_{1\epsilon}=\mathrm{Cov}(u_1,\epsilon)$ and $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the density and distribution functions of the standard normal.

If there is a variable that appears in both $X_{1i}$ and $Z_i$ (say in the $j$th position for each), then

$$\frac{\partial E(y_{1i} \mid I_i=1)}{\partial X_{1ij}} = \beta_{1j} + \gamma_j\sigma_{1\epsilon}\frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)}\left(Z_i\gamma + \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)}\right)$$

Note that the sign changes on $\sigma_{1\epsilon}$ in our equations, as compared with those in the study of Poirier and Rudd, arise from the way equation (9.63) is defined.

Poirier and Rudd pointed out that given that sex is a variable included in both $X_{1i}$ and $Z_i$, Lee and Trost (1978, p. 374) incorrectly argued from the sign of $\beta_{1j}$ that "females tend to spend more than males if they own houses." However, a reading of Lee and Trost's study indicates that what they had in mind all along was potential expenditures, and thus they meant to say "if they were to own houses" rather than "if they own houses." There is, thus, no confusion or misinterpretation. The substantive issue here is what type of inference is of practical interest in this problem. Is it inference from the marginal distribution (potential expenditures on housing) or from the conditional distributions (actual expenditures on housing)? The answer to this question is that it depends on the problem at hand. If we are considering the effects of tax incentives, as did Rosen (1979), then we have to consider potential as well as actual expenditures. Poirier and Rudd (1981, p. 283) argued that if only conditional inferences are needed, then one should model directly the conditional densities and not bother about the selectivity model. But certainly the housing-decision model that they cited is not one in which our interest is only in conditional inferences. We can perhaps find examples in which only conditional inferences are of interest. In the next section we shall give examples in which the selectivity model makes sense but the mixture model does not.

## 9.8 When can the selection model be used, but not the mixture model?

In the preceding section we argued that the main difference between the selection model and the mixture model is in the specification of the distributions of the disturbances over the entire population versus subpopulations. We shall now give examples of cases in which the mixture model is not applicable.

To be specific, the selection model ($S$) is

$$I^* = Z\gamma - \epsilon$$
$$Y_1 = X_1\beta_1 + u_1$$
$$Y_2 = X_2\beta_2 + u_2$$

where $\epsilon$, $u_1$, and $u_2$ have well-defined distributions on the whole population. The mixture-distribution model ($M$) is

$$I^* = Z\gamma - \epsilon$$
$$Y_1 = X_1\beta_1 + u_1$$
$$Y_2 = X_2\beta_2 + u_2$$

where $\epsilon$ has a well-defined distribution on the whole population. The distribution of $u_1$ is defined only on the subpopulation for which $I=1$, and the distribution of $u_2$ is defined only on the subpopulation for which $I=0$.

The $M$ model is not appropriate for modeling the problems in which $y_1$ and $y_2$ are explicitly factors in the decision process, as in the union-decision model of Lee (1978). Here we have

$$I^* = \alpha(y_1 - y_2) + Z\gamma - \epsilon \tag{9.69}$$

This implies

$$I^* = \alpha(X_1\beta_1 - X_2\beta_2) + Z\gamma - v$$

where $v=\epsilon-\alpha(u_1-u_2)$. The disturbance term $v$ does not have a well-defined distribution in the $M$ model, because the distributions of $u_1$ and $u_2$ are defined only on subpopulations. The main interest in such models centers on the decision process, in particular the significance or nonsignificance of $\alpha$ in (9.69); see, for instance, the work of Willis and Rosen (1979). These situations can be modeled only by the selection models, not the mixture models. For each individual, $y_{1i}-y_{2i}$ represents the net gain (or net loss) from the choice between the two options. If $y_{1i}$ is the return of the outcome from choosing option 1, $y_{2i}$ will be the foregone

outcome from option 2 (and vice versa). Any econometric models involving discrete choice and foregone outcomes (or earnings) are meaningless to be modeled as $M$ models. The selection model is the appropriate one to be used.

The selection models are also useful for evaluating many government programs. These cannot be analyzed by the mixture-distribution model. These problems have been discussed in section 9.2.

In many activities (or choices) involving productivity or earnings, such as job choices, we observe that individuals engage in one activity rather than others. A possible reason from the productivity (or earnings) point of view, as suggested by Roy (1951), Sattinger (1975), Rosen (1978), and Willis and Rosen (1979), is that the individual has comparative advantage in an activity that, as compared with the other options, increases the well-being of the individual. To infer the implications of comparative advantage in discrete-choice behavior, one needs to have information on the potential outcomes from the unchosen alternative options. For this purpose, the selectivity model is useful. The mixture-distribution model does not permit any inferences to be made in such cases. An example cited by Roy (1951) was discussed in section 9.1.

## 9.9 Summary and conclusions

The preceding discussion suggests the usefulness of the selectivity model in a number of situations. The selectivity model has been applied in the following types of studies, among others:

1. Studies of participation in the labor force: Heckman (1974, 1979), Nelson (1977), Cogan (1980), Hanoch (1980*a*, *b*)
2. Studies of retirement decisions: Gordon and Blinder (1980)
3. Studies of returns to education: Griliches et al. (1978), Kenny et al. (1979), Willis and Rosen (1979)
4. Studies of the effects of unions on wages: Lee (1978), Abowd and Farber (1982)
5. Studies of the effects of employment services: Katz (1977)
6. Studies of migration and income: Nakosteen and Zimmer (1980)
7. Studies of physician behavior: Poirier (1981); studies of lawyer behavior: Weisbrod (1980)
8. Studies of electric utility rates: Roberts et al. (1978)
9. Studies of tenure choice and the demand for housing: Trost (1977), Lee and Trost (1978), Rosen (1979), King (1980)

We shall discuss the union-and-wages problem in Chapter 11. The studies on labor supply are far too numerous to review here. Regarding

the area of tenure choice and demand for housing, the traditional literature treats the discrete tenure choice and the continuous housing-demand choice separately. Trost (1977), Lee and Trost (1978), and Rosen (1979) recognized that the two decisions are interdependent, and they specified error terms of the discrete- and continuous-choice models to be correlated. King (1980) extended this analysis further in two important directions. First, he noted that because tenure choice and housing demand are based on maximization of the same utility function, the two models can involve some of the same parameters. In that case, joint estimation will involve imposing cross-equation constraints on the parameters of the tenure-choice and housing-demand equations, as well as recognizing that error terms are correlated. Second, King incorporated into the model estimates of the impact of rationing in the mortgage market and in the local-authority rental market in the United Kingdom. Because going through these aspects would involve reproducing most of King's study, it will not be done here.