

Limited Dependent and Sample Selection Models

Our discussion so far has been based on the sample being randomly drawn. A lot of times the observed samples are not drawn randomly. Moreover, the range of realized values may not take any values between $-\infty$ and ∞ , but may be constrained. For instance, in his pioneering work on household expenditure on durable goods, Tobin (1958) used a regression model that specifically took account of the fact that the expenditure (the dependent variable of his regression model) cannot be negative. Tobin called this type of model the model of limited dependent variables. It and its various generalizations are known as Tobit models because of their similarities to Probit models.¹ When a sample consists only of measurement with a range, it is called a *truncated sample*. When a sample consists of those with exact measurements if they fall within a range and those with no exact measurements but only indications that they fall outside of this range, it is called a *censored sample*. In this chapter we consider linear regression models when the observed sample is truncated or censored or nonrandomly drawn. Section 7.1 provides a brief review of the cross-sectional approach. Section 7.2 gives an example showing that attrition in panel data could lead to a nonrandom sample even when the initial wave is a random sample. Section 3 considers panel random-effects models. Section 7.4 considers fixed-effects models. An example on housing expenditure is given in Section 7.5. Section 7.6 considers issues of panel dynamic limited dependent and sample selection models.

7.1 CROSS-SECTIONAL DATA APPROACH

7.1.1 Truncated or Censored Regression Models

It is more convenient to relate an observed sample y that is subject to truncation or selection through a noncensored latent response function,

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u, \quad (7.1.1)$$

where \mathbf{x} is a $K \times 1$ vector of exogenous variables and u is the error term that can take any value between $-\infty$ and ∞ . Without loss of generality, suppose that the *observed* y is related to the *latent variable* y^* by

$$y = \begin{cases} y^*, & \text{if } y^* > 0, \\ 0, & \text{if } y^* \leq 0. \end{cases} \quad (7.1.2)$$

¹ See Amemiya (1985) and Maddala (1983) for extensive discussions of various types of Tobit models.

Models of the form (7.1.1) and (7.1.2) are called *censored regression* models if the data consist of those points of (y_i^*, \mathbf{x}_i) if $y_i^* > 0$ and $(0, \mathbf{x}_i)$ if $y_i^* \leq 0$ for $i = 1, \dots, N$. The *truncated data* consist only of points of (y_i^*, \mathbf{x}_i) where $y_i^* > 0$.

The conditional expectation of y given \mathbf{x} for truncated data is equal to

$$E(y \mid \mathbf{x}, y > 0) = E(y^* \mid \mathbf{x}, y^* > 0) = \mathbf{x}'\boldsymbol{\beta} + E(u \mid u > -\mathbf{x}'\boldsymbol{\beta}). \quad (7.1.3)$$

The conditional expectation of y given \mathbf{x} for censored data is equal to

$$\begin{aligned} E(y \mid \mathbf{x}) &= \text{Prob}(y = 0) \cdot 0 + \text{Prob}(y > 0 \mid \mathbf{x}) \cdot E(y \mid y > 0, \mathbf{x}) \\ &= \text{Prob}(u \leq -\mathbf{x}'\boldsymbol{\beta}) \cdot 0 + \text{Prob}(u > -\mathbf{x}'\boldsymbol{\beta})E(y^* \mid \mathbf{x}; u > -\mathbf{x}'\boldsymbol{\beta}) \\ &= \text{Prob}(u > -\mathbf{x}'\boldsymbol{\beta})[\mathbf{x}'\boldsymbol{\beta} + E(u \mid u > -\mathbf{x}'\boldsymbol{\beta})]. \end{aligned} \quad (7.1.4)$$

If u is independently normally distributed with mean 0 and variance σ_u^2 , then

$$\text{Prob}(u > -\mathbf{x}'\boldsymbol{\beta}) = 1 - \Phi\left(\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right) = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right), \quad (7.1.5)$$

and

$$E(u \mid u > -\mathbf{x}'\boldsymbol{\beta}) = \sigma_u \cdot \frac{\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right)}{\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right)}, \quad (7.1.6)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote standard normal density, $N(0, 1)$, and cumulative (or integrated) normal, $\Phi(a) = \int_{-\infty}^a \phi(u)du$, respectively. Equations (7.1.3) and (7.1.4) show that truncation or censoring of the dependent variables introduces dependence between the error term and the regressors for the model

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon, \quad (7.1.7)$$

where the error is

$$\epsilon = v + E(y \mid \mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}. \quad (7.1.8)$$

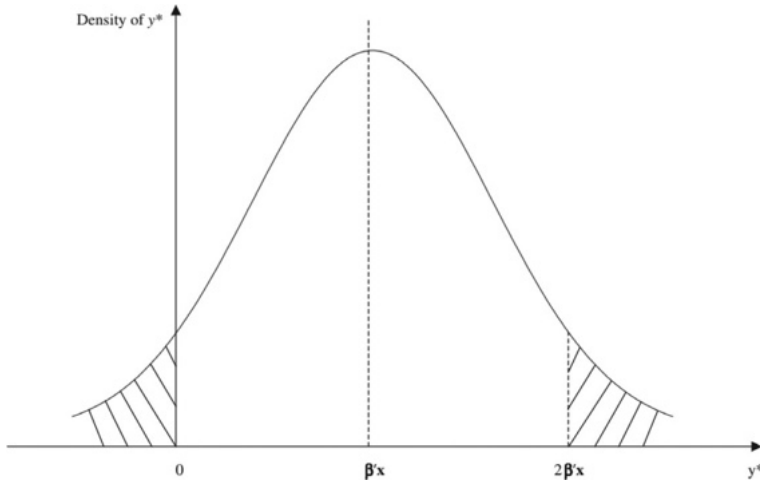
Although $v = y - E(y \mid \mathbf{x})$ has $E(v \mid \mathbf{x}) = 0$, but $E(\epsilon \mid \mathbf{x}) \neq 0$. Therefore, the least squares estimator of (7.1.7) is biased and inconsistent.

The likelihood function of the truncated data is equal to

$$L_1 = \prod_{i \in \psi} [\text{Prob}(y_i > 0 \mid \mathbf{x}_i)]^{-1} f(y_i) \quad (7.1.9)$$

where $f(\cdot)$ denotes the density of y_i^* (or u_i) and $\psi = \{i|y_i^* > 0\}$. Let $\bar{\psi} = \{i|y_i^* \leq 0\}$, the likelihood function of the censored data is equal to

$$\begin{aligned}
L_2 &= \left\{ \prod_{i \in \bar{\psi}} \text{Prob}(y_i = 0 \mid \mathbf{x}_i) \cdot \prod_{i \in \psi} \text{Prob}(y_i > 0 \mid \mathbf{x}_i) \right\} \\
&\quad \cdot \left\{ \prod_{i \in \psi} [\text{Prob}(y_i > 0 \mid \mathbf{x}_i)]^{-1} f(y_i) \right\} \\
&= \prod_{i \in \bar{\psi}} \text{Prob}(y_i = 0 \mid \mathbf{x}_i) \prod_{i \in \psi} f(y_i).
\end{aligned} \tag{7.1.10}$$

Figure 7.1. Density of y^* censored or truncated at 0.

To make this approach more explicit, consider first the case in which the dependent variable is truncated at zero. In such a truncated sample, data points for which $u_i \leq -\mathbf{x}'_i\boldsymbol{\beta}$ when $\mathbf{x}'_i\boldsymbol{\beta} > 0$ are omitted. But if data points with $u_i \geq \mathbf{x}'_i\boldsymbol{\beta}$ are also excluded from the sample, then any remaining observations would have error terms lying within the interval $(-\mathbf{x}'_i\boldsymbol{\beta}, \mathbf{x}'_i\boldsymbol{\beta})$. (Any observations for which $\mathbf{x}'_i\boldsymbol{\beta} \leq 0$ are automatically deleted.) If u is symmetrically distributed, the residuals for the “symmetrically truncated” sample will also be symmetrically distributed about zero. The corresponding dependent variable would take values between zero and $2\mathbf{x}'_i\boldsymbol{\beta}$, as shown in the region AOB of Figure 7.2. In other words, points b and c in Figure 7.2 are thrown away (point a is not observed). Therefore, the moment conditions

$$E[1(y < 2\mathbf{x}'\boldsymbol{\beta})(y - \mathbf{x}'\boldsymbol{\beta}) | \mathbf{x}] = 0, \quad (7.1.12)$$

and

$$E[1(y < 2\mathbf{x}'\boldsymbol{\beta})(y - \mathbf{x}'\boldsymbol{\beta})\mathbf{x}] = \mathbf{0}, \quad (7.1.13)$$

hold, where $1(A)$ denotes the indicator function that takes the value 1 if A occurs and 0 otherwise.

The sample analog of (7.1.13) is

$$\frac{1}{N} \sum_{i=1}^N 1(y_i < 2\mathbf{x}'_i\hat{\boldsymbol{\beta}})(y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}})\mathbf{x}_i = \mathbf{0} \quad (7.1.14)$$

which is the first-order condition of applying the least squares principle to symmetrically trimmed truncated data falling in the region AOB.

Definition of the symmetrically trimmed estimator for a censored sample is similarly motivated. The error terms of the censored regression model are of the form $u_i^* = \max\{u_i, -\mathbf{x}'_i\boldsymbol{\beta}\}$, (i.e., point a in Figure 7.2 is moved to the corresponding circled point a'). “Symmetric censoring” would replace u_i^* with $\min\{u_i^*, \mathbf{x}'_i\boldsymbol{\beta}\}$ whenever $\mathbf{x}'_i\boldsymbol{\beta} > 0$, and would delete the observation otherwise. In other words, the dependent variable $y_i = \max\{0, y_i^*\}$ is replaced with $\min\{y_i, 2\mathbf{x}'_i\boldsymbol{\beta}\}$ as the points a, b, c in Figure 7.2 have been moved

The least squares method yields the mean. The least absolute deviation method yields the median (e.g., Amemiya 1984). When $E(y^* | \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, censoring affects the mean, $E(y | \mathbf{x})$, but does not affect the median; therefore, Powell (1984) suggests a least absolute deviation estimator of $\boldsymbol{\beta}$ by minimizing

$$E(y_i | d_i = 1) = \mathbf{x}_i' \boldsymbol{\beta} + E(u_i | v_i > -\mathbf{w}_i' \mathbf{a}) \quad (7.1.25)$$

The likelihood function of (y_i, d_i) is

$$\begin{aligned} L &= \prod_c \text{Prob} (d_i = 0) \prod_{\bar{c}} f(y_i \mid d_i = 1) \text{Prob} (d_i = 1), \\ &= \prod_c \text{Prob} (d_i = 0) \cdot \prod_{\bar{c}} \text{Prob} (d_i^* > 0 \mid y_i) f(y_i), \end{aligned} \quad (7.1.26)$$

where $c = \{i \mid d_i = 0\}$, and \bar{c} denotes its complement. If the joint distribution of (u, v) is specified, one can estimate this model by the MLE. For instance, if (u, v) is jointly normally distributed with mean $(0, 0)$ and covariance matrix $\begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{vu} & 1 \end{pmatrix}$, then

$$E(u \mid v > -\mathbf{w}'\mathbf{a}) = \sigma_{uv} \frac{\phi(\mathbf{w}'\mathbf{a})}{\Phi(\mathbf{w}'\mathbf{a})}, \quad (7.1.27)$$

$$\text{Prob}(d = 0) = [1 - \Phi(w'a)] = \Phi(-w'a), \quad (7.1.28)$$

$$\text{Prob}(d = 1 \mid y) = \Phi \left\{ \frac{w'a + \frac{\sigma_{uv}}{\sigma_u}(y - x'\beta)}{\sigma_u} \right\}. \quad (7.1.29)$$

Alternatively, Heckman's (1979) two-stage method can be applied. First, estimate \mathbf{a} by a Probit MLE of $d_i, i = 1, \dots, N$. Evaluate $\phi(\mathbf{a}'\mathbf{w}_i)/\Phi(\mathbf{a}'\mathbf{w}_i)$ using the estimated \mathbf{a} . Second, regress y_i on \mathbf{x}_i and $\phi(\hat{\mathbf{a}}'\mathbf{w}_i)/\Phi(\hat{\mathbf{a}}'\mathbf{w}_i)$ using data corresponding to $d_i = 1$ only.

Just like the standard Tobit model, the consistency and asymptotic normality of the MLE and Heckman two-stage estimator for the endogenously determined selection depend critically on the correct assumption of the joint probability distribution of (u, v) . When the distribution of (u, v) is unknown, the coefficients of \mathbf{x} which are not overlapping with \mathbf{w} can be estimated by a semiparametric method.

For ease of exposition, suppose that there are no variables appearing in both \mathbf{x} and \mathbf{w} . Then, as noted by Robinson (1988b), the model of (7.1.1), (7.1.23), and (7.1.24) conditional on $d_i = 1$ becomes a partially linear model of the form:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \lambda(\mathbf{w}_i) + \epsilon_i, \quad (7.1.30)$$

where $\lambda(\mathbf{w}_i)$ denotes the unknown selection factor. The expectation of y_i conditional on \mathbf{w}_i and $d_i = 1$ is equal to

$$E(y_i \mid \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}' E(\mathbf{x}_i \mid \mathbf{w}_i, d_i = 1) + \lambda(\mathbf{w}_i). \quad (7.1.31)$$

Subtracting (7.1.31) from (7.1.30) yields

$$y_i - E(y_i \mid \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'(x_i - E(x_i \mid \mathbf{w}_i, d_i = 1)) + \epsilon_i, \quad (7.1.32)$$

where $E(\epsilon_i \mid \mathbf{w}_i, \mathbf{x}_i, d_i = 1) = 0$. Thus, Robinson (1988b) suggests estimating β by

$$\beta = \{E(x - E(x | w))[x - E(x | w)]'\}^{-1} E[(x - E(x | w))[y - E(y | w)]], \quad (7.1.33)$$

using the truncated sample.

The first-stage conditional expectation for the estimator (7.1.31) can be estimated by the nonparametric method. For instance, one may use the kernel method to estimate the density of y at y_a (e.g., Härdle 1990; Robinson 1989):

$$\hat{f}(y_a) = \frac{1}{Nh_N} \sum_{i=1}^N k\left(\frac{y_i - y_a}{h_N}\right), \quad (7.1.34)$$

where h_N is a positive number called the “bandwidth” or “smoothing” parameter that tends to zero as $N \rightarrow \infty$, $k(u)$ is a kernel function that is a bounded symmetric probability density function (pdf) that integrates to 1. Similarly, one can construct a kernel estimator of a multivariate pdf at \mathbf{w}_a , $f(\mathbf{w}_a)$ by

$$\hat{f}(\mathbf{w}_a) = \frac{1}{N |H_m|} \sum_{i=1}^N k_m(H_m^{-1}(\mathbf{w}_i - \mathbf{w}_a)), \quad (7.1.35)$$

where \mathbf{w} is a $m \times 1$ vector of random variables, k_m is a kernel function on m dimensional space, and H_m is a positive definite matrix. For instance, $k_m(\mathbf{u})$ can be the multivariate normal density function or $k_m(\mathbf{u}) = \prod_{j=1}^m k(u_j)$, $\mathbf{u}' = (u_1, \dots, u_m)$, $H_m = \text{diag}(h_{1N}, \dots, h_{mN})$.

Kernel estimates of a conditional pdf $f(y_a | \mathbf{w}_a)$ or conditional expectations $Eg(y | \mathbf{w}_a)$ may be derived from the kernel estimates of the joint pdf and marginal pdf. Thus, the conditional pdf may be estimated by

$$\hat{f}(y_a | \mathbf{w}_a) = \frac{\hat{f}(y_a, \mathbf{w}_a)}{\hat{f}(\mathbf{w}_a)} \quad (7.1.36)$$

and the conditional expectation by

$$\hat{E}g(y | \mathbf{w}_a) = \frac{1}{N |H_m|} \sum_{i=1}^N g(y_i) k_m(H_m^{-1}(\mathbf{w}_i - \mathbf{w}_a)) / \hat{f}(\mathbf{w}_a). \quad (7.1.37)$$

The differencing equation (7.1.32) eliminates all variables that appear in both \mathbf{x}_i and \mathbf{w}_i . The Robinson (1988b) approach cannot identify the parameters of variables that appear both in the regression equation, \mathbf{x} , and the selection equation, \mathbf{w} . When there are variables appearing in both \mathbf{x} and \mathbf{w} , Newey (2009) suggests a two-step series method of estimating β provided that the selection correction term of (7.1.30), $\lambda(\mathbf{w}_i, d_i = 1)$, is a function of the single index, $\mathbf{w}'_i \mathbf{a}$,

$$\lambda(\mathbf{w}, d = 1) = E[u | v(\mathbf{w}'\mathbf{a}), d = 1]. \quad (7.1.38)$$

The first step of Newey's method uses the distribution-free method discussed in Chapter 6 or Klein and Spady (1993) to estimate \mathbf{a} . The second step consists of a linear regression of $d_i y_i$ on $d_i \mathbf{x}_i$ and the approximations of $\lambda(\mathbf{w}_i)$. Newey suggests approximating $\lambda(\mathbf{w}_i)$ by either a polynomial function of $(\mathbf{w}'_i \hat{\mathbf{a}})$ or a spline function, $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a}) = (P_{1K}(\mathbf{w}'\mathbf{a}), P_{2K}(\mathbf{w}'\mathbf{a}), \dots, P_{KK}(\mathbf{w}'\mathbf{a}))'$ with the property that for large K , a linear combination of $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a})$ can approximate an unknown function of $\lambda(\mathbf{w}'\mathbf{a})$ well. Newey (2009) shows that the two-step series estimation of β is consistent and asymptotically normally distributed when $N \rightarrow \infty$, $K \rightarrow \infty$, and $\sqrt{N} K^{-s-t+1} \rightarrow 0$ where $s \geq 5$ and $K^7/N \rightarrow 0$ if $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a})$ is a power series or $m \geq t - 1$, $s \geq 3$, and $K^4/N \rightarrow 0$ if $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a})$ is a spline of degree m in $(\mathbf{w}'\mathbf{a})$.²

² For instance, a spline of degree m in $(\mathbf{w}'\hat{\mathbf{a}})$ with L evenly spaced knots on $[-1, 1]$ can be based on

$$P_{kK} = (\mathbf{w}'\mathbf{a})^{k-1}, 1 \leq k \leq m+1, \\ = \{[(\mathbf{w}'\mathbf{a}) + 1 - 2(k-m-1)/(L+1)]_+\}^m, m+2 \leq k \leq m+1+L \equiv K,$$

where $b_+ \equiv 1(b > 0) \cdot b$.

$$\begin{aligned}
& E(u \mid v > -\mathbf{w}'\mathbf{a}) \text{Prob}(v > -\mathbf{w}'\mathbf{a}) - E(u \mid v > \mathbf{w}'\mathbf{a}) \text{Prob}(v > \mathbf{w}'\mathbf{a}) \\
&= \int_{-\infty}^{\infty} \int_{-\mathbf{w}'\mathbf{a}}^{\infty} u f(u, v) du dv - \int_{-\infty}^{\infty} \int_{\mathbf{w}'\mathbf{a}}^{\infty} u f(u, v) du dv \\
&= \int_{-\infty}^{\infty} \int_{-\mathbf{w}'\mathbf{a}}^{\mathbf{w}'\mathbf{a}} u f(u, v) du dv = 0,
\end{aligned} \tag{7.1.42}$$

where, without loss of generality, we let $\mathbf{w}'\mathbf{a} > 0$. It follows that

$$\begin{aligned}
& E[d_i y_i - d_j y_j - (d_i x_i - d_j x_j)' \boldsymbol{\beta} \mid \mathbf{w}'_i \mathbf{a} = -\mathbf{w}'_j \mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] \\
&= E[d_i u_i - d_j u_j \mid \mathbf{w}'_i \mathbf{a} = -\mathbf{w}'_j \mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] = 0.
\end{aligned} \tag{7.1.43}$$

Because $E[d_i - d_j \mid \mathbf{w}'_i \mathbf{a} = -\mathbf{w}'_j \mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] = 2 \text{Prob}(d_i = 1 \mid \mathbf{w}'_i \mathbf{a}) - 1 \neq 0$ and the conditioning is on $\mathbf{w}'_i \mathbf{a} = -\mathbf{w}'_j \mathbf{a}$, not on $\mathbf{w}'_i \mathbf{a} = \mathbf{w}'_j \mathbf{a}$, the moment condition (7.1.43) allows the identification of the intercept and the slope parameters without the need to impose the exclusion restriction that at least one component of \mathbf{x} is excluded from \mathbf{w} . Therefore, Chen (1999) suggests a \sqrt{N} consistent instrumental variable estimator for the intercept and the slope parameters as

$$\hat{\boldsymbol{\beta}}_c = \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N K \left(\frac{(\mathbf{w}_i + \mathbf{w}_j)' \hat{\mathbf{a}}}{h_N} \right) (d_i x_i - d_j x_j)(z_i - z_j)' \right]^{-1} \tag{7.1.44}$$

$$\left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N K \left(\frac{(\mathbf{w}_i + \mathbf{w}_j)' \hat{\mathbf{a}}}{h_N} \right) (z_i - z_j)' (d_i y_i - d_j y_j) \right],$$

where z_i are the instruments for $d_i x_i$. In the case when y are unobservable, but the corresponding x are observable, the natural instrument will be $E(d \mid \mathbf{w}'\mathbf{a})\mathbf{x}$. An efficient method for estimating binary choice models that contain an intercept term suggested by Chen (2000) can be used to obtain the first stage estimate of \mathbf{a} .

7.2 A SAMPLE SELECTION EXAMPLE: NONRANDOMLY MISSING DATA IN A PANEL

Attrition is a problem in any panel survey. For instance, by 1981, all four of the national longitudinal surveys started in the 1960s had lost at least one-fourth of their original samples. In the Gary income maintenance project, 206 of the sample of 585 black, male-headed households, or 35.2%, did not complete the experiment. If data are randomly missing, there is no sample truncation or selection. However, the major problem in econometrics is not simply missing data but the possibility that they are missing for a variety of self-selection reasons. For instance, in a social experiment, although the first wave of the data could be randomly drawn, some individuals may decide that keeping the detailed records that the experiments require is not worth the payment and opt to drop out. Some may move or may be inducted into the military. In some experiments, persons with large earnings receive no experimental-treatment benefit and thus drop out of the experiment altogether. This attrition may negate the randomization in the initial experiment design. If the probability of attrition is correlated with experimental response, then traditional statistical techniques will lead to biased and inconsistent estimates of the experimental effect. In this section we show how models of limited dependent variables (e.g., see the surveys of Amemiya 1984; Heckman 1976; and Maddala 1983) can provide

both the theory and computational techniques for analyzing nonrandomly missing data (Griliches, Hall, and Hausman 1978; Hausman and Wise 1979).³

7.2.1 A Probability Model of Attrition and Sample Selection Bias

Suppose that the structural model is

$$y_{it} = \beta' x_{it} + v_{it}, \quad i = 1, \dots, N, \\ t = 1, \dots, T, \quad (7.2.1)$$

where the error term v_{it} is assumed to follow a conventional random effects error-components formulation $v_{it} = \alpha_i + u_{it}$, as discussed in Chapter 2. For ease of exposition, we assume that $T = 2$.

If attrition occurs in the second period, a common practice is to discard those observations for which y_{i2} is missing. But suppose that the probability of observing y_{i2} varies with its value, as well as the values of other variables; then the probability of observing y_{i2} will depend on v_{i2} . Least squares of (7.2.1) based on observed y will lead to biased estimates of the underlying structural parameters and the experimental response.

To formalize the argument, let the indicator variable $d_i = 1$ if y_{i2} is observed in period 2, and $d_i = 0$ if y_{i2} is not observed; in other words, attrition occurs. Suppose that y_{i2} is observed ($d_i = 1$) if the latent variable

$$d_i^* = \gamma y_{i2} + \theta' x_{i2} + \delta' w_i + \epsilon_i^* \geq 0, \quad (7.2.2)$$

where w_i is a vector of variables that do not enter the conditional expectation of y but affect the probability of observing y , and θ and δ are vectors of parameters. Suppose (v_i, ϵ_i^*) are jointly normally distributed. Substituting for y_{i2} leads to the reduced-form specification

$$d_i^* = (\gamma \beta' + \theta') x_{i2} + \delta' w_i + \gamma v_{i2} + \epsilon_i^* \\ = \pi' x_{i2} + \delta' w_i + \epsilon_i \\ = a' R_i + \epsilon_i, \quad (7.2.3)$$

where $\epsilon_i = \gamma v_{i2} + \epsilon_i^*$, and $R_i = (x'_{i2}, w'_i)'$, and $a' = (\pi', \delta')$. We normalize the variance of $\epsilon_i, \sigma_{\epsilon}^2$, equal to 1. Then the probabilities of retention and attrition are Probit functions given, respectively, by

$$\text{Prob}(d_i = 1) = \Phi(a' R_i), \quad \text{and} \\ \text{Prob}(d_i = 0) = 1 - \Phi(a' R_i), \quad (7.2.4)$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Suppose we estimate the model (7.2.1) using only complete observations. The conditional expectation of y_{i2} , given that it is observed, is

$$E(y_{i2} \mid x_{i2}, w_i, d_i = 1) = \beta' x_{i2} + E(v_{i2} \mid x_{i2}, w_i, d_i = 1). \quad (7.2.5)$$

³ Another example is the analysis of event histories in which responses are at nonequally spaced points in time (e.g., Heckman and Singer 1984, Lancaster 1990). Some people choose to model event histories in discrete time using sequences of binary indicators. Then the subject becomes very much like the discrete panel data analysis discussed in Chapter 6.

$$\begin{aligned} E(v_{i2} \mid \mathbf{w}_i, d_i = 1) &= \sigma_{2\epsilon} E(\epsilon_i \mid \mathbf{w}_i, d_i = 1) \\ &= \frac{\sigma_{2\epsilon}}{\Phi(\mathbf{a}' R_i)} \int_{-\mathbf{a}' R_i}^{\infty} \epsilon \cdot \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2} d\epsilon \\ &= \sigma_{2\epsilon} \frac{\phi(\mathbf{a}' R_i)}{\Phi(\mathbf{a}' R_i)}, \end{aligned} \quad (7.2.6)$$
$$E(y_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}' \mathbf{x}_{i2} + \sigma_{2\epsilon} \frac{\phi(\mathbf{a}' R_i)}{\Phi(\mathbf{a}' R_i)}. \quad (7.2.7)$$

When $d_i = 1$, the joint density of $d_i = 1$, y_{i1} , and y_{i2} is given by

$$\begin{aligned}
 f(d_i = 1, y_{i1}, y_{i2}) &= \text{Prob}(d_i = 1 \mid y_{i1}, y_{i2}) f(y_{i1}, y_{i2}) \\
 &= \text{Prob}(d_i = 1 \mid y_{i2}) f(y_{i1}, y_{i2}) \\
 &= \Phi \left\{ \frac{a' R_i + \left(\frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right) (y_{i2} - \beta' x_{i2})}{\left[1 - \frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right\} \\
 &\quad \cdot [2\pi \sigma_u^2 (\sigma_u^2 + 2\sigma_\alpha^2)]^{-1/2} \\
 &\quad \cdot \exp \left\{ -\frac{1}{2\sigma_u^2} \left[\sum_{t=1}^2 (y_{it} - \beta' x_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \right. \right. \\
 &\quad \left. \left. \cdot \left(\sum_{t=1}^2 (y_{it} - \beta' x_{it}) \right)^2 \right] \right\},
 \end{aligned} \tag{7.2.8}$$

$$\begin{aligned}
 f(d_i = 0, y_{i1}) &= \text{Prob}(d_i = 0 \mid y_{i1}) f(y_{i1}) \\
 &= \left\{ 1 - \Phi \left[\frac{\mathbf{a}' R_i + \frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})}{\left[1 - \frac{\sigma_{1\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right] \right\} \\
 &\quad \cdot [2\pi(\sigma_u^2 + \sigma_\alpha^2)]^{-1/2} \\
 &\quad \cdot \exp \left\{ -\frac{1}{2(\sigma_u^2 + \sigma_\alpha^2)} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})^2 \right\}.
 \end{aligned} \tag{7.2.9}$$

The right-hand side of (7.2.9) follows from the fact that $f(\epsilon_i | v_{i1})$ is normal, with mean $[\sigma_{1\epsilon}/(\sigma_u^2 + \sigma_\alpha^2)]v_{i1}$ and variance $1 - \sigma_{1\epsilon}^2/(\sigma_u^2 + \sigma_\alpha^2)$, where $\sigma_{1\epsilon}$ is the covariance between v_{i1} and ϵ_i , which is equal to $\sigma_{2\epsilon} = \sigma_\alpha^2/(\sigma_u^2 + \sigma_\alpha^2)$.

The likelihood function follows from (7.2.8) and (7.2.9). Order the observations so that the first N_1 observations correspond to $d_i = 1$, and the remaining $N - N_1$ correspond to $d_i = 0$; then the log-likelihood function is given by

$$\begin{aligned} \log L = & -N \log 2\pi - \frac{N_1}{2} \log \sigma_u^2 - \frac{N_1}{2} \log (\sigma_u^2 + 2\sigma_\alpha^2) \\ & - \frac{N - N_1}{2} \log (\sigma_u^2 + \sigma_\alpha^2) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^{N_1} \left\{ \sum_{t=1}^2 (y_{it} - \beta' x_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \left[\sum_{t=1}^2 (y_{it} - \beta' x_{it}) \right]^2 \right\} \\ & + \sum_{i=1}^{N_1} \log \Phi \left\{ \frac{a' R_i + \frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} (y_{i2} - \beta' x_{i2})}{\left[1 - \frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right\} \\ & - \frac{1}{2(\sigma_u^2 + \sigma_\alpha^2)} \sum_{i=N_1+1}^N (y_{i1} - \beta' x_{i1})^2 \\ & + \sum_{i=N_1+1}^N \log \left\{ 1 - \Phi \left[\frac{a' R_i + \frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} (y_{i1} - \beta' x_{i1})}{\left[1 - \frac{\sigma_{1\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right] \right\}. \quad (7.2.10) \end{aligned}$$

The critical parameter for attrition bias is $\sigma_{2\epsilon}$. If $\sigma_{2\epsilon} = 0$, so does $\sigma_{1\epsilon}$. The likelihood function (7.2.10) then separates into two parts. One corresponds to the variance-components specification for y . The other corresponds to the Probit specification for attrition. Thus, if attrition bias is not present, this is identical with the random missing-data situations. Generalized-least-squares techniques used to estimate (7.2.1) will lead to consistent and asymptotically efficient estimates of the structural parameters of the model.

The Hausman–Wise two-period model of attrition can be extended in a straightforward manner to more than two periods and to simultaneous-equations models with selection bias. When $T > 2$, an attrition equation can be specified for each period. If attrition occurs, the individual does not return to the sample; then a series of conditional densities analogous to (7.2.8) and (7.2.9) result. The last period for which the individual appears in the sample gives information on which the random term in the attrition equations is conditioned. For periods in which the individual remains in the sample, an equation like (7.2.8) is used to specify the joint probability of no attrition and the observed values of the dependent variables.

In the case of simultaneous-equations models, all the attrition model leads to is simply to add an equation for the probability of observing an individual in the sample. Then the joint density of observing in-sample respondents becomes the product of the conditional probability of the observation being in the sample, given the joint dependent variable \mathbf{y} , and the marginal density of \mathbf{y} . The joint density of incomplete respondents becomes the product of the conditional probability of the observation being out-of-sample, given the

The employment of probability equations to specify the status of individuals can be very useful in analyzing the general problems of changing compositions of the sample over time, in particular when changes are functions of individual characteristics. For instance, in addition to the problem of attrition in the national longitudinal surveys' samples of young men, there is also the problem of sample accretion, that is, entrance into the labor force of the fraction of the sample originally enrolled in school. The literature on switching regression models can be used as a basis for constructing behavioral models for analyzing the changing status of individuals over time.⁴

The Gary income-maintenance project focused on the impact of alternative sets of income-maintenance structures on work-leisure decisions. The basic project design was to randomly divide individuals into two groups: "controls" and "experimentals." The controls were not on an experimental-treatment plan but received nominal payments for completing periodic questionnaires. The experimentals were randomly assigned to one of several income-maintenance plans. The experiment had four basic plans defined by an income guarantee and a tax rate. The two guarantee levels were \$4,300 and \$3,300 for a family of four and were adjusted up for larger families and down for smaller families. The two marginal tax rates were 0.6 and 0.4. Retrospective information of individuals in the experiments was also surveyed for a pre-experimental period (normally just prior to the beginning of the experimental period) so that the behavior of experimentals during the experiment could be compared with their own pre-experimental behavior and also compared with that of the control group to obtain estimates of the effects of treatment plans.

If attrition is random, as will be discussed in Section 9.6, it is not a major problem. What matters is that data are missing for a variety of self-selection reasons. In this case it is easy to imagine that attrition is related to endogenous variables. Beyond a break-even point, experimentals receive no benefits from the experimental treatment. The break-even point occurs when the guarantee minus taxes paid on earnings (wage rate times hours worked) is zero. Individuals with high earnings receive no treatment payment and may be much like controls vis-à-vis their incentive to remain in the experiment. But because high earnings are caused in part by the unobserved random term of the structural equation (7.2.1), attrition may well be related to it.

⁴ See Quandt (1982) for a survey of switching regression models.

Table 7.1. *Parameter estimates of the earnings-function structural model with and without a correction for attrition*

Variables	With attrition correction: maximum likelihood estimates (standard errors)		Without attrition correction: Generalized least squares estimates (standard errors): earnings-function parameters
	Earnings-function parameters	Attrition parameters	
Constant	5.8539 (0.0903)	−0.6347 (0.3351)	5.8911 (0.0829)
Experimental effect	−0.0822 (0.0402)	0.2414 (0.1211)	−0.0793 (0.0390)
Time trend	0.0940 (0.0520)	— ^a —	0.0841 (0.0358)
Education	0.0209 (0.0052)	−0.0204 (0.0244)	0.0136 (0.0050)
Experience	0.0037 (0.0013)	−0.0038 (0.0061)	0.0020 (0.0013)
Nonlabor income	−0.0131 (0.0050)	0.1752 (0.0470)	−0.0115 (0.0044)
Union	0.2159 (0.0362)	1.4290 (0.1252)	0.2853 (0.0330)
Poor health	−0.0601 (0.0330)	0.2480 (0.1237)	−0.0578 (0.0326)
$\hat{\sigma}_u^2 = 0.1832$ (0.0057)			$\hat{\sigma}_u^2 = 0.1236$
$\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = 0.2596$ (0.0391)			$\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = 0.2003$
$\frac{\hat{\sigma}_{2\epsilon}}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = -0.1089$ (0.0429)			

^a Not estimated.

Source: Hausman and Wise (1979, Table IV).

did not differ, the coefficient of this variable provided a reasonable indicator of the effect of experimental treatment on hours worked.

Because only three observations were available during the experiment, each for a one-month period, they concentrated on a two-period model: a period for the pre-experiment average monthly earnings and a period for the average earning of the three monthly observations of the experimental period. Their generalized least squares estimates of the structural parameters that were not corrected for attrition and the maximum likelihood estimates that incorporated the effects of attrition, (7.2.1) and (7.2.3), are presented in Table 7.1.

The attrition-bias parameter $\sigma_{2\epsilon}/(\sigma_u^2 + \sigma_\alpha^2)$ was estimated to be -0.1089 . This indicates a small but statistically significant correlation between earnings and the probability of attrition. The estimate of the experimental effect was very close whether or not the attrition bias was corrected for. However, the experimental-effect coefficient did increase in magnitude from -0.079 to -0.082 , an increase of 3.6%. Some of the other coefficients showed more pronounced changes. The effect of non-labor-family income on earnings (hence hours worked) decreased by 23% from the generalized least squares estimates, and the effect of another year of education increased by 43%. These results demonstrate that attrition bias was a potentially important problem in the Gary experiment. For other examples, see Ridder (1990), Nijman and Verbeek (1992), and Verbeek and Nijman (1996).

The Hausman–Wise (HW) model assumes that the contemporaneous values affect the probability of responding. Alternatively, the decision on whether to respond may be related to past experiences – if in the first period the effort in responding was high, an individual

may be less inclined to respond in the second period. When the probability of attrition depends on lagged but not on contemporaneous variables, and if v_{it} and ϵ_i^* are mutually independent; then individuals are missing at random (MAR) (Little and Rubin 1987; Rubin 1976) and the missing data are ignorable (this case is sometimes referred to as selection on observables; e.g., Moffitt, Fitzgerald, and Gottschalk 1997).

Both sets of models are often used to deal with attrition in panel data sets. However, they rely on fundamentally different restrictions on the dependence of the attrition process on time path of the variables and can lead to very different inferences. In a two-period model one cannot introduce dependence on y_{i2} in the MAR model, or dependence on y_{i1} in the HW model, without relying heavily on functional form and distributional assumptions. However, when missing data are augmented by replacing the units that have dropped out with new units randomly sampled from the original population, called refreshment samples by Ridder (1992), it is possible to test between these two types of models nonparametrically as well as to estimate more general models (e.g., Hirano, Imbens, Ridder, and Rubin 2001).

7.3 TOBIT MODELS WITH RANDOM INDIVIDUAL EFFECTS

The most typical concern in empirical work using panel data has been the presence of unobserved heterogeneity.⁵ Thus, a linear latent response function is often written in the form

$$y_{it}^* = \alpha_i + \beta' x_{it} + u_{it}, \quad \begin{matrix} i = 1, \dots, N, \\ t = 1, \dots, T, \end{matrix} \quad (7.3.1)$$

with the error term assumed to be independent of x_{it} and is i.i.d. over time and across individuals, where the observed value y_{it} is equal to y_{it}^* if $y_{it}^* > 0$ and is unobserved for $y_{it}^* \leq 0$ when data is truncated and is equal to zero when data is censored. Under the assumption that α_i is randomly distributed with density function $g(\alpha)$ (or $g(\alpha | x)$), the likelihood function of the standard Tobit model for the truncated data is of the form

$$\prod_{i=1}^N \int \left[\prod_{t=1}^T [1 - F(-\beta' x_{it} - \alpha_i)]^{-1} f(y_{it} - \beta' x_{it} - \alpha_i) \right] g(\alpha_i) d\alpha_i, \quad (7.3.2)$$

where $f(\cdot)$ denotes the density function of u_{it} and $F(a) = \int_{-\infty}^a f(u) du$. The likelihood function of the censored data takes the form

$$\prod_{i=1}^N \int \left[\prod_{t \in c_i} F(-\beta' x_{it} - \alpha_i) \prod_{t \in \bar{c}_i} f(y_{it} - \alpha_i - \beta' x_{it}) \right] g(\alpha_i) d\alpha_i, \quad (7.3.3)$$

where $c_i = \{t | y_{it} = 0\}$ and \bar{c}_i denotes its complement. Maximizing (7.3.2) or (7.3.3) with respect to unknown parameters yields a consistent and asymptotically normally distributed estimator.

Similarly, for the type II Tobit model, we may specify a sample selection equation

$$d_{it}^* = w_{it}' a + \eta_i + v_{it}, \quad (7.3.4)$$

with the observed (y_{it}, d_{it}) following the rule of $d_{it} = 1$ if $d_{it}^* > 0$ and zero otherwise as in (7.1.23) and $y_{it} = y_{it}^*$ if $d_{it} = 1$ and unknown otherwise, as in (7.1.24). Suppose that the

⁵ In this chapter we consider only the case involving the presence of individual specific effects. For some generalization to the estimation of the random-coefficients sample selection model, see Chen (1999).

7.4.1.1 Truncated Regression

We assume a model of (7.3.1) and (7.1.3), except that now the individual effects are assumed fixed. The disturbance u_{it} is again assumed to be independently distributed over i and independently, identically distributed (i.i.d.) over t conditional on \mathbf{x}_i and α_i .

We note that where data are truncated or censored, first differencing does not eliminate the individual-specific effects from the specification. To see this, suppose that the data are truncated. Let

$$y_{it} = E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) + \epsilon_{it}, \quad (7.4.1)$$

where

$$E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}). \quad (7.4.2)$$

Since $\mathbf{x}_{it} \neq \mathbf{x}_{is}$, in general,

$$\begin{aligned} & E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) - E(y_{is} \mid \mathbf{x}_{is}, \alpha_i, y_{is} > 0) \\ &= (\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}' \boldsymbol{\beta}) \\ &\quad - E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}' \boldsymbol{\beta}), \end{aligned} \tag{7.4.3}$$

In other words,

$$\begin{aligned} (y_{it} - y_{is}) &= (\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}' \boldsymbol{\beta}) \\ &\quad - E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}' \boldsymbol{\beta}) + (\epsilon_{it} - \epsilon_{is}). \end{aligned} \quad (7.4.4)$$

The truncation correction term, $E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}'_{it}\boldsymbol{\beta})$, which is a function of the individual-specific effects α_i , remains after first differencing. However, we may eliminate the truncation correction term through first differencing if we restrict our analysis to observations where $y_{it} > (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$ and $y_{is} > -(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$. To see this, suppose that $(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} < 0$, then

$$\begin{aligned} E(y_{is} \mid \alpha_i, \mathbf{x}_{it}, \mathbf{x}_{is}, y_{is} > -(\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta}) \\ = \alpha_i + \mathbf{x}'_{is} \boldsymbol{\beta} + E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}'_{is} \boldsymbol{\beta} - (\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta}). \end{aligned} \quad (7.4.5)$$

Since u_{it} conditional on \mathbf{x}_i and α_i is assumed to be i.i.d.,

$$E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}'_{it}\boldsymbol{\beta}) = E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}'_{it}\boldsymbol{\beta}). \quad (7.4.6)$$

Similarly, if $(\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta} > 0$,

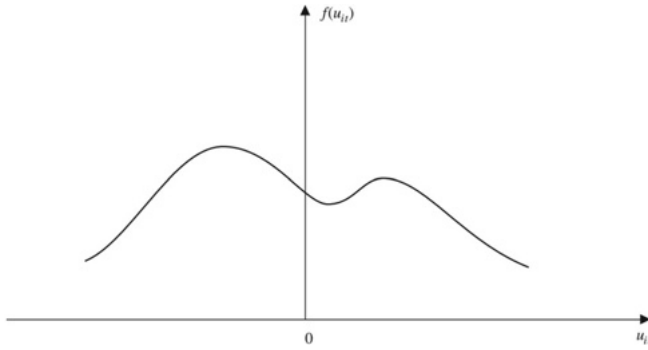
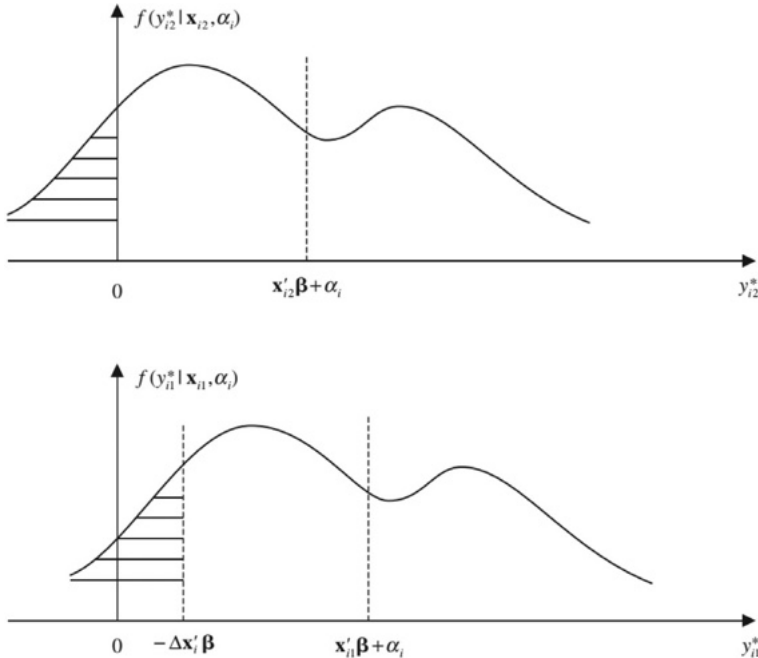
$$\begin{aligned} E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta} + (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}) \\ &= E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}) \\ &= E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}). \end{aligned} \tag{7.4.7}$$

Therefore, by confining our analysis to the truncated observations where $y_{it} > (\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta}$, $y_{is} > -(\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta}$, $y_{it} > 0$, $y_{is} > 0$, we have

$$(y_{it} - y_{is}) = (\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta} + (\epsilon_{it} - \epsilon_{is}), \quad (7.4.8)$$

which no longer involves the incidental parameter, α_i . Since $E[(\epsilon_{it} - \epsilon_{is}) \mid \mathbf{x}_{it}, \mathbf{x}_{is}] = 0$, applying least squares to (7.4.8) will yield a consistent estimator of β .

The idea of restoring symmetry of the error terms of the pairwise differencing equation $(y_{it} - y_{is})$ by throwing away observations where $y_{it} < (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$ and $y_{is} < -(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$ can be seen by considering the following graphs, assuming that $T = 2$. Suppose that the probability density function of u_{it} is of the shape shown in Figure 7.3. Since u_{i1} and u_{i2}

Figure 7.3. Probability density of u_{it} .Figure 7.4. Conditional densities of y_{i1}^* and y_{i2}^* given $(x_{i1}, x_{i2}, \alpha_i)$, assuming $\Delta x'_i \beta < 0$.

are i.i.d. conditional on $(x_{i1}, x_{i2}, \alpha_i)$, the probability density of y_{i1}^* and y_{i2}^* conditional on $(x_{i1}, x_{i2}, \alpha_i)$ should have the same shape except for the location. The top and bottom figures of Figure 7.4 postulate the probability density of y_{i1}^* and y_{i2}^* conditional on $(x_{i1}, x_{i2}, \alpha_i)$, respectively, assuming that $\Delta x'_i \beta < 0$, where $\Delta x_i = x_{i2} - x_{i1}$. The truncated data correspond to those sample points where y_{it}^* or $y_{it} > 0$. Because $x'_{i1} \beta \neq x'_{i2} \beta$, the probability density of y_{i1} is different from that of y_{i2} . However, the probability density of y_{i1}^* given $y_{i1}^* > -\Delta x'_i \beta$ (or y_{i1} given $y_{i1} > -\Delta x'_i \beta$) is identical to the probability density of y_{i2}^* given $y_{i2}^* > 0$ (or y_{i2} given $y_{i2} > 0$), as shown in Figure 7.4. Similarly, if $\Delta x'_i \beta > 0$, the probability density of y_{i1}^* given $y_{i1}^* > 0$ (or y_{i1} given $y_{i1} > 0$) is identical to the probability density of y_{i2}^* given $y_{i2}^* > \Delta x'_i \beta$, as shown in Figure 7.5.⁷

⁷ I owe this exposition to the suggestion of J. L. Powell. Note that with panel data we can relax the assumption of $f(u_{it})$ to be unimodal and symmetric.

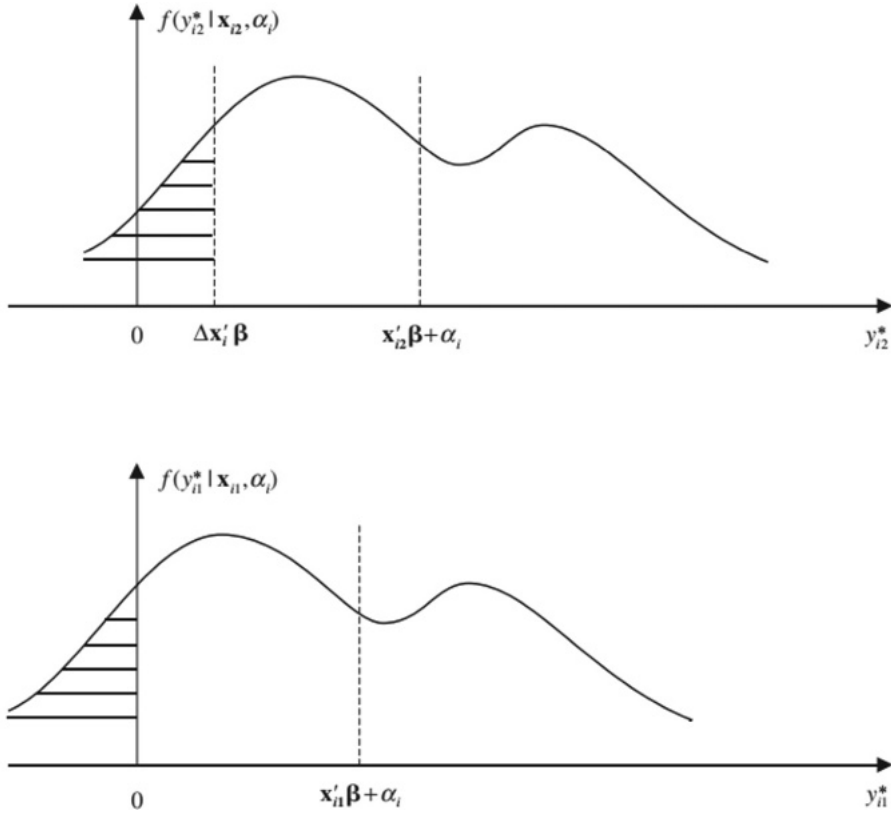


Figure 7.5. Conditional densities of y_{i1}^* and y_{i2}^* given $(x_{i1}, x_{i2}, \alpha_i)$, assuming $\Delta x_i' \beta > 0$.

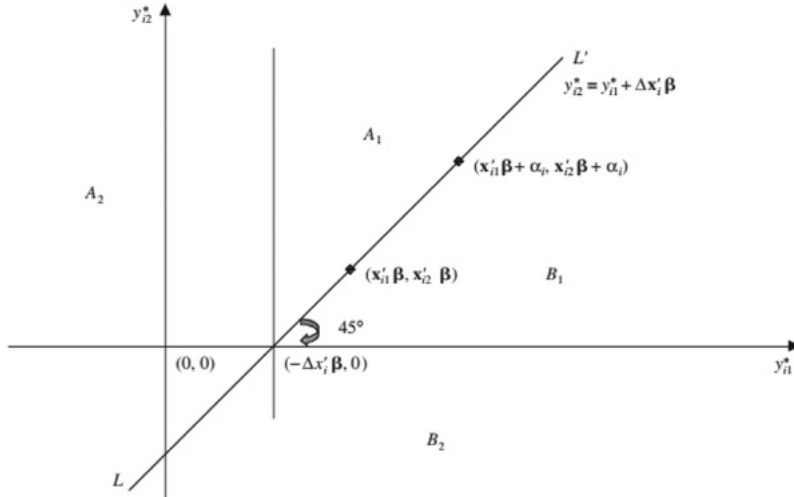


Figure 7.6. The distribution of (y_{i1}^*, y_{i2}^*) assuming $\Delta x_i' \beta < 0$.

$$\begin{aligned}
 A_1 &= \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta x_i' \beta, y_{i2}^* > y_{i1}^* + \Delta x_i' \beta\}, \\
 A_2 &= \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq -\Delta x_i' \beta, y_{i2}^* > 0\}, \\
 B_1 &= \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta x_i' \beta, 0 < y_{i2}^* < y_{i1}^* + \Delta x_i' \beta\}, \\
 B_2 &= \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta x_i' \beta, y_{i2}^* \leq 0\}.
 \end{aligned}$$

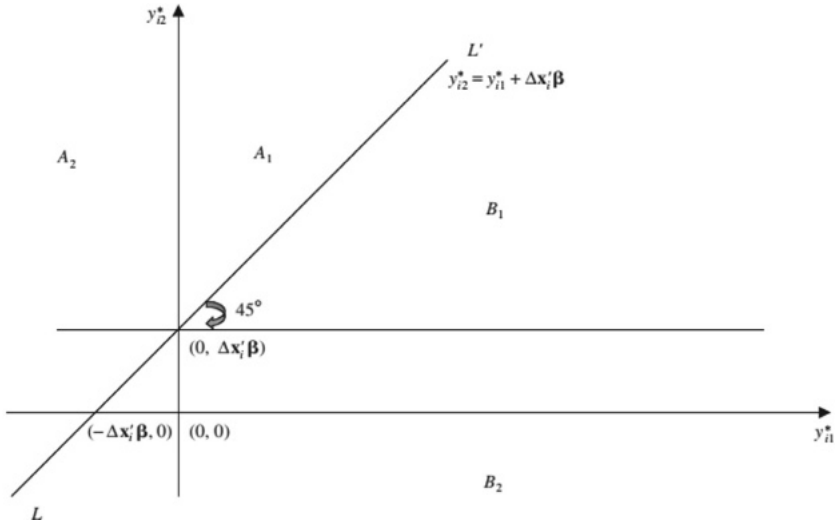


Figure 7.7. The distribution of (y_{i1}^*, y_{i2}^*) assuming $\Delta x_i' \beta > 0$.

$$A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* > y_{i1}^* + \Delta x_i' \beta\},$$

$$A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq 0, y_{i2}^* > \Delta x_i' \beta\},$$

$$B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, \Delta x_i' \beta < y_{i2}^* < y_{i1}^* + \Delta x_i' \beta\},$$

$$B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* \leq \Delta x_i' \beta\}.$$

In the two-dimensional diagram of (y_{i1}^*, y_{i2}^*) of Figures 7.6 and 7.7, (y_{i1}^*, y_{i2}^*) conditional on $(x_{i1}, x_{i2}, \alpha_i)$ is symmetrically distributed around the 45-degree line through $(x_{i1}'\beta + \alpha_i, x_{i2}'\beta + \alpha_i)$ or equivalently around the 45-degree line through $(x_{i1}'\beta, x_{i2}'\beta)$ or $(-\Delta x_i' \beta, 0)$ as the line LL' . Since this is true for any value of α_i , the same statement is true for the distribution of (y_{i1}^*, y_{i2}^*) conditional on (x_{i1}, x_{i2}) . When $\Delta x_i' \beta < 0$, the symmetry of the distribution of (y_{i1}^*, y_{i2}^*) around LL' means that the probability that (y_{i1}^*, y_{i2}^*) falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta x_i' \beta, y_{i2}^* > y_{i1}^* + \Delta x_i' \beta\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta x_i' \beta, 0 < y_{i2}^* < y_{i1}^* + \Delta x_i' \beta\}$. (Figure 7.6). When $\Delta x_i' \beta > 0$, the probability that (y_{i1}^*, y_{i2}^*) falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* > y_{i1}^* + \Delta x_i' \beta\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, \Delta x_i' \beta < y_{i2}^* < y_{i1}^* + \Delta x_i' \beta\}$. (Figure 7.7). That is, points in the regions A_1 and B_1 are not affected by the truncation. On the other hand, points falling into the region $(0 < y_{i1}^* < -\Delta x_i' \beta, y_{i2}^* > 0)$ in Figure 7.6 (which correspond to points $(y_{i1} < -\Delta x_i' \beta, y_{i2})$) and $(y_{i1}^* > 0, 0 < y_{i2}^* < \Delta x_i' \beta)$ in Figure 7.7 (which correspond to points $(y_{i1}, y_{i2} < \Delta x_i' \beta)$) will have to be thrown away to restore symmetry.

Let $C = \{i \mid y_{i1} > -\Delta x_i' \beta, y_{i2} > \Delta x_i' \beta\}$; then $(y_{i1} - x_{i1}'\beta - \alpha_i)$ and $(y_{i2} - x_{i2}'\beta - \alpha_i)$ for $i \in C$ are symmetrically distributed around zero. Therefore, $E[(y_{i2} - y_{i1}) - (x_{i2} - x_{i1})'\beta \mid x_{i1}, x_{i2}, i \in C] = 0$. In other words,

$$\begin{aligned} E[\Delta y_i - \Delta x_i' \beta \mid y_{i1} > -\Delta x_i' \beta, y_{i2} > \Delta x_i' \beta] \\ = E[\Delta y_i - \Delta x_i' \beta \mid y_{i1}^* > 0, y_{i1}^* > -\Delta x_i' \beta, y_{i2}^* > 0, y_{i2}^* > \Delta x_i' \beta] = 0, \end{aligned} \quad (7.4.9a)$$

and

$$E[(\Delta y_i - \Delta x_i' \beta) \Delta x_i \mid y_{i1} > -\Delta x_i' \beta, y_{i2} > \Delta x_i' \beta] = 0, \quad (7.4.9b)$$

where $\Delta y_i = \Delta y_{i2} = y_{i2} - y_{i1}$. However, there could be multiple roots that satisfy (7.4.9b). To ensure a unique solution for β , Honoré (1992) suggests the trimmed LAD and LS estimators as those $\hat{\beta}$ and $\tilde{\beta}$ that minimize the objective functions

$$\begin{aligned}\hat{\Gamma}_1^{(j,k)} = \frac{1}{h_N} & \left[\frac{1}{N} \sum_{i=1}^N (1\{\Delta y_i < \Delta \mathbf{x}'_i(\hat{\boldsymbol{\beta}} + h_N \mathbf{i}_k) < y_{i2}\} \right. \\ & - 1\{-y_{i1} < \Delta \mathbf{x}'_i(\hat{\boldsymbol{\beta}} + h_N \mathbf{i}_k) < \Delta y_i\}) \Delta \mathbf{x}_i^{(j)} \\ & + \frac{1}{N} \sum_{i=1}^N (-1\{\Delta y_i < \Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}} < y_{i2}\} \\ & \left. - 1\{-y_{i1} < \Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}} < \Delta y_i\}) \Delta \mathbf{x}_i^{(j)} \right],\end{aligned}\quad (7.4.18)$$

$$\begin{aligned}\hat{\Gamma}_2^{(j,k)} = \frac{1}{h_N} & \left[\frac{1}{N} \sum_{i=1}^N 1\{-y_{i1} < \Delta \mathbf{x}'_i(\tilde{\boldsymbol{\beta}} + h_N \mathbf{i}_k) < y_{i2}\} \right. \\ & \times \left(\Delta y_i - \Delta \mathbf{x}'_i(\tilde{\boldsymbol{\beta}} + h_N \mathbf{i}_k) \right) \Delta \mathbf{x}_i^{(j)} \\ & \left. - \frac{1}{N} \sum_{i=1}^N 1\{-y_{i1} < \Delta \mathbf{x}'_i \tilde{\boldsymbol{\beta}} < y_{i2}\} (\Delta y_i - \Delta \mathbf{x}'_i \tilde{\boldsymbol{\beta}}) \Delta \mathbf{x}_i^{(j)} \right],\end{aligned}\quad (7.4.19)$$

where $\Gamma_\ell^{(j,k)}$ denotes the (j,k) th element of Γ_ℓ , for $\ell = 1, 2$, $\Delta \mathbf{x}_i^{(j)}$ denotes the j th coordinate of $\Delta \mathbf{x}_i$, \mathbf{i}_k is a unit vector with 1 in its k th place and h_N decreases to zero with the speed of $N^{-\frac{1}{2}}$. The bandwidth factor h_N appears in (7.4.18) and (7.4.19) because Γ_ℓ is a function of densities and conditional expectations of y (Honoré 1992).

7.4.1.2 Censored Regressions

When data are censored, observations $\{y_{it}, \mathbf{x}_{it}\}$ are available for $i = 1, \dots, N, t = 1, \dots, T$, where $y_{it} = \max\{0, y_{it}^*\}$. In other words, y_{it} can now be either 0 or a positive number rather than just a positive number as in the case of truncated data. Of course, we can throw away observations of $(y_{it}, \mathbf{x}_{it})$ that correspond to $y_{it} = 0$ and treat the censored regression model as the truncated regression model using the methods of Section 7.4.1.1. But this will lead to a loss of information. In the case that data are censored, in addition to the relation (7.4.9a,b), the joint probability of $y_{i1} \leq -\boldsymbol{\beta}' \Delta \mathbf{x}_i$ and $y_{i2} > 0$ is identical to the joint probability of $y_{i1} > -\boldsymbol{\beta}' \Delta \mathbf{x}_i$ and $y_{i2} = 0$, when $\boldsymbol{\beta}' \Delta \mathbf{x}_i < 0$, as shown in Figure 7.6, region A_2 and B_2 , respectively. When $\boldsymbol{\beta}' \Delta \mathbf{x}_i > 0$, the joint probability of $y_{i1} = 0$ and $y_{i2} > \boldsymbol{\beta}' \Delta \mathbf{x}_i$ is identical to the joint probability of $y_{i1} > 0$ and $y_{i2} \leq \boldsymbol{\beta}' \Delta \mathbf{x}_i$ as shown in Figure 7.7. In other words, (y_{i1}^*, y_{i2}^*) conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$ is symmetrically distributed around the 45-degree line through $(\mathbf{x}'_{i1} \boldsymbol{\beta} + \alpha_i, \mathbf{x}'_{i2} \boldsymbol{\beta} + \alpha_i)$ or equivalently around the 45-degree line through $(-\Delta \mathbf{x}'_i \boldsymbol{\beta}, 0)$ as the line LL' in Figures 7.6 and 7.7. Since this is true for any value of α_i , the same statement is true for the distribution of (y_{i1}^*, y_{i2}^*) conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$. When $\Delta \mathbf{x}'_i \boldsymbol{\beta} < 0$, the symmetry of the distribution of (y_{i1}^*, y_{i2}^*) around LL' means that the probability that (y_{i1}^*, y_{i2}^*) falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta \mathbf{x}'_i \boldsymbol{\beta}, y_{i2}^* > y_{i1}^* + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta \mathbf{x}'_i \boldsymbol{\beta}, 0 < y_{i2}^* < y_{i1}^* + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$. Similarly, the probability that (y_{i1}^*, y_{i2}^*) falls in the region $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* < -\Delta \mathbf{x}'_i \boldsymbol{\beta}, y_{i2}^* > 0\}$ equals the probability that it falls in the region $B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta \mathbf{x}'_i \boldsymbol{\beta}, y_{i2}^* \leq 0\}$, as shown in Figure 7.6. When $\Delta \mathbf{x}'_i \boldsymbol{\beta} > 0$, the probability that (y_{i1}^*, y_{i2}^*) falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* > y_{i1}^* + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, \Delta \mathbf{x}'_i \boldsymbol{\beta} < y_{i2}^* < y_{i1}^* + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$, and the probability that it falls in

the region $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq 0, y_{i2}^* > \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* \leq \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$ as in Figure 7.7. Therefore, the probability of (y_{i1}^*, y_{i2}^*) conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$ falling in $A = (A_1 \cup A_2)$ equals the probability that it falls in $B = (B_1 \cup B_2)$. As neither of these probabilities is affected by censoring, the same is true in the censored sample. This implies that

$$E \left[(1\{(y_{i1}, y_{i2}) \in A\} - 1\{(y_{i1}, y_{i2}) \in B\}) \Delta \mathbf{x}_i \right] = \mathbf{0}. \quad (7.4.20)$$

In other words, to restore symmetry of censored observations around their expected values, observations correspond to $(y_{i1} = 0, y_{i2} < \Delta \mathbf{x}_i' \boldsymbol{\beta})$ or $(y_{i1} < -\Delta \mathbf{x}_i' \boldsymbol{\beta}, y_{i2} = 0)$ will have to be thrown away.

By the same argument, conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$, the expected vertical distance from a (y_{i1}, y_{i2}) in A to the boundary of A equals the expected horizontal distance from a (y_{i1}, y_{i2}) in B to the boundary of B . For (y_{i1}, y_{i2}) in A_1 , the vertical distance to LL' is $(\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta})$. For (y_{i1}, y_{i2}) in B_1 , the horizontal distance to LL' is $y_{i1} - (y_{i2} - \Delta \mathbf{x}_i' \boldsymbol{\beta}) = -(\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta})$. For (y_{i1}, y_{i2}) in A_2 , the vertical distance to the boundary of A_2 is $y_{i2} - \max(0, \Delta \mathbf{x}_i' \boldsymbol{\beta})$. For (y_{i1}, y_{i2}) in B_2 , the horizontal distance is $y_{i1} - \max(0, -\Delta \mathbf{x}_i' \boldsymbol{\beta})$. Therefore,

$$E \left[\left(1\{(y_{i1}, y_{i2}) \in A_1\} (\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta}) + 1\{(y_{i1}, y_{i2}) \in A_2\} (y_{i2} - \max(0, \Delta \mathbf{x}_i' \boldsymbol{\beta})) \right. \right. \\ \left. \left. + 1\{(y_{i1}, y_{i2}) \in B_1\} (\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta}) \right. \right. \\ \left. \left. - 1\{(y_{i1}, y_{i2}) \in B_2\} (y_{i1} - \max(0, -\Delta \mathbf{x}_i' \boldsymbol{\beta})) \right) \Delta \mathbf{x}_i \right] = \mathbf{0}. \quad (7.4.21)$$

The pairwise trimmed LAD and LS estimators, $\hat{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$, for the estimation of the censored regression model proposed by Honoré (1992) are obtained by minimizing the objective functions

$$\mathcal{Q}_N^*(\boldsymbol{\beta}) = \sum_{i=1}^N \left[1 - 1\{y_{i1} \leq -\Delta \mathbf{x}_i' \boldsymbol{\beta}, y_{i2} \leq 0\} \right] \left[1 - 1\{y_{i2} \leq \Delta \mathbf{x}_i' \boldsymbol{\beta}, y_{i1} \leq 0\} \right] \\ \times |\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta}| \\ = \sum_{i=1}^N \psi^*(y_{i1}, y_{i2}, \Delta \mathbf{x}_i \boldsymbol{\beta}), \quad (7.4.22)$$

$$R_N^*(\boldsymbol{\beta}) = \sum_{i=1}^N \left\{ \left[\max\{y_{i2}, \Delta \mathbf{x}_i' \boldsymbol{\beta}\} - \max\{y_{i1}, -\Delta \mathbf{x}_i' \boldsymbol{\beta} - \Delta \mathbf{x}_i' \boldsymbol{\beta}\} \right]^2 \right. \\ \left. - 2 \times 1\{y_{i1} < -\Delta \mathbf{x}_i' \boldsymbol{\beta}\} (y_{i1} + \Delta \mathbf{x}_i' \boldsymbol{\beta}) y_{i2} \right. \\ \left. - 2 \times 1\{y_{i2} < \Delta \mathbf{x}_i' \boldsymbol{\beta}\} (y_{i2} - \Delta \mathbf{x}_i' \boldsymbol{\beta}) y_{i1} \right\} \\ = \sum_{i=1}^N \chi(y_{i1}, y_{i2}, \Delta \mathbf{x}_i \boldsymbol{\beta}), \quad (7.4.23)$$

where

$$\psi^*(w_1, w_2, c) \\ = \begin{cases} 0, & \text{for } w_1 \leq \max(0, -c) \text{ and } w_2 \leq \max(0, c), \\ |w_2 - w_1 - c|, & \text{otherwise,} \end{cases}$$

and

$$\chi(w_1, w_2, c) = \begin{cases} w_1^2 - 2w_1(w_2 - c) & \text{for } w_2 \leq c, \\ (w_2 - w_1 - c)^2 & \text{for } -w_1 < c < w_2, \\ w_2^2 - 2w_2(c + w_1) & \text{for } c \leq -w_1, \end{cases}$$

which is convex in c . The first-order conditions of (7.4.22) and (7.4.23) are the sample analogs of (7.4.20) and (7.4.21), respectively. For instance, when $(y_{i1}, y_{i2}) \in (A_1 \cup B_1)$, the corresponding terms in R_N^* become $(\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta})^2$. When $(y_{i1}, y_{i2}) \in A_2$, the corresponding terms become $y_{i2}^2 - 2 \times 1\{y_{i1} < -\Delta \mathbf{x}_i' \boldsymbol{\beta}\} (y_{i1} + \Delta \mathbf{x}_i' \boldsymbol{\beta}) y_{i2}$. When $(y_{i1}, y_{i2}) \in B_2$, the corresponding terms become $y_{i1}^2 - 2 \times 1\{y_{i2} < \Delta \mathbf{x}_i' \boldsymbol{\beta}\} (y_{i2} - \Delta \mathbf{x}_i' \boldsymbol{\beta}) y_{i1}$. The partial derivative of the first term with respect to $\boldsymbol{\beta}$ converges to $E\{1\{(y_{i1}, y_{i2}) \in A_1\}(\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta}) + 1\{(y_{i1}, y_{i2}) \in B_1\}(\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta})\Delta \mathbf{x}_i\}$. The partial derivative of the second and third terms with respect to $\boldsymbol{\beta}$ yield $-2E\{1[(y_{i1}, y_{i2}) \in A_2]y_{i2}\Delta \mathbf{x}_i - 1[(y_{i1}, y_{i2}) \in B_2]y_{i1}\Delta \mathbf{x}_i\}$. Because $Q_N^*(\boldsymbol{\beta})$ is piecewise linear and convex and $R_N^*(\boldsymbol{\beta})$ is continuously differentiable and convex and twice differentiable except at a finite number of points, the censored pairwise trimmed LAD and LS estimators, $\hat{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$, are computationally simpler than the truncated estimators $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

Honoré (1992) shows that $\hat{\beta}^*$ and $\tilde{\beta}^*$ are consistent and asymptotically normally distributed. The asymptotic covariance matrix of $\sqrt{N}(\hat{\beta}^* - \beta)$ is equal to

$$\text{Asy. Cov}(\sqrt{N}(\hat{\beta}^* - \beta)) = \Gamma_3^{-1} V_3 \Gamma_3^{-1}, \quad (7.4.24)$$

and of $\sqrt{N}(\tilde{\beta}^* - \beta)$ is equal to

$$\text{Asy. Cov}(\sqrt{N}(\tilde{\beta}^* - \beta)) = \Gamma_4^{-1} V_4 \Gamma_4^{-1}, \quad (7.4.25)$$

where V_3, V_4, Γ_3 and Γ_4 may be approximated by

$$\begin{aligned} \hat{V}_3 = & \frac{1}{N} \sum_{i=1}^N 1 \left\{ \left[\Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}}^* < \Delta y_i, y_{i2} > \max(0, \Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}}^*) \right] \right. \\ & \left. \cup \left[\Delta y_i < \Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}}^*, y_{i1} > \max(0, -\Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}}^*) \right] \right\} \Delta \mathbf{x}_i \Delta \mathbf{x}'_i, \end{aligned} \quad (7.4.26)$$

$$\begin{aligned} \hat{V}_4 = & \frac{1}{N} \sum_{i=1}^N \left[y_{i2}^2 1\{\Delta \mathbf{x}'_i \tilde{\boldsymbol{\beta}}^* \leq -y_{i1}\} + y_{i1}^2 1\{y_{i2} \leq \Delta \mathbf{x}'_i \tilde{\boldsymbol{\beta}}^*\} \right. \\ & \left. + (\Delta y_i - \Delta \mathbf{x}'_1 \tilde{\boldsymbol{\beta}}^*)^2 1\{-y_{i1} < \Delta \mathbf{x}'_i \tilde{\boldsymbol{\beta}}^* < y_{i2}\} \right] \Delta \mathbf{x}_i \Delta \mathbf{x}'_i, \end{aligned} \quad (7.4.27)$$

$$\begin{aligned} \hat{\Gamma}_3^{(j,k)} = & \frac{-1}{h_N} \left\{ \frac{1}{N} \sum_{i=1}^N \left[1\{y_{i2} > 0, y_{i2} > y_{i1} + \Delta \mathbf{x}'_i (\hat{\boldsymbol{\beta}}^* + h_N \mathbf{i}_k)\} \right. \right. \\ & - 1\{y_{i1} > 0, y_{i1} > y_{i2} - \Delta \mathbf{x}'_i (\hat{\boldsymbol{\beta}}^* + h_N \mathbf{i}_k)\} \left. \right] \Delta \mathbf{x}_i^{(j)} \\ & - \frac{1}{N} \sum_{i=1}^N \left[1\{y_{i2} > 0, y_{i2} > y_{i1} + \Delta \mathbf{x}'_i \hat{\boldsymbol{\beta}}^*\} \right. \\ & \left. \left. - 1\{y_{i1} > 0, y_{i1} > y_{i2} - \Delta \mathbf{x}_i \hat{\boldsymbol{\beta}}^*\} \right] \Delta \mathbf{x}_i^{(j)} \right\}, \end{aligned} \quad (7.4.28)$$

and

$$\hat{\Gamma}_4 = \frac{1}{N} \sum_{i=1}^N 1\{-y_{i1} < \Delta \mathbf{x}'_i \tilde{\boldsymbol{\beta}}^* < y_{i2}\} \Delta \mathbf{x}_i \Delta \mathbf{x}'_i. \quad (7.4.29)$$

where \mathbf{i}_k is a unit vector with 1 in its k th place and h_N decreases to zero at the speed of $N^{-\frac{1}{2}}$.

Both the truncated and censored estimators are presented assuming that $T = 2$. They can be easily modified to cover the case where $T > 2$. For instance, (7.4.23) can be modified to be the estimator that minimizes

$$\sum_{i=1}^N \sum_{t=2}^T \chi(y_{i,t-1}, y_{it}, (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \boldsymbol{\beta}) \quad (7.4.30)$$

when $T > 2$.

The argument of the objective function (7.4.30) is based on $(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})$. If \mathbf{x}_{it} contains time-invariant variables \mathbf{z}_i , then they are removed from (7.4.30). Only the coefficients of time-varying explanatory variables can be estimated. To obtain the coefficients of $\mathbf{z}_i, \boldsymbol{\gamma}$, just like the binary choice model discussed in Section 6.4.3, Honoré and Kesina (2017) assume conditional on $\mathbf{x}'_{it} \boldsymbol{\beta} = \mathbf{x}'_{js} \boldsymbol{\beta}, \alpha_i + u_i$ are independently identically distributed over i and suggest to estimate $\boldsymbol{\gamma}$ by minimizing

$$\left(\frac{N}{2}\right)^{-1} \frac{1}{h_N} \sum_{i < j} \sum_{t=1}^{T_i} \sum_{s=1}^{T_j} K \left(\frac{(\mathbf{x}_{it} - \mathbf{x}_{js})' \hat{\boldsymbol{\beta}}}{h_N} \right) \chi(y_{it}, y_{js}, (\mathbf{z}_i - \mathbf{z}_j)' \boldsymbol{\gamma}), \quad (7.4.31)$$

where $K(\cdot)$ is a kernel and h_N a bandwidth chosen so that in the limit only pairs with $\mathbf{x}'_{it} \boldsymbol{\beta} = \mathbf{x}'_{js} \boldsymbol{\beta}$ are included in the estimation of $\boldsymbol{\gamma}$ (Honoré and Powell 1994).

7.4.2 A Semiparametric Two-Step Estimator for the Endogenously Determined Sample Selection Model

In this subsection we consider the estimation of the endogenously determined sample selection model in which the sample selection rule is determined by the binary response model (7.3.4) and (7.1.24) for the linear regression model (7.3.1), where $y_{it} = y_{it}^*$ if $d_{it} = 1$ and unknown if $d_{it} = 0$, as in (7.1.24). We assume that both (7.3.1) and (7.3.4) contain unobserved fixed individual-specific effects α_i and η_i that may be correlated with the observed explanatory variables in an arbitrary way. Following the spirit of Heckman's (1979) two-step estimation procedure for the parametric model, Kyriazidou (1997) proposed a two-step semiparametric method for estimating the main regression of interest (7.3.4). In the first step, the unknown coefficients of the "selection" equation (7.3.4), \mathbf{a} , are consistently estimated by some semi-parametric method. In the second step, these estimates are substituted into the equation of interest (7.3.1) conditional on $d_{it} = 1$ and estimate it by a weighted least squares method. The fixed effect from the main equation is eliminated by taking time differences on the observed y_{it} . The selection effect is eliminated by conditioning time differencing of y_{it} and y_{is} on those observations where $\mathbf{w}'_{it} \hat{\mathbf{a}} \simeq \mathbf{w}'_{is} \hat{\mathbf{a}}$ because the magnitude of the selection effect is the same if the impact of the observed variables determining selection remains the same over time.

We note that without sample selectivity, that is, $d_{it} = 1$ for all i and t , or if u_{it} and v_{it} are uncorrelated conditional on α_i and \mathbf{x}_{it} , then (7.3.1) and (7.1.24) correspond to the standard variable intercept model for panel data discussed in Chapter 2 with balanced

panel or randomly missing data.⁸ If u_{it} and v_{it} are correlated, sample selection will arise because $E(u_{it} | \mathbf{x}_{it}, \mathbf{w}_{it}, \alpha_i, d_{it} = 1) \neq 0$. Let $\lambda(\cdot)$ denote the conditional expectation of u conditional on $d = 1, \mathbf{x}, \mathbf{w}, \alpha$ and η , then (7.3.1) and (7.1.24) conditional on $d_{it} = 1$ can be written as

$$y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + \lambda(\eta_i + \mathbf{w}_{it}' \mathbf{a}) + \epsilon_{it}, \quad (7.4.32)$$

where $E(\epsilon_{it} | \mathbf{x}_{it}, d_{it} = 1) = 0$. The form of the selection function $\lambda(\cdot)$ is derived from the joint distribution of u and v . For instance, if u and v are bivariate normal, then we have

the Heckman sample selection correction of $\lambda(\eta_i + \mathbf{a}' \mathbf{w}_{it}) = \frac{\sigma_{uv}}{\sigma_v} \frac{\phi\left(\frac{\eta_i + \mathbf{w}_{it}' \mathbf{a}}{\sigma_v}\right)}{\Phi\left(\frac{\eta_i + \mathbf{w}_{it}' \mathbf{a}}{\sigma_v}\right)}$. Therefore, in

the presence of sample selection or attrition with short panels, regressing y_{it} on \mathbf{x}_{it} using only the observed information is invalidated by two problems – first, the presence of the unobserved effects α_i which introduces the incidental parameter problem, and second, the “selection bias” arising from the fact that

$$E(u_{it} | \mathbf{x}_{it}, d_{it} = 1) = \lambda(\eta_i + \mathbf{w}_{it}' \mathbf{a}).$$

The presence of individual-specific effects in (7.3.1) is easily solved by time differencing those individuals that are observed for two time periods t and s , i.e., who have $d_{it} = d_{is} = 1$. However, the sample selectivity factors are not eliminated by time differencing. However, conditional on given i , if (u_{it}, v_{it}) are stationary and $\mathbf{w}_{it}' \mathbf{a} = \mathbf{w}_{is}' \mathbf{a}$, $\lambda(\eta_i + \mathbf{w}_{it}' \mathbf{a}) = \lambda(\eta_i + \mathbf{w}_{is}' \mathbf{a})$. Then the difference of (7.4.32) between t and s if both y_{it} and y_{is} are observable no longer contains the individual-specific effects, α_i , and the selection factor, $\lambda(\eta_i + \mathbf{w}_{it}' \mathbf{a})$,

$$\Delta y_{its} = y_{it} - y_{is} = (\mathbf{x}_{it} - \mathbf{x}_{is})' \beta + (\epsilon_{it} - \epsilon_{is}) = \Delta \mathbf{x}_{its}' \beta + \Delta \epsilon_{its}. \quad (7.4.33)$$

As shown by Ahn and Powell (1993), if λ is a sufficiently “smooth” function, and $\hat{\mathbf{a}}$ is a consistent estimator of \mathbf{a} , observations for which the difference $(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}$ is close to zero should have $\lambda_{it} - \lambda_{is} \simeq 0$. Therefore, Kyriazidou (1997) generalizes the pairwise difference concept of Ahn and Powell (1993) and proposes to estimate the fixed-effects sample selection models in two steps: In the first step, estimate \mathbf{a} by either the Andersen (1970) and Chamberlain (1980) conditional maximum likelihood approach or the Horowitz (1992) and Lee (1999) smoothed version of the Manski (1975) maximum score method discussed in Chapter 6. In the second step, the estimated $\hat{\mathbf{a}}$ is used to estimate β based on pairs of observations for which $d_{it} = d_{is} = 1$ and for which $(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}$ is “close” to zero. This last requirement is operationalized by weighting each pair of observations with a weight that depends inversely on the magnitude of $(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}$, so that pairs with larger differences in the selection effects receive less weight in the estimation. The Kyriazidou (1997) estimator takes the form:

$$\hat{\beta}_K = \left\{ \sum_{i=1}^N \frac{1}{T_i - 1} \sum_{1 \leq s < t \leq T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{x}_{it} - \mathbf{x}_{is})' K \left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}}{h_N} \right] d_{it} d_{is} \right\}^{-1} \left\{ \sum_{i=1}^N \frac{1}{T_i - 1} \sum_{1 \leq s < t \leq T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(y_{it} - y_{is}) K \left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}}{h_N} \right] d_{it} d_{is} \right\} \quad (7.4.34)$$

⁸ Linear panel data with randomly missing data will be discussed in Section 9.6.

where T_i denotes the number of positively observed y_{it} for the i th individual, $K(\cdot)$ is a kernel density function which tends to zero as the magnitude of its argument increases, and h_N is a positive constant or bandwidth that decreases to zero as $N \rightarrow \infty$. The effect of multiplying the Kernel function $K(\cdot)$ is to give more weight to observations with $\frac{1}{h_N}(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}} \simeq 0$ and less weight to those observations that $\mathbf{w}_{it}'\hat{\mathbf{a}}$ is different from $\mathbf{w}_{is}'\hat{\mathbf{a}}$ so that in the limit only observations with $\mathbf{w}_{it}'\mathbf{a} = \mathbf{w}_{is}'\mathbf{a}$ are used in (7.4.34). Under appropriate regularity conditions (7.4.34) is consistent, but the rate of convergence is proportional to $\sqrt{Nh_N}$, much slower than the standard square root of the sample size.

When $T = 2$, the asymptotic covariance matrix of the Kyriazidou (1997) estimator (7.4.34) may be approximated by the Eicker (1963)–White (1980) formulae of the asymptotic covariance matrix of the least squares estimator for the linear regression model with heteroscedasticity,

$$\left(\sum_{i=1}^N \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i' \right)^{-1} \sum_{i=1}^N \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i' \Delta \hat{e}_i^2 \left(\sum_{i=1}^N \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i' \right)^{-1}, \quad (7.4.35)$$

where $\hat{\mathbf{x}}_i = K\left(\frac{\Delta \mathbf{w}_i' \hat{\mathbf{a}}}{h_N}\right)^{1/2} \Delta \mathbf{x}_i (d_{i2} d_{i1})$ and $\Delta \hat{e}_i$ is the estimated residual of (7.4.33).

In the case that only a truncated sample is observed, the first-stage estimation of $\hat{\mathbf{a}}$ cannot be implemented. However, a sufficient condition to ensure only observations with $\Delta \mathbf{w}_{its}' \mathbf{a} = 0$ are used is to replace $K\left[\frac{\Delta \mathbf{w}_{its}' \hat{\mathbf{a}}}{h_N}\right]$ by a multivariate Kernel function $K\left(\frac{\mathbf{w}_{it} - \mathbf{w}_{is}}{h_N}\right)$ in (7.4.34). However, the speed of convergence of (7.4.34) to the true $\boldsymbol{\beta}$ will be $\sqrt{Nh_N^k}$, where k denotes the dimension of \mathbf{w}_{it} . This is much slower speed than $\sqrt{Nh_N}$ since h_N needs to converge to zero as $N \rightarrow \infty$.

7.5 AN EXAMPLE: HOUSING EXPENDITURE

Charlier, Melenberg, and van Soest (2001) use Dutch Socio-Economic Panel (SEP) 1987–89 waves to estimate the following endogenous switching regression model for the share of housing expenditure in total expenditure:

$$d_{it} = 1(\mathbf{w}_{it}' \mathbf{a} + \eta_i + v_{it} > 0), \quad (7.5.1)$$

$$y_{1it} = \boldsymbol{\beta}'_1 \mathbf{x}_{it} + \alpha_{1i} + u_{1it}, \text{ if } d_{it} = 1, \quad (7.5.2)$$

$$y_{2it} = \boldsymbol{\beta}'_2 \mathbf{x}_{it} + \alpha_{2i} + u_{2it}, \text{ if } d_{it} = 0, \quad (7.5.3)$$

where d_{it} denotes the tenure choice between owning and renting, with 1 for owners and 0 for renters; y_{1it} and y_{2it} are the budget shares spent on housing for owners and renters, respectively; \mathbf{w}_{it} and \mathbf{x}_{it} are vectors of explanatory variables; η_i, α_{1i} , and α_{2i} are unobserved household specific effects; and v_{it}, u_{1it} , and u_{2it} are the error terms. The budget share spent on housing is defined as the fraction of total expenditure spent on housing. Housing expenditure for renters is just the rent paid by a family. The owners' expenditure on housing consists of net interest costs on mortgages, net rent paid if the land is not owned, taxes on owned housing, costs of insuring the house, opportunity cost of housing equity (which is set at 4% of the value of house minus the mortgage value), and maintenance cost, minus the increase of the value of the house. The explanatory variables considered are the education level of the head of household (DOP), age of the head of the household (AGE), age squared (AGE2), marital status (DMAR), logarithm of monthly family income (LINC) and its square (L2INC), monthly total family expenditure (EXP), logarithm of monthly total family expenditure (LEXP) and its square (L2EXP), number of children

(NCH), logarithm of constant quality price of rental housing (LRP), logarithm of constant quality price of owner occupied housing after tax (LOP), and LRP-LOP. The variables that are excluded from the tenure choice equation (7.5.1) are DOP, LEXP, L2EXP, LRP, and LOP. The variables excluded from the budget share equations (7.5.2 and 7.5.3) are DOP, LINC, L2INC, EXP, NCH, and LRP-LOP.

The random-effects and fixed-effects models with and without selection are estimated. However, since \mathbf{x} includes LEXP and L2EXP, and they could be endogenous, Charlier et al. (2001) also estimate this model by the instrumental variable (IV) method. For instance, the Kyriazidou (1997) weighted least squares estimator is modified as:

$$\hat{\beta}_{KN} = \left\{ \sum_{i=1}^N \sum_{1 \leq s < t \leq T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{z}_{it} - \mathbf{z}_{is})' K \left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}}{h_N} \right] d_{it} d_{is} \right\}^{-1} \cdot \left\{ \sum_{i=1}^N \sum_{1 \leq s < t \leq T_i} (\mathbf{z}_{it} - \mathbf{z}_{is})(y_{it} - y_{is}) K \left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})' \hat{\mathbf{a}}}{h_N} \right] d_{it} d_{is} \right\}, \quad (7.5.4)$$

to take account of the potential endogeneity issue of LEXP and L2EXP, where \mathbf{z}_{it} is a vector of instruments.

Tables 7.2 and 7.3 present the fixed effects and random effects estimation results for the budget share equations without and with correction for selection, respectively. The Kyriazidou (1997) estimator is based on the first-stage logit estimation of the tenure choice equation (7.5.1). The random-effects estimator is based on Newey's (2009) series

Table 7.2. *Estimation results for the budget share equations without correction for selection (standard errors in parentheses)^a*

Variable	Pooled random effects	Pooled IV random effects	Linear model fixed effects	Linear model IV ^b fixed effects
<i>Owners</i>				
Constant	4.102** (0.238)	4.939** (0.712)		
AGE	0.045** (0.009)	0.029** (0.010)	-0.073 (0.041)	-0.063 (0.044)
AGE2	-0.005** (0.001)	-0.003** (0.001)	0.009** (0.004)	0.009* (0.004)
LEXP	-0.977** (0.059)	-1.271** (0.178)	-0.769** (0.049)	-1.345** (0.269)
L2EXP	0.052** (0.003)	0.073** (0.011)	0.036** (0.003)	0.070** (0.016)
DMAR	0.036** (0.004)	0.027** (0.005)		
Dummy87			-0.001 (0.003)	-0.000 (0.004)
Dummy88			-0.002 (0.001)	-0.001 (0.002)
LOP	0.068** (0.010)	0.108** (0.010)	0.065** (0.016)	0.050** (0.018)
<i>Renters</i>				
Constant	2.914** (0.236)	3.056** (0.421)		
AGE	0.038** (0.007)	0.027** (0.007)	0.114** (0.034)	0.108** (0.035)
AGE2	-0.004** (0.000)	-0.003** (0.001)	-0.009* (0.004)	-0.009* (0.004)
LEXP	-0.772** (0.055)	-0.820** (0.106)	-0.800** (0.062)	-0.653** (0.219)
L2EXP	0.040** (0.003)	0.045** (0.006)	0.039** (0.004)	0.031* (0.014)
DMAR	0.011** (0.002)	0.001** (0.003)		
Dummy87			-0.004 (0.003)	-0.003 (0.003)
Dummy88			-0.002 (0.002)	-0.002 (0.002)
LRP	0.119* (0.017)	0.112** (0.017)	0.057** (0.020)	0.060** (0.020)

^a *Means significant at the 5% level. ** Means significant at the 1% level.

^b In IV estimation AGE, AGE2, LINC, L2INC, Dummy87, Dummy88, and either LOP (for owners) or LRP (for renters) are used as instruments.

Source: Charlier, Melenberg, and van Soest (2001, Table 3).

Table 7.3. Estimation results for the budget share equations using panel data models taking selection into account (standard errors in parentheses)^a

Variable	Pooled random effects ^b	Pooled IV random effects ^c	Kyriazidou OLS estimates	Kyriazidou IV ^d estimates
<i>Owners</i>				
Constant	2.595 ^e	3.370 ^e		
AGE	−0.040** (0.013)	−0.020 (0.015)	0.083 (0.083)	0.359** (0.084)
AGE2	0.004** (0.001)	0.002 (0.001)	−0.008 (0.008)	−0.033** (0.009)
LEXP	−0.594** (0.142)	−0.821 (0.814)	−0.766** (0.102)	−0.801** (0.144)
L2EXP	0.026** (0.008)	0.042 (0.050)	0.036** (0.006)	0.036** (0.008)
DMAR	0.006 (0.007)	0.012 (0.007)		
LOP	0.126** (0.012)	0.121** (0.011)	0.006 (0.030)	0.001 (0.029)
Dummy87			−0.006 (0.007)	−0.013 (0.007)
Dummy88			−0.004 (0.004)	−0.008 (0.004)
<i>Renters</i>				
Constant	2.679 ^d	1.856 ^d		
AGE	−0.037** (0.012)	−0.027* (0.012)	0.127* (0.051)	0.082 (0.080)
AGE2	0.004** (0.001)	0.003* (0.001)	−0.018** (0.006)	−0.014 (0.007)
LEXP	−0.601** (0.091)	−0.417 (0.233)	−0.882** (0.087)	−0.898** (0.144)
L2EXP	0.027** (0.005)	0.016 (0.015)	0.044** (0.005)	0.044** (0.009)
DMAR	−0.021** (0.005)	−0.019** (0.005)		
LRP	0.105** (0.016)	0.106** (0.016)	0.051 (0.028)	0.024 (0.030)
Dummy87			−0.024** (0.007)	−0.023 (0.013)
Dummy88			−0.009* (0.004)	−0.012 (0.007)

^a * Means significant at the 5% level. ** Means significant at the 1% level.

^b Series approximation using single index ML probit in estimating the selection equation.

^c IV using AGE, AGE2, LINC, L2INC, DMAR and either LOP (for owners) or LRP (for renters) as instruments.

^d In IV estimation AGE, AGE2, LINC, L2INC, Dummy87, and Dummy88 are used as instruments.

^e Estimates include the estimate for the constant term in the series approximation.

Source: Charlier, Melenberg, and van Soest (2001, Table 4).

expansion method (Charlier et al. 2000). The differences among these different formulations are quite substantial. For instance, the parameters related to AGE, AGE2, LEXP, L2EXP, and the prices are substantially different from their random-effects counterparts based on IV. They also lead to very different conclusions on the elasticities of interest. The price elasticities for the average renters and owners are about −0.5 in the random-effects model, but are close to −1 for owners and −0.8 for renters in the fixed-effects models.

The Hausman-type specification tests of endogeneity of LEXP and L2EXP are inconclusive. But a test for the presence of selectivity bias based on the difference between the Kyriazidou IV and linear panel data estimates have test statistics of 88.2 for owners and 23.7 for renters which are significant at the 5% level for the chi-square distribution with seven degrees of freedom. This indicates that the model that does not allow for correlation between the error terms in the share equations (7.5.2 and 7.5.3) and the error term in the selection equation (7.5.1) is probably misspecified.

The Hausman (1978)-type specification test of no correlation between the household specific effects and the x 's based on the difference between the Newey IV and the Kyriazidou IV estimates have test statistics of 232.1 for owners and 37.8 for renters. These are significant at the 5% level for the chi-square distribution with five degrees of freedom, thus rejecting the random-effects model that does not allow for correlation between the household-specific effects and the explanatory variables. These results indicate that the random-effects linear panel models or linear panel data models which allow for only very specific selection mechanisms (both of which can be estimated with just the cross-sectional data) are probably too restrictive.

7.6 DYNAMIC TOBIT MODELS

7.6.1 Dynamic Censored Models

In the section we consider dynamic Tobit models in which the observed y_{it} takes the form⁹

$$y_{it} = \begin{cases} y_{it}^*, & \text{if } y_{it}^* > 0, \\ 0, & \text{if } y_{it}^* \leq 0. \end{cases} \quad (7.6.1)$$

There could be two types of dynamic dependence for y_{it}^* :

$$y_{it}^* = \gamma y_{i,t-1}^* + \beta' x_{it} + \alpha_i + u_{it}, \quad (7.6.2)$$

or

$$y_{it}^* = \gamma y_{i,t-1} + \beta' x_{it} + \alpha_i + u_{it}, \quad (7.6.3)$$

where the error term u_{it} is independently distributed over i and independently identically distributed over t (i.e., we allow $\text{Var}(u_{it}) = \sigma_i^2$).

For model (7.6.2), when $y_{i,t-1} = 0$, $y_{i,t-1}^*$ could be any value between $-\infty$ and 0. If there are no individual-specific effects α_i (or $\alpha_i = 0$ for all i), panel data actually allow the possibility of ignoring the censoring effects in the lagged dependent variables by concentrating on the subsample where $y_{i,t-1} > 0$. Since if $y_{i,t-1} > 0$, $y_{i,t-1} = y_{i,t-1}^*$, (7.6.1) and (7.6.2) with $\alpha_i = 0$ become

$$\begin{aligned} y_{it}^* &= \gamma y_{i,t-1}^* + \beta' x_{it} + u_{it} \\ &= \gamma y_{i,t-1} + \beta' x_{it} + u_{it}. \end{aligned} \quad (7.6.4)$$

Thus, by treating $y_{i,t-1}$ and x_{it} as predetermined variables that are independent of the error, u_{it} , the censored estimation techniques for the static model discussed in Section 7.1 can be applied to the subsample where (7.6.4) holds.

When random individual-specific effects α_i are present in (7.6.2), y_{it}^* and α_i are correlated for all s even if α_i can be assumed to be uncorrelated with x_i . To implement the MLE approach, not only does one have to make assumptions on the distribution of individual effects and initial observations, but computation may become unwieldy. To reduce the computational complexity, Arellano, Bover, and Labeaga (1999) suggest a two-step approach. The first step estimates the reduced form of y_{it}^* by projecting y_{it}^* on all previous $y_{i0}^*, y_{i1}^*, \dots, y_{i,t-1}^*$ and x_{i1}, \dots, x_{it} . The second step estimates (γ, β') from the reduced form parameters of the y_{it}^* equation, π_t , by a minimum distance estimator of the form (2.9.14). To avoid the censoring problem in the first step, they suggest that for the i th individual, only the string $(y_{is}, y_{i,s-1}, \dots, y_{i0})$ where $y_{i0} > 0, \dots, y_{i,s-1} > 0$ is used. However, in order to derive the estimates of π_t , the conditional distribution of y_{it}^* given $y_{i0}^*, \dots, y_{i,t-1}^*$ will have to be assumed. Moreover, the reduced form parameters π_t are related to (γ, β') in a highly nonlinear way. Thus, the second-stage estimator is not easily derivable. Therefore, in this section we shall bypass the issue of fixed or random α_i and only discuss the Honoré (1993) and Hu (2002) trimmed estimator.

For model (7.6.2), if $y_{i,t-1} = 0$ (i.e., $y_{i,t-1}^* < 0$), there is no one-to-one correspondence between u_{it} and y_{it}^* given $(y_{i,t-1}, x_{it}, \alpha_i)$. On the other hand, for model (7.6.3) there is still a one-to-one correspondence between u_{it} and y_{it}^* given $(y_{i,t-1}, x_{it}, \alpha_i)$, whether $y_{i,t-1} = 0$ or > 0 . Therefore, we may split the observed sample for model (7.6.2) into two groups. For the group where $y_{i,t-1} = 0$, the estimation method discussed in Section 7.5 can be

⁹ See Honoré (1993) for a discussion of the model $y_{it}^* = \gamma y_{i,t-1} + \beta' x_{it} + \alpha_i + u_{it}$.

$$E \left\{ (y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B \right\} \cdot \left[1 \left\{ y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}' \Delta \mathbf{x}_{i2} > 0 \right\} \right. \\ \left. - 1 \left\{ y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}' \Delta \mathbf{x}_{i2} < 0 \right\} \right] = 0. \quad (7.6.6)$$
$$E \left[1 \left\{ (y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B \right\} (y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}' \Delta \mathbf{x}_{i2}) \right] = 0. \quad (7.6.7)$$
$$E \left[1 \{ (y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B \} \cdot \xi(y_{i1} - \gamma y_{i0} + \boldsymbol{\beta}' \Delta \mathbf{x}_{i2}, \tilde{y}_{i2}(\gamma)) \right. \\ \left. \cdot h(y_{i0}, \mathbf{x}_{i1}, \mathbf{x}_{i2}) \right] = 0, \quad (7.6.8)$$
$$\begin{aligned}
1\{(y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B\} &\equiv 1\{y_{i0} > 0, y_{i1} > 0, y_{i2} > 0\} \\
&\cdot \left[1\{\gamma y_{i0} - \boldsymbol{\beta}' \Delta \mathbf{x}_{i2} > 0\} \cdot 1\{y_{i1} > \gamma y_{i0} \right. \\
&\quad \left. - \boldsymbol{\beta}' \Delta \mathbf{x}_{i2} - \gamma(\tilde{y}_{i2}(\gamma) + \gamma y_{i0} - \boldsymbol{\beta}' \Delta \mathbf{x}_{i2})\} \right. \\
&\cdot 1\{\tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}' \Delta \mathbf{x}_{i2}\} + 1\{\gamma y_{i0} - \boldsymbol{\beta}' \Delta \mathbf{x}_{i2} < 0\} \\
&\cdot 1\{\tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}' \Delta \mathbf{x}_{i2}\} \left. \right]. \tag{7.6.9}
\end{aligned}$$
$$\begin{aligned} W_{its}(\boldsymbol{\beta}', \gamma) &= \max \{0, (\mathbf{x}_{it} - \mathbf{x}_{is})' \boldsymbol{\beta}, y_{it} - \gamma y_{i,t-1}\} - \mathbf{x}_{it}' \boldsymbol{\beta} \\ &= \max \{-\mathbf{x}_{it}' \boldsymbol{\beta}, -\mathbf{x}_{is}' \boldsymbol{\beta}, \alpha_i + u_{it}\}, \end{aligned} \quad (7.6.10)$$
$$\begin{aligned} W_{ist}(\boldsymbol{\beta}', \gamma) &= \max \{0, (\mathbf{x}_{is} - \mathbf{x}_{it})' \boldsymbol{\beta}, y_{is} - \gamma y_{i,s-1}\} - \mathbf{x}'_{is} \boldsymbol{\beta} \\ &= \max \{-\mathbf{x}'_{is} \boldsymbol{\beta}, -\mathbf{x}'_{it} \boldsymbol{\beta}, \alpha_i + u_{is}\}, \end{aligned} \quad (7.6.11)$$
$$E \left[1 \{ y_{it-1} > 0, y_{it} > 0, y_{i,s-1} > 0, y_{is} > 0 \} \right. \\ \left. \cdot \xi(W_{its}(\boldsymbol{\beta}', \gamma), W_{ist}(\boldsymbol{\beta}', \gamma)) \cdot h(\mathbf{x}_{it}, \mathbf{x}_{is}) \right] = 0, \quad (7.6.12)$$

for any function $h(\cdot)$. When $T \geq 3$, the symmetric trimming procedure (7.6.12) requires weaker assumptions than the one based on three consecutive uncensored observations since the conditioning variables do not involve the initial value y_{i0} . However, this approach also leads to more severe trimming.

Based on the orthogonality conditions (7.6.8) or (7.6.12), Hu (2002) suggests a GMM estimator of $\theta = (\beta', \gamma')'$ by minimizing $\mathbf{m}_N(\theta)' A_N \mathbf{m}_N(\theta)$ where $\mathbf{m}_N(\theta)$ is the sample analog of (7.6.8) or (7.6.12), and A_N is a positive definite matrix that converges to a constant matrix A as $N \rightarrow \infty$. The GMM estimator will have the limiting distribution of the form

$$\sqrt{N}(\hat{\theta}_{GMM} - \theta) \longrightarrow N(\mathbf{0}, (\Gamma' \Lambda \Gamma)^{-1} [\Gamma' A V A \Gamma] (\Gamma' A \Gamma)^{-1}), \quad (7.6.13)$$

where $\Gamma = \frac{\partial}{\partial \theta} E[\mathbf{m}(\theta)]$, $V = E[\mathbf{m}(\theta) \mathbf{m}(\theta)']$. When the optimal weighting matrix $A = V^{-1}$ is used, the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta}_{GMM} - \theta)$ becomes $(\Gamma' V^{-1} \Gamma)^{-1}$.

However, the true value of θ is not the only value that satisfies the orthogonality conditions (7.6.6)–(7.6.8) or (7.6.12). For instance, those orthogonality conditions can be trivially satisfied when the parameter values are arbitrarily large. To see this, note that for a given value of γ , when the value of $\delta_{it} = \mathbf{x}_{it}' \beta$ goes to infinity, the number of observations falling in the (nontruncated) region $A \cup B$ in Figures 7.8 and 7.9 approaches zero. Thus, the moment conditions can be trivially satisfied. To overcome this possible lack of identification of GMM estimates based on the minimization of the criterion function, Hu (2002) suggests using a subset of the moments that exactly identify β for given γ to provide the estimates of β , then testing whether the rest of the moment conditions are satisfied by these estimates for a sequence of γ values ranging from 0 to 0.9 with an increment of 0.01. Among the values of γ at which the test statistics are not rejected, the one which yields the smallest test statistic is chosen as the estimate of γ . Hu (2002) uses this estimation method to study earnings dynamics, using matched data from the Current Population Survey and Social Security Administration (CPS-SSA) Earnings Record for a sample of men who were born in 1930–1939 and living in the South during the period of 1957–1973. The SSA earnings are top-coded at the maximum Social Security taxable level, namely, $y_{it} = \min(y_{it}^*, c_t)$, where c_t is the Social Security maximum taxable earnings level in period t . This censoring at the top can be easily translated into censoring at zero by considering $\tilde{y}_{it} = c_t - y_{it}$, then $\tilde{y}_{it} = \max(0, c_t - y_{it}^*)$.

Table 7.4 presents the estimates of the coefficient of the lagged log real annual earnings coefficient of an AR(1) model based on a sample of 226 black and 1883 white men with and without correction for censoring. When censoring is ignored, the model is estimated by the linear GMM method. When censoring is taken into account, Hu uses an unbalanced panel of observations with positive SSA earnings in three consecutive time periods.

Table 7.4. *Estimates of AR(1) coefficients of log real annual earnings (in thousands)^a*

Linear GMM (assuming no censoring)		Nonlinear GMM with correction for censoring	
Black	White	Black	White
0.379 (0.030)	0.399 (0.018)	0.210 (0.129)	0.380 (0.051)

^a Standard errors in parenthesis.
Source: Hu (2002).

The estimated γ are very similar for black and white men when censoring is ignored. However, when censoring is taken into account, the estimated autoregressive parameter γ is much higher for white men than for black men. The higher persistence of the earnings process for white men than for black men is consistent with the notion that white men held jobs that had better security and were less vulnerable to economic fluctuation than black men during the period 1957–1973.

7.6.2 Dynamic Sample Selection Models

When the selection rule is endogenously determined as given by (7.2.4) and y_{it}^* is given by (7.6.2) or (7.6.3), with \mathbf{w}_{it} and \mathbf{x}_{it} being nonoverlapping vectors of strictly exogenous explanatory variables (with possibly common elements), the model under consideration has the form:¹⁰

$$y_{it} = d_{it}y_{it}^*, \quad (7.6.14)$$

$$d_{it} = 1\{\mathbf{w}'_{it}\mathbf{a} + \eta_i + v_{it}\}, \quad \begin{matrix} i = 1, \dots, N, \\ t = 1, \dots, T, \end{matrix} \quad (7.6.15)$$

where $(d_{it}, \mathbf{w}_{it})$ is always observed, and $(y_{it}^*, \mathbf{x}_{it})$ is observed only if $d_{it} = 1$. For notational ease, we assume that d_{i0} and y_{i0} are also observed.

In the static case of $\gamma = 0$, Kyriazidou (1997) achieves the identification of β by replying on the conditional pairwise exchangeability of the error vector (u_{it}, v_{it}) given the entire path of the exogenous variables $(\mathbf{x}_i, \mathbf{w}_i)$ and the individual effects (α_i, η_i) . However, the consistency of the Kyriazidou estimator (7.4.34) breaks down in the presence of the lagged dependent variable in (7.6.2) or (7.6.3). The reason is the same as in linear dynamic panel data models where first differencing generates nonzero correlation between $y_{i,t-1}^*$ and the transformed error term (see Chapter 3). However, just as in the linear case, estimators based on linear and nonlinear moment conditions on the correlation structure of the unobservables with the observed variables can be used to obtain consistent estimators of γ and β .

Under the assumption that $\{u_{it}, v_{it}\}$ are independently identically distributed over time for all i conditional on $\xi_i \equiv (\mathbf{w}'_i, \alpha_i, \eta_i, y_{i0}^*, d_{i0})$, where $\mathbf{w}_i = (\mathbf{w}'_{i1}, \dots, \mathbf{w}'_{iT})'$, Kyriazidou (2001) notes that by conditioning on the event that $\Delta \mathbf{w}'_{it}\mathbf{a} = 0$, the following moment conditions hold:¹¹

$$E(d_{it}d_{i,t-1}d_{i,t-2}d_{i,t-j}y_{i,t-j}\Delta u_{it} \mid \Delta \mathbf{w}'_{it}\mathbf{a} = 0) = 0, j = 2, \dots, t, \quad (7.6.16)$$

and

$$E(d_{is}d_{it}d_{i,t-1}d_{i,t-2}\mathbf{x}_{is}\Delta u_{it} \mid \Delta \mathbf{w}'_{it}\mathbf{a} = 0) = 0, \text{ for } t = 2, \dots, T; s = 1, \dots, T. \quad (7.6.17)$$

This is because for an individual i when the selection index $\mathbf{w}'_{it}\mathbf{a} = \mathbf{w}'_{i,t-1}\mathbf{a}$, the magnitude of the sample selection effects in the two periods, $\lambda(\eta_i + \mathbf{w}'_{it}\mathbf{a})$ and $\lambda(\eta_i + \mathbf{w}'_{i,t-1}\mathbf{a})$, will also be the same. Thus, by conditioning on $\Delta \mathbf{w}'_{it}\mathbf{a} = 0$, the sample selection effects and the individual effects are eliminated by first differencing.

¹⁰ The assumption that \mathbf{x}_{it} and \mathbf{w}_{it} do not coincide rules out the censored regression model as a special case of (7.6.14) and (7.6.15).

¹¹ Kyriazidou (2001) shows that these moment conditions also hold if $d_{it}^* = \phi d_{i,t-1} + \mathbf{w}'_{it}\mathbf{a} + \eta_i + v_{it}$.

Let $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}')'$, $\mathbf{z}'_{it} = (y_{i,t-1}, \mathbf{x}'_{it})$, and

$$m_{1it}(\theta) = d_{it}d_{i,t-1}d_{i,t-2}d_{i,t-j}y_{i,t-j}(\Delta y_{it} - \Delta \mathbf{z}'_{it}\theta), t = 2, \dots, T; j = 2, \dots, t, \quad (7.6.18)$$

$$m_{2it,k}(\boldsymbol{\theta}) = d_{is}d_{it}d_{i,t-1}d_{i,t-2}x_{is,k}(\Delta y_{it} - \Delta z'_{it}\boldsymbol{\theta}),$$

$$t = 2, \dots, T; s = 1, \dots, T; k = 1, \dots, K. \quad (7.6.19)$$

Kyriazidou (2001) suggests a kernel weighted generalized method of moments estimator (KGMM) that minimizes the following quadratic form

$$\hat{G}_N(\boldsymbol{\theta})' A_N \hat{G}_N(\boldsymbol{\theta}), \quad (7.6.20)$$

where A_N is a stochastic matrix that converges in probability to a finite nonstochastic limit A , and $\hat{G}_N(\theta)$ is the vector of stacked sample moments with rows of the form

$$\frac{1}{N} \sum_{i=1}^N \frac{1}{h_N} K \left(\frac{\Delta \mathbf{w}'_{it} \hat{\mathbf{a}}}{h_N} \right) m_{\ell it}(\boldsymbol{\theta}), \quad (7.6.21)$$

where m_{lit} denotes the elements of (7.6.18) and (7.6.19), $K(\cdot)$ is a kernel density function, $\hat{\mathbf{a}}$ is some consistent estimator of \mathbf{a} , and h_N is a bandwidth that shrinks to zero as $N \rightarrow \infty$. Under appropriate conditions, Kyriazidou (2001) proves that the KGMM estimator is consistent and asymptotically normal. The rate of convergence is the same as in univariate nonparametric density and regression function estimation, i.e., at the speed of $\sqrt{N h_N}$.