Analysis of Panel Data

Second Edition

CHENG HSIAO
University of Southern California

CAMBRIDGE
UNIVERSITY PRESS

CHAPTER 8

Truncated and Censored Data

8.1 INTRODUCTION

In economics, the ranges of dependent variables are often constrained in some way. 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 generalization are known as *Tobit* models because of their similarities to probit models.[1] In statistics they are known as truncated or censored regression models. The model is called *truncated* if the observations outside a specific range are totally lost, and is called *censored* if we can at least observe some of the explanatory variables.

Consider a latent response function,

$$y^* = \boldsymbol{\beta}'\mathbf{x} + u, \tag{8.1.1}$$

where $\mathbf{x}$ is a $K \times 1$ vector of exogenous variables and u is the error term that is independently, identically distributed (i.i.d.) with mean 0 and variance σ_u^2. Without loss of generality, suppose that the observed y are related to y^* by

$$y = \begin{cases} y^* & \text{if} \quad y^* > 0, \\ 0 & \text{if} \quad y^* \leq 0. \end{cases} \tag{8.1.2}$$

Models of the form (8.1.1) and (8.1.2) are called censored regression models because the data consist of those points of the form $(y_i^*, \mathbf{x}_i)$ if $y_i^* > 0$ and $(0, \mathbf{x}_i)$ if $y_i^* \leq 0$ for $i = 1, \ldots, N$. The truncated data only consist of points of the form $(y_i^*, \mathbf{x}_i)$ where $y_i^* > 0$.

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

$$E(y \mid y > 0) = E(y^* \mid y^* > 0) = \mathbf{x}'\boldsymbol{\beta} + E(u \mid u > -\mathbf{x}'\boldsymbol{\beta}). \tag{8.1.3}$$

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

$$\begin{aligned} E(y \mid \mathbf{x}) &= \text{Prob}(y = 0) \cdot 0 + \text{Prob}(y > 0 \mid \mathbf{x}) \cdot E(y \mid y > 0, \mathbf{x}) \\ &= \text{Prob}(u \leq -\mathbf{x}'\boldsymbol{\beta}) \cdot 0 + \text{Prob}(u > -\mathbf{x}'\boldsymbol{\beta}) E(y^* \mid \mathbf{x}; u > -\mathbf{x}'\boldsymbol{\beta}) \\ &= \text{Prob}(u > -\mathbf{x}'\boldsymbol{\beta})[\mathbf{x}'\boldsymbol{\beta} + E(u \mid u > -\mathbf{x}'\boldsymbol{\beta})]. \end{aligned} \tag{8.1.4}$$

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

$$\text{Prob}(u > -\mathbf{x}'\boldsymbol{\beta}) = 1 - \Phi\left(\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right) = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right), \tag{8.1.5}$$

and

$$E(u \mid u > -\mathbf{x}'\boldsymbol{\beta}) = \sigma_u \cdot \frac{\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right)}{\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right)}, \tag{8.1.6}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are standard normal density and cumulative (or integrated) normal, respectively. Equations (8.1.3) and (8.1.4) show that truncation or censoring of the dependent variables introduces dependence between the error term and the regressors for the model

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon, \tag{8.1.7}$$

where the error

$$\epsilon = \nu + E(y \mid \mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}. \tag{8.1.8}$$

Although $\nu = y - E(y \mid \mathbf{x})$ has $E(\nu \mid \mathbf{x}) = 0$, we have $E(\epsilon \mid \mathbf{x}) \neq 0$. Therefore, the least-squares estimator of (8.1.7) is biased and inconsistent.

For a sample of N independent individuals, the likelihood function of the truncated data is equal to

$$L_1 = \prod_1 [\text{Prob}(y_i > 0 \mid \mathbf{x}_i)]^{-1} f(y_i), \tag{8.1.9}$$

where $f(\cdot)$ denotes the density of y_i^* (or u_i), and $\prod_1$ means the product over those i for which $y_i > 0$. The likelihood function of the censored data is equal to

$$\begin{aligned} L_2 &= \left\{\prod_0 \text{Prob}(y_i = 0 \mid \mathbf{x}_i) \cdot \prod_1 \text{Prob}(y_i > 0 \mid \mathbf{x}_i)\right\} \\ &\quad \times \left\{\prod_1 [\text{Prob}(y_i > 0 \mid \mathbf{x}_i)]^{-1} f(y_i)\right\} \\ &= \prod_0 \text{Prob}(y_i = 0 \mid \mathbf{x}_i) \prod_1 f(y_i), \end{aligned} \tag{8.1.10}$$

where $\prod_0$ means the product over those i for which $y_i^* \leq 0$. In the case that u_i is independently normally distributed with mean 0 and variance σ_u^2, we have $f(y_i) = (2\pi)^{-\frac{1}{2}} \sigma_u^{-1} \exp\{-(1/2\sigma_u^2)(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\}$ and $\text{Prob}(y_i = 0 \mid \mathbf{x}_i) = \Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma_u) = 1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma_u)$.

Maximizing (8.1.9) or (8.1.10) with respect to $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \sigma_u^2)$ yields the maximum likelihood estimator (MLE). The MLE, $\hat{\boldsymbol{\theta}}$, is consistent and is asymptotically normally distributed. The asymptotic covariance matrix of the MLE, asy $\text{cov}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]$, is equal to the inverse of the information matrix, $[-E(1/N)\partial^2 \log L_j / \partial\boldsymbol{\theta}\, \partial\boldsymbol{\theta}']^{-1}$, which may be approximated by

$[-(1/N)\partial^2 \log L_j/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'\,|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}]^{-1}$, $j = 1, 2$. However, the MLE is highly nonlinear. A Newton–Raphson iterative scheme may have to be used to obtain the MLE. Alternatively, if u is normally distributed, Heckman (1976a) suggests the following two-step estimator:

1. Maximize the first factor in braces in the likelihood function (8.1.10) by probit MLE with respect to $\boldsymbol{\delta} = (1/\sigma_u)\boldsymbol{\beta}$, yielding $\hat{\boldsymbol{\delta}}$.
2. Substitute $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$ into the truncated model

$$\begin{aligned} y_i &= E(y_i \mid \mathbf{x}_i; y_i > 0) + \eta_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + \sigma_u \frac{\phi(\mathbf{x}_i'\boldsymbol{\delta})}{\Phi(\mathbf{x}_i'\boldsymbol{\delta})} + \eta_i \quad \text{for those } i \text{ such that } y_i > 0, \end{aligned} \tag{8.1.11}$$

where $E(\eta_i \mid \mathbf{x}_i) = 0$, $\text{Var}(\eta_i \mid \mathbf{x}_i) = \sigma_u^2[1 - (\mathbf{x}_i'\boldsymbol{\delta})\lambda_i - \lambda_i^2]$, and $\lambda_i = \phi(\mathbf{x}_i'\boldsymbol{\delta})/\Phi(\mathbf{x}_i'\boldsymbol{\delta})$. Regress y_i on $\mathbf{x}_i$ and $\phi(\mathbf{x}_i'\hat{\boldsymbol{\delta}})/\Phi(\mathbf{x}_i'\hat{\boldsymbol{\delta}})$ by least squares, using only the positive observations of y_i.

The Heckman two-step estimator is consistent. The formula for computing the asymptotic variance–covariance matrix of Heckman's estimator is given by Amemiya (1978b). But the Heckman two-step estimator is not as efficient as the MLE.

Both the MLE of (8.1.10) and the Heckman two-step estimator (8.1.11) are consistent only if u is independently normally distributed with constant variance. Of course, the idea of the MLE and the Heckman two-step estimator can still be implemented with proper modification if the identically distributed density function of u is correctly specified. A lot of times an investigator does not have the knowledge of the density function of u, or u is not identically distributed. Under the assumption that it is symmetrically distributed around 0, Powell (1986) proves that applying the least-squares method to the symmetrically censored or truncated data yields a consistent estimator which is robust to the assumption of the probability density function of u and heteroscedasticity of the unknown form.

The problem of censoring or truncation is that conditional on $\mathbf{x}$, y is no longer symmetrically distributed around $\mathbf{x}'\boldsymbol{\beta}$ even though u is symmetrically distributed around zero. Data points for which $u_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$ are either censored or omitted. However, we can restore symmetry by censoring or throwing away observations with $u_i \geq \mathbf{x}_i'\boldsymbol{\beta}$ or $y_i \geq 2\mathbf{x}_i'\boldsymbol{\beta}$, as shown in Figure 8.1, so that the remaining observations fall between $(0, 2\mathbf{x}'\boldsymbol{\beta})$. Because of the symmetry of u, the corresponding dependent variables are again symmetrically distributed about $\mathbf{x}'\boldsymbol{\beta}$ (Hsiao (1976)).

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 $u_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$ 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.) Because of the symmetry of the distribution

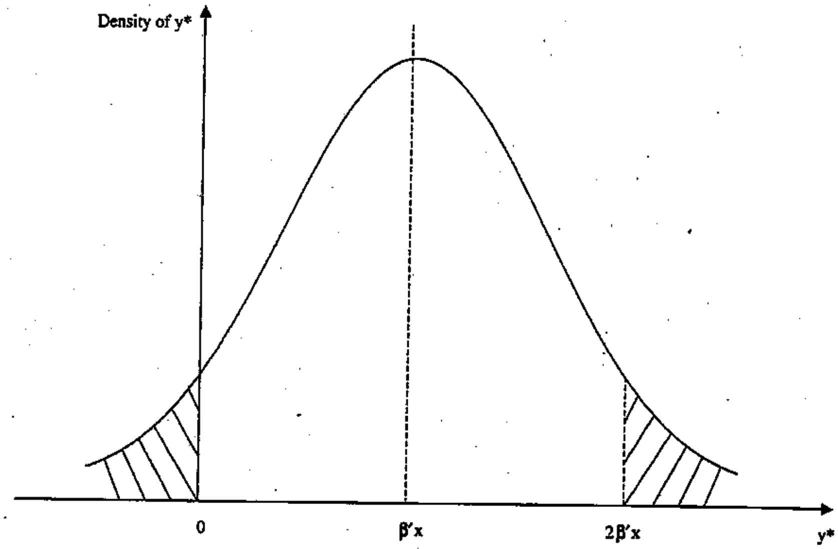

Fig. 8.1 Density of y^* censored or truncated at 0.

of u, 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 8.2. In other words, points b and c in Figure 8.2 are thrown away (point a is not observed).

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 8.2 is moved to the corresponding circle 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 8.2 have been moved to the corresponding circle points (a', b', c').

Applying the least-squares principle to the symmetrically trimmed truncated data is equivalent to requiring the observations falling in the region AOB to satisfy the following first-order condition:

$$\frac{1}{N}\sum_{i=1}^{N} 1(y_i < 2\boldsymbol{\beta}'\mathbf{x}_i)(y_i - \boldsymbol{\beta}'\mathbf{x}_i)\mathbf{x}_i = \mathbf{0}, \tag{8.1.12}$$

in the limit, where $1(A)$ denotes the indicator function of the event A, which takes the value 1 if A occurs and 0 otherwise. Applying the least-squares principle to the symmetrically censored data is equivalent to requiring the observations in the region AOB and the boundary OA and OB (the circle points in Figure 8.2) to satisfy the first-order condition,

$$\frac{1}{N}\sum_{i=1}^{N} 1(\boldsymbol{\beta}'\mathbf{x}_i > 0)(\min\{y_i, 2\boldsymbol{\beta}'\mathbf{x}_i\} - \boldsymbol{\beta}'\mathbf{x}_i)\mathbf{x}_i = \mathbf{0}, \tag{8.1.13}$$

in the limit. Therefore, Powell (1986) proposes the symmetrically trimmed least-squares estimator as the $\hat{\boldsymbol{\beta}}$ that minimizes

$$R_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{y_i - \max\left(\tfrac{1}{2}y_i, \mathbf{x}_i'\boldsymbol{\beta}\right)\right\}^2 \tag{8.1.14}$$

for the truncated data, and

$$\begin{aligned} S_N(\boldsymbol{\beta}) = {} & \sum_{i=1}^{N} \left\{y_i - \max\left(\tfrac{1}{2}y_i, \boldsymbol{\beta}'\mathbf{x}_i\right)\right\}^2 \\ & + \sum_{i=1}^{N} 1(y_i > 2\mathbf{x}'\boldsymbol{\beta})\left\{\left(\tfrac{1}{2}y_i\right)^2 - [\max(0, \mathbf{x}_i'\boldsymbol{\beta})]^2\right\} \end{aligned} \tag{8.1.15}$$

for the censored data. The motivation for $R_N(\boldsymbol{\beta})$ is that if $y > 2\boldsymbol{\beta}'\mathbf{x}$, it will have zero weight in the first-order condition (8.1.12) for the truncated sample. The motivation for $S_N(\boldsymbol{\beta})$ is that observations greater than $2\boldsymbol{\beta}'\mathbf{x}_i$ if $\boldsymbol{\beta}'\mathbf{x} > 0$ and all observations corresponding to $\mathbf{x}'\boldsymbol{\beta} < 0$ will have zero weight in the first-order condition (8.1.13) for the censored sample. Powell (1986) shows that the symmetrically trimmed least-squares estimator is consistent and asymptotically normally distributed as $N \to \infty$.

The exogenously determined limited-dependent-variable models can be generalized to consider a variety of endogenously determined sample selection issues. For instance, in Gronau (1976) and Heckman's (1976a) female-labor-supply model the hours worked are observed only for those women who decide to participate in the labor force. In other words, instead of being exogenously given, the truncating or censoring value is endogenously and stochastically

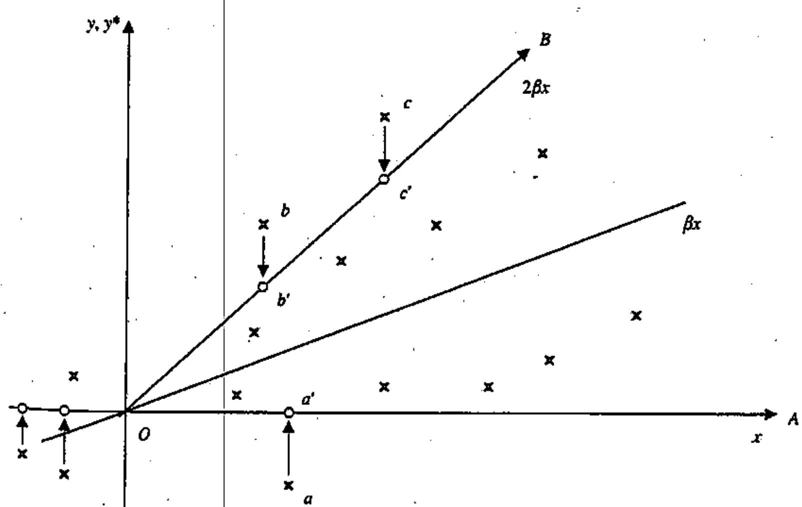

Fig. 8.2 Distribution of y and y^* under symmetric trimming.

determined by a selection equation (e.g. Duncan (1980))

$$d_i^* = \mathbf{w}_i'\mathbf{a} + v_i, \qquad i = 1, \ldots, N, \tag{8.1.16}$$

where $\mathbf{w}_i$ is a vector of exogenous variables, $\mathbf{a}$ is the parameter vector, and v_i is the random error term, assumed to be i.i.d. with mean 0 and variance normalized to 1. The samples (y_i, d_i), $i = 1, \ldots, N$, are related to y_i^* and d_i^* by the rule

$$d = \begin{cases} 1 & \text{if} \quad d^* > 0, \\ 0 & \text{if} \quad d^* \leq 0, \end{cases} \tag{8.1.17}$$

$$y = \begin{cases} y^* & \text{if} \quad d = 1, \\ 0 & \text{if} \quad d = 0. \end{cases} \tag{8.1.18}$$

The model (8.1.1), (8.1.16)–(8.1.18) is called the type II Tobit model by Amemiya (1985). Then

$$E(y_i \mid d_i = 1) = \mathbf{x}_i'\boldsymbol{\beta} + E(u_i \mid v_i > -\mathbf{w}_i'\mathbf{a}). \tag{8.1.19}$$

The likelihood function of (y_i, d_i) is

$$\begin{aligned} L &= \prod_c \operatorname{Prob}(d_i = 0) \prod_{\bar{c}} f(y_i^* \mid d_i = 1) \operatorname{Prob}(d_i = 1) \\ &= \prod_c \operatorname{Prob}(d_i = 0) \prod_{\bar{c}} \operatorname{Prob}(d_i^* > 0 \mid y_i) f(y_i), \end{aligned} \tag{8.1.20}$$

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

$$E(u \mid v > -\mathbf{w}'\mathbf{a}) = \sigma_{uv} \frac{\phi(\mathbf{w}'\mathbf{a})}{\Phi(\mathbf{w}'\mathbf{a})}, \tag{8.1.21}$$

$$\operatorname{Prob}(d = 0) = [1 - \Phi(\mathbf{w}'\mathbf{a})] = \Phi(-\mathbf{w}'\mathbf{a}), \tag{8.1.22}$$

$$\operatorname{Prob}(d_i = 1 \mid y_i) = \Phi\left\{\mathbf{w}'\mathbf{a} + \frac{\sigma_{uv}}{\sigma_u}(y - \mathbf{x}'\boldsymbol{\beta})\right\}. \tag{8.1.23}$$

Alternatively, Heckman's (1979) two-stage method can be applied: First estimate $\mathbf{a}$ by a probit MLE of d_i, $i = 1, \ldots, N$. Evaluate $\phi(\mathbf{a}'\mathbf{w}_i)/\Phi(\mathbf{a}'\mathbf{w}_i)$ using the estimated 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 as in 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 (8.1.1), (8.1.17), (8.1.18) conditional on $d_i = 1$ becomes a partially linear model of the form

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \lambda(\mathbf{w}_i) + \epsilon_i, \tag{8.1.24}$$

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

$$E(y_i \mid \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'E(\mathbf{x}_i \mid \mathbf{w}_i, d_i = 1) + \lambda(\mathbf{w}_i). \tag{8.1.25}$$

Subtracting (8.1.25) from (8.1.24) yields

$$y_i - E(y_i \mid \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'(\mathbf{x}_i - E(\mathbf{x}_i \mid \mathbf{w}_i, d_i = 1)) + \epsilon_i, \tag{8.1.26}$$

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

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \{&E[\mathbf{x} - E(\mathbf{x} \mid \mathbf{w})][\mathbf{x} - E(\mathbf{x} \mid \mathbf{w})]'\}^{-1} \\ &\times E[(\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}))][y - E(y \mid \mathbf{w})],\end{aligned} \tag{8.1.27}$$

using the truncated sample.

The first-stage conditional expectation for the estimator (8.1.27) 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)):

$$\hat{f}(y_a) = \frac{1}{Nh_N}\sum_{i=1}^{N} k\left(\frac{y_i - y_a}{h_N}\right), \tag{8.1.28}$$

where h_N is a positive number, called the *bandwidth* or *smoothing parameter*, that tends to zero as $N \to \infty$, and $k(u)$ 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

$$\hat{f}(\mathbf{w}_a) = \frac{1}{N\,|H_m|}\sum_{i=1}^{N} k_m\big(H_m^{-1}(\mathbf{w}_i - \mathbf{w}_a)\big), \tag{8.1.29}$$

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 one can have $k_m(\mathbf{u}) = \prod_{j=1}^{m} k(u_j)$, $\mathbf{u}' = (u_1, \ldots, u_m)$, $H_m = \operatorname{diag}(h_{1N}, \ldots, h_{mN})$.

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

$$\hat{f}(y_a \mid \mathbf{w}_a) = \frac{\hat{f}(y_a, \mathbf{w}_a)}{\hat{f}(\mathbf{w}_a)}, \tag{8.1.30}$$

and the conditional expectation by

$$E\hat{g}(y \mid \mathbf{w}_a) = \frac{1}{N\,|H_m|}\sum_{i=1}^{N} g(y_i)k_m\big(H_m^{-1}(\mathbf{w}_i - \mathbf{w}_a)\big)\big/\hat{f}(\mathbf{w}_a). \tag{8.1.31}$$

Robinson's (1988b) approach does not allow the identification of the parameters of variables that appear both in the regression equation ($\mathbf{x}$) and in the selection equation ($\mathbf{w}$). When there are variables appearing in both $\mathbf{x}$ and $\mathbf{w}$, Newey (1999) suggests a two-step series method of estimating $\boldsymbol{\beta}$ provided that the selection correction term of (8.1.25), $\lambda(\mathbf{w}_i, d_i = 1)$, is a function of the single index $\mathbf{w}_i'\mathbf{a}$:

$$\lambda(\mathbf{w}, d = 1) = E[u \mid v(\mathbf{w}'\mathbf{a}), d = 1]. \tag{8.1.32}$$

The first step of Newey's method uses the distribution-free methods discussed in Chapter 7 and in 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 $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a}) = (P_{1K}(\mathbf{w}'\mathbf{a}), P_{2K}(\mathbf{w}'\mathbf{a}), \ldots, P_{KK}(\mathbf{w}'\mathbf{a}))'$ 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 (1999) shows that the two-step series estimation of $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed when $N \to \infty$, $K \to \infty$, and $\sqrt{N}K^{-s-t+1} \to 0$, where $s \geq 5$, and where $K^7/N \to 0$ if $P_N^K(\mathbf{w}'\mathbf{a})$ is a power series or $m \geq t - 1$, $s \geq 3$, and $K^4/N \to 0$ if $P^K(\mathbf{w}'\mathbf{a})$ is a spline of degree m in $(\mathbf{w}'\mathbf{a})$.[2]

If the selection factor $\lambda(\mathbf{w}_i)$ is a function of a *single index* $\mathbf{w}_i'\mathbf{a}$, and the components of $\mathbf{w}_i$ are not a subset of $\mathbf{x}_i$, then instead of subtracting (8.1.26) from (8.1.25) to eliminate the unknown selection factor $\lambda(\mathbf{w}_i)$, Ahn and Powell (1993) note that for those individuals with $\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}$, one has $\lambda(\mathbf{w}_i'\mathbf{a}) = \lambda(\mathbf{w}_j'\mathbf{a})$. Thus, conditional on $\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}$, $d_i = 1$, $d_j = 1$,

$$(y_i - y_j) = (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta} + (\epsilon_i - \epsilon_j), \tag{8.1.33}$$

where the error term $(\epsilon_j - \epsilon_j)$ is symmetrically distributed around zero. They show that if λ is a sufficiently "smooth" function and $\hat{\mathbf{a}}$ is a consistent estimator of $\mathbf{a}$, observations for which the difference $(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}}$ is close to zero should have $\lambda(\mathbf{x}_i'\hat{\mathbf{a}}) - \lambda(\mathbf{w}_j'\hat{\mathbf{a}}) \simeq 0$. Therefore, Powell (2001) proposes a two-step procedure. In the first step, consistent semiparametric estimates of the coefficients of the *selection* equation are obtained. The result is used to obtain estimates of the single index $(\mathbf{x}_i'\mathbf{a})$ variables characterizing the selectivity bias in the equation of interest. The second step of the approach estimates the parameters of the interest by a weighted least-squares (or instrumental) variables regression of pairwise differences in dependent variables in the sample on the corresponding differences in explanatory variables:

$$\hat{\boldsymbol{\beta}}_{\text{AP}} = \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right) \cdot (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)' d_i d_j\right]^{-1} \times \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right) \cdot (\mathbf{x}_i - \mathbf{x}_j)(y_i - y_j) d_i d_j\right], \tag{8.1.34}$$

where $K(\cdot)$ is a kernel density weighting function that is bounded, is symmetric, and tends to zero as the absolute value of its argument increases, and h_N is a positive constant (or bandwidth) that decreases to zero such that $N(h_N)^6 \to 0$, and $N(h_N)^8 \to 0$ as $N \to \infty$. Often, standard normal density is used as a kernel function. The effect of multiplying by $K(\cdot)$ is to give more weights to observations with $(1/h_N)(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}} \simeq 0$ and less weight to observations for which $\mathbf{w}_i'\hat{\mathbf{a}}$ is different from $\mathbf{w}_j'\hat{\mathbf{a}}$, so that in the limit only observations with $\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}$ are used in (8.1.34), and (8.1.34) converges to a weighted least-squares estimator for the truncated data,

$$\hat{\boldsymbol{\beta}}_{\text{AP}} \to \{E\{f(\mathbf{w}'\mathbf{a})[\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}'\mathbf{a})][\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}'\mathbf{a})]'\}\}^{-1} \times \{E\{f(\mathbf{w}'\mathbf{a})[\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}'\mathbf{a})][y - E(y \mid \mathbf{w}'\mathbf{a})]\}\}, \tag{8.1.35}$$

where $f(\mathbf{w}'\mathbf{a})$ denotes the density function of $\mathbf{w}'\mathbf{a}$, which is assumed to be continuous and bounded above.

Both the Robinson (1988b) semiparametric estimator and the Powell-type pairwise differencing estimator converge to the true value at the speed of $N^{-1/2}$. However, neither method can provide estimate of the intercept term, because differencing the observation conditional on $\mathbf{w}$ or $\mathbf{w}'\mathbf{a}$, although it eliminates the selection factor $\lambda(\mathbf{w})$, also eliminates the constant term, nor can $\mathbf{x}$ and $\mathbf{w}$ be identical. Chen (1999) notes that if (u, v) are jointly symmetrical and $\mathbf{w}$ includes a constant term,

$$\begin{aligned}
&E(u \mid v > -\mathbf{w}'a)\,\text{Prob}(v > -\mathbf{w}'a) - E(u \mid v > \mathbf{w}'\mathbf{a})\,\text{Prob}(v > \mathbf{w}'\mathbf{a}) \\
&\quad = \int_{-\infty}^{\infty}\int_{-\mathbf{w}'\mathbf{a}}^{\infty} uf(u,v)\,du\,dv - \int_{-\infty}^{\infty}\int_{\mathbf{w}'\mathbf{a}}^{\infty} uf(u,v)\,du\,dv \\
&\quad = \int_{-\infty}^{\infty}\int_{-\mathbf{w}'\mathbf{a}}^{\mathbf{w}'\mathbf{a}} uf(u,v)\,du\,dv = 0,
\end{aligned} \tag{8.1.36}$$

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

$$\begin{aligned}
&E[d_i y_i - d_j y_j - (d_i \mathbf{x}_i - d_j \mathbf{x}_j)'\boldsymbol{\beta} \mid \mathbf{w}_i'\mathbf{a} = -\mathbf{w}_j'\mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] \\
&\quad = E[d_i u_i - d_j u_j \mid \mathbf{w}_i'\mathbf{a} = -\mathbf{w}_j'\mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] = 0.
\end{aligned} \tag{8.1.37}$$

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 (8.1.37) 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

$$\hat{\boldsymbol{\beta}}_c = \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i + \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right)(d_i\mathbf{x}_i - d_j\mathbf{x}_j)(\mathbf{z}_i - \mathbf{z}_j)'\right]^{-1} \times \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i + \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right)(\mathbf{z}_i - \mathbf{z}_j)'(d_i y_i - d_j y_j)\right], \tag{8.1.38}$$

where $\mathbf{z}_i$ are the instruments for $d_i\mathbf{x}_i$. In the case when y are unobservable, but the corresponding $\mathbf{x}$ 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 $\mathbf{a}$.

8.2 AN EXAMPLE – NONRANDOMLY MISSING DATA

8.2.1 Introduction

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 percent, did not complete the experiment. In Section 9.2 we shall discuss procedures to handle randomly missing data. 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 such as the New Jersey or Gary negative-income-tax experiments, some individuals may decide that keeping the detailed records that the experiments require is not worth the payment. Also, 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 (1976a); and Maddala (1983)] can provide both the theory and the computational techniques for analyzing nonrandomly missing data (Griliches, Hall, and Hausman (1978); Hausman and Wise (1979)).[3]

8.2.2 A Probability Model of Attrition and Selection Bias

Suppose that the structural model is

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{8.2.1}$$

where the error term v_{it} is assumed to follow a conventional error-components formulation $v_{it} = \alpha_i + u_{it}$. 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 (8.2.1) 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

$$d_i^* = \gamma y_{i2} + \boldsymbol{\theta}'\mathbf{x}_{i2} + \boldsymbol{\delta}'\mathbf{w}_i + \epsilon_i^* \geq 0, \tag{8.2.2}$$

where $\mathbf{w}_i$ is a vector of variables that do not enter the conditional expectation of y but affect the probability of observing y; $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are vectors of parameters; and (v_i, ϵ_i^*) are jointly normally distributed. Substituting for y_{i2} leads to the reduced-form specification

$$\begin{aligned} d_i^* &= (\gamma\boldsymbol{\beta}' + \boldsymbol{\theta}')\mathbf{x}_{i2} + \boldsymbol{\delta}'\mathbf{w}_i + \gamma v_{i2} + \epsilon_i^* \\ &= \boldsymbol{\pi}'\mathbf{x}_{i2} + \boldsymbol{\delta}'\mathbf{w}_i + \epsilon_i \\ &= \mathbf{a}'R_i + \epsilon_i, \end{aligned} \tag{8.2.3}$$

where $\epsilon_i = \gamma v_{i2} + \epsilon_i^*$, $R_i = (\mathbf{x}_{i2}', \mathbf{w}_i')'$, and $\mathbf{a}' = (\boldsymbol{\pi}', \boldsymbol{\delta}')$. We further assume that v_{it} are also normally distributed, and we normalize the variance σ_ϵ^2 of ϵ_i to 1. Then the probabilities of retention and attrition are probit functions given, respectively, by

$$\begin{aligned} \text{Prob}(d_i = 1) &= \Phi(\mathbf{a}'R_i), \\ \text{Prob}(d_i = 0) &= 1 - \Phi(\mathbf{a}'R_i), \end{aligned} \tag{8.2.4}$$

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

Suppose we estimate the model (8.2.1) using only complete observations. The conditional expectation of y_{i2}, given that it is observed, is

$$E(y_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'\mathbf{x}_{i2} + E(v_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1). \tag{8.2.5}$$

From $v_{i2} = \sigma_{2\epsilon}\epsilon_i + \eta_i$, where $\sigma_{2\epsilon}$ is the covariance between v_{i2} and ϵ_i, and η_i

is independent of ϵ_i (Anderson (1958, Chapter 2)), we have

$$\begin{aligned} E(v_{i2} \mid \mathbf{w}_i, d_i = 1) &= \sigma_{2\epsilon} E(\epsilon_i \mid \mathbf{w}_i, d_i = 1) \\ &= \frac{\sigma_{2\epsilon}}{\Phi(\mathbf{a}'R_i)} \int_{-\mathbf{a}'R_i}^{\infty} \epsilon \cdot \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2}\, d\epsilon \\ &= \sigma_{2\epsilon} \frac{\phi(\mathbf{a}'R_i)}{\Phi(\mathbf{a}'R_i)}, \end{aligned} \tag{8.2.6}$$

where $\phi(\cdot)$ denotes the standard normal density function. The last equality of (8.2.6) follows from the fact that the derivative of the standard normal density function $\phi(\epsilon)$ with respect to ϵ is $-\epsilon\phi(\epsilon)$. Therefore,

$$E(y_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'\mathbf{x}_{i2} + \sigma_{2\epsilon} \frac{\phi(\mathbf{a}'R_i)}{\Phi(\mathbf{a}'R_i)}. \tag{8.2.7}$$

Thus, estimating (8.2.1) using complete observations will lead to biased and inconsistent estimates of $\boldsymbol{\beta}$ unless $\sigma_{2\epsilon} = 0$. To correct for selection bias, one can use either Heckman's two-stage method (1979) (see Section 8.1) or the maximum likelihood method.

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

$$\begin{aligned} f(d_i = 1, y_{i1}, y_{i2}) &= \text{Prob}(d_i = 1 \mid y_{i1}, y_{i2}) f(y_{i1}, y_{i2}) \\ &= \text{Prob}(d_i = 1 \mid y_{i2}) f(y_{i1}, y_{i2}) \\ &= \Phi\left\{ \frac{\mathbf{a}'R_i + \left(\frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2}\right)(y_{i2} - \boldsymbol{\beta}'\mathbf{x}_{i2})}{\left[1 - \frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2}\right]^{1/2}} \right\} \\ &\quad \times \left[2\pi\sigma_u^2(\sigma_u^2 + 2\sigma_\alpha^2)\right]^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma_u^2}\left[\sum_{t=1}^{2}(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \right.\right. \\ &\qquad \left.\left. \times \left(\sum_{t=1}^{2}(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it}) \right)^2 \right] \right\}, \end{aligned} \tag{8.2.8}$$

where the first factor follows from the fact that the conditional density of $f(\epsilon_i \mid v_{i2})$ is normal, with mean $[\sigma_{2\epsilon}/(\sigma_u^2 + \sigma_\alpha^2)]v_{i2}$ and variance $1 - \sigma_{2\epsilon}^2/(\sigma_u^2 + \sigma_\alpha^2)$. When $d_i = 0$, y_{i2} is not observed and must be integrated out. In this instance, the joint density of $d_i = 0$ and y_{i1} is given by

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

$$\times \left[2\pi\left(\sigma_u^2+\sigma_\alpha^2\right)\right]^{-1/2}$$
$$\times \exp\left\{-\frac{1}{2(\sigma_u^2+\sigma_\alpha^2)}(y_{i1}-\boldsymbol{\beta}'\mathbf{x}_{i1})^2\right\}. \qquad (8.2.9)$$

The second equality of (8.2.9) follows from the fact that $f(\epsilon_i \mid v_{i1})$ 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 (8.2.8) and (8.2.9). 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

$$\begin{aligned}
\log L = &-N\log 2\pi - \frac{N_1}{2}\log\sigma_u^2 - \frac{N_1}{2}\log\left(\sigma_u^2+2\sigma_\alpha^2\right)\\
&-\frac{N-N_1}{2}\log\left(\sigma_u^2+\sigma_\alpha^2\right)\\
&-\frac{1}{2\sigma^2}\sum_{i=1}^{N_1}\left\{\sum_{t=1}^{2}(y_{it}-\boldsymbol{\beta}'\mathbf{x}_{it})^2-\frac{\sigma_\alpha^2}{\sigma_u^2+2\sigma_\alpha^2}\left[\sum_{t=1}^{2}(y_{it}-\boldsymbol{\beta}'\mathbf{x}_{it})\right]^2\right\}\\
&+\sum_{i=1}^{N_1}\log\Phi\left\{\frac{\mathbf{a}'R_i+\frac{\sigma_{2\epsilon}}{\sigma_u^2+\sigma_\alpha^2}(y_{i2}-\boldsymbol{\beta}'\mathbf{x}_{i2})}{\left[1-\frac{\sigma_{2\epsilon}^2}{\sigma_u^2+\sigma_\alpha^2}\right]^{1/2}}\right\}\\
&-\frac{1}{2(\sigma_u^2+\sigma_\alpha^2)}\sum_{i=N_1+1}^{N}(y_{i1}-\boldsymbol{\beta}'\mathbf{x}_{i1})^2\\
&+\sum_{i=N_1+1}^{N}\log\left\{1-\Phi\left[\frac{\mathbf{a}'R_i+\frac{\sigma_{1\epsilon}}{\sigma_u^2+\sigma_\alpha^2}(y_{i1}-\boldsymbol{\beta}'\mathbf{x}_{i1})}{\left[1-\frac{\sigma_{1\epsilon}^2}{\sigma_u^2+\alpha_\alpha^2}\right]^{1/2}}\right]\right\}. \qquad (8.2.10)
\end{aligned}$$

The critical parameter for attrition bias is $\sigma_{2\epsilon}$. If $\sigma_{2\epsilon}=0$, so does $\sigma_{1\epsilon}$. The likelihood function (8.2.10) 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 (8.2.1) 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, as discussed in Section 8.2. 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 (8.2.8) and (8.2.9) 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 (8.2.8) 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 does is 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 **y**, and the marginal density of **y**. The joint density of incomplete respondents becomes the product of the conditional probability of the observation being out of the sample, given the before-dropping-out values of **y**, and the marginal density of the previous periods' **y**. The likelihood function is simply the product of these two joint densities; see Griliches, Hall, and Hausman (1978) for a three-equation model.

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]

8.2.3 Attrition in the Gary Income-Maintenance Experiment

The Gary income-maintenance project focused on the effect 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 preexperimental 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 preexperimental behavior and also compared with that of the control group to obtain estimates of the effects of treatment plans.

Two broad groups of families were studied in the Gary experiment: Black female-headed households and black male-headed households. There was little attrition among the first group, but the attrition among male-headed families was substantial. Of the sample of 334 experimentals used by Hausman and

Wise (1979), the attrition rate was 31.1 percent. Among the 251 controls, 40.6 percent failed to complete the experiment.

If attrition is random, as will be discussed in Section 9.2, 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 breakeven point, experimentals receive no benefits from the experimental treatment. The breakeven 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 with respect to their incentive to remain in the experiment. But because high earnings are caused in part by the unobserved random term of the structural equation (8.2.1), attrition may well be related to it.

Hausman and Wise (1979) estimated structural models of earnings with and without correcting for attrition. The logarithm of earnings was regressed against time trend, education, experience, union membership, health status, and the logarithm of nonlabor family income. To control for the effects of the treatment, they also used a dummy variable that was 1 if for that period the household was under one of the four basic income-maintenance plans, and 0 otherwise. Because hourly wages for experimentals and controls 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 preexperiment 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, (8.2.1) and (8.2.3), are presented in Table 8.1.

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 percent. Some of the other coefficients showed more pronounced changes. The effect of nonlabor family income on earnings (hence hours worked) decreased by 23 percent from the generalized-least-squares estimates, and the effect of another year of education increased by 43 percent. 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, individuals are *missing at random* (MAR) (Rubin (1976); Little and Rubin (1987)) 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 who 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 et al. (2001)).

8.3 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

$$y_{it}^* = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{8.3.1}$$

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

$$\prod_{i=1}^{N} \int \left[\prod_{t=1}^{T} [1 - F(-\boldsymbol{\beta}'\mathbf{x}_{it} - \alpha_i)]^{-1} f(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it} - \alpha_i) \right] g(\alpha_i)\, d\alpha_i, \tag{8.3.2}$$

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

$$\prod_{i=1}^{N} \int \left[\prod_{t \in c_i} F(-\boldsymbol{\beta}'\mathbf{x}_{it} - \alpha_i) \prod_{t \in \bar{c}_i} f(y_{it} - \alpha_i - \boldsymbol{\beta}'\mathbf{x}_{it}) \right] g(\alpha_i)\, d\alpha_i, \tag{8.3.3}$$

Table 8.1. *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)
	$\hat{\sigma}_u^2 = \underset{(0.0057)}{0.1832}$		$\hat{\sigma}_u^2 = 0.1236$
	$\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = \underset{(0.0391)}{0.2596}$	$\frac{\hat{\sigma}_{2\epsilon}}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = \underset{(0.0429)}{-0.1089}$	$\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = 0.2003$

[a]Not estimated.
Source: Hausman and Wise (1979, Table IV).

where $c_i = \{t \mid y_{it} = 0\}$ and $\bar{c}_i$ denotes its complement. Maximizing (8.3.2) or (8.3.3) with respect to unknown parameters yields consistent and asymptotically normally distributed estimators.

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

$$d_{it}^* = \mathbf{w}_{it}'\mathbf{a} + \eta_i + \nu_{it}, \tag{8.3.4}$$

with the observed (y_{it}, d_{it}) following the rule $d_{it} = 1$ if $d_{it}^* > 0$ and zero otherwise, as in (8.1.17), and $y_{it}^* = y_{it}$ if $d_{it} = 1$ and unknown otherwise, as in (8.1.18). Suppose that the joint density of (α_i, η_i) is given by $g(\alpha, \eta)$. Then the likelihood function of the type II Tobit model takes the form

$$\begin{aligned}
&\prod_{i=1}^{N} \int \Bigg[\prod_{t \in c_i} \text{Prob}(d_{it} = 0 \mid \mathbf{w}_{it}, \alpha_i) \prod_{t \in \bar{c}_i} \text{Prob}(d_{it} = 1 \mid \mathbf{w}_{it}, \alpha_i) \\
&\qquad\qquad \times f(y_{it} \mid \mathbf{x}_{it}, \mathbf{w}_{it}, \alpha_i, \eta_i, d_{it} = 1) \Bigg] g(\alpha_i, \eta_i)\, d\alpha_i\, d\eta_i \\
&= \prod_{i=1}^{N} \int \Bigg[\prod_{t \in c_i} \text{Prob}(d_{it} = 0 \mid \mathbf{w}_{it}, \alpha_i) \prod_{t \in \bar{c}_i} \text{Prob}(d_{it} = 1 \mid \mathbf{w}_{it}, \eta_i, \alpha_i, y_{it}, \mathbf{x}_{it}) \\
&\qquad\qquad \times f(y_{it} \mid \mathbf{x}_{it}, \alpha_i) \Bigg] g(\alpha_i, \eta_i)\, d\alpha_i\, d\eta_i.
\end{aligned} \tag{8.3.5}$$

Maximizing the likelihood function (8.3.2), (8.3.3), or (8.3.5) with respect to unknown parameters yields consistent and asymptotically normally distributed estimator of $\boldsymbol{\beta}$ when either N or T or both tend to infinity. However, the computation is quite tedious even with a simple parametric specification of the individuals effects α_i and η_i, because it involves multiple integration.[6] Neither is a generalization of the Heckman (1976a) two-stage estimator easily implementable (e.g., Nijman and Verbeek (1992); Ridder (1990); Vella and Verbeek (1999); Wooldridge (1999)). Moreover, both the MLE and the Heckman two-step estimators are sensitive to the exact specification of the error distribution. However, if the random effects α_i and η_i are independent of $\mathbf{x}_i$, then the Robinson (1988b) and Newey (1999) estimators ((8.1.27) and (8.1.32)) can be applied to obtain consistent and asymptotically normally distributed estimators of $\boldsymbol{\beta}$. Alteratively, one may ignore the randomness of α_i and η_i and apply the Honoré (1992) fixed-effects trimmed least-squares or least-absolute-deviation estimator for the panel data censored and truncated regression models, or the Kyriazidou (1997) two-step semiparametric estimator for the panel data sample selection model, to estimate $\boldsymbol{\beta}$ (see Section 8.4).

8.4 FIXED-EFFECTS ESTIMATOR

8.4.1 Pairwise Trimmed Least-Squares and Least-Absolute-Deviation Estimators for Truncated and Censored Regressions

When the effects are fixed and if $T \to \infty$, the MLEs of $\boldsymbol{\beta}'$ and α_i are straightforward to implement and are consistent. However, panel data often involve many individuals observed over few time periods, so that the MLE, in general, will be inconsistent as described in Chapter 7. In this section, we consider the pairwise trimmed least-squares (LS) and least-absolute-deviation (LAD) estimators of Honoré (1992) for panel data censored and truncated regression models that are consistent without the need to assume a parametric form for the disturbances u_{it}, nor homoscedasticity across individuals.

8.4.1.a Truncated Regression

We assume a model (8.3.1) and (8.1.2) except that now the individual effects are assumed fixed. The disturbance u_{it} is again assumed to be independently distributed over i and i.i.d. over t conditional on $\mathbf{x}_i$ and α_i.

We note that when 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

$$y_{it} = E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) + \epsilon_{it}, \tag{8.4.1}$$

where

$$E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}). \tag{8.4.2}$$

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

$$\begin{aligned} &E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) - E(y_{is} \mid \mathbf{x}_{is}, \alpha_i, y_{is} > 0) \\ &\quad = (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &\quad\quad - E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}). \end{aligned} \tag{8.4.3}$$

In other words,

$$\begin{aligned} (y_{it} - y_{is}) &= (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &\quad - E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}) + (\epsilon_{it} - \epsilon_{is}). \end{aligned} \tag{8.4.4}$$

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

$$\begin{aligned}&E(y_{is}\mid\alpha_i,\mathbf{x}_{it},\mathbf{x}_{is},y_{is}>-(\mathbf{x}_{it}-\mathbf{x}_{is})'\boldsymbol{\beta})\\&\quad=\alpha_i+\mathbf{x}_{is}'\boldsymbol{\beta}+E(u_{is}\mid u_{is}>-\alpha_i-\mathbf{x}_{is}'\boldsymbol{\beta}-(\mathbf{x}_{it}-\mathbf{x}_{is})'\boldsymbol{\beta}).\end{aligned}\tag{8.4.5}$$

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

$$E(u_{it}\mid u_{it}>-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta})=E(u_{is}\mid u_{is}>-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta}).\tag{8.4.6}$$

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

$$\begin{aligned}&E(u_{it}\mid u_{it}>-\alpha_i-\mathbf{x}_{it}'\boldsymbol{\beta}+(\mathbf{x}_{it}-\mathbf{x}_{is})'\boldsymbol{\beta})\\&\quad=E(u_{it}\mid u_{it}>-\alpha_i-\mathbf{x}_{is}'\boldsymbol{\beta})\\&\quad=E(u_{is}\mid u_{is}>-\alpha_i-\mathbf{x}_{is}'\boldsymbol{\beta}).\end{aligned}\tag{8.4.7}$$

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

$$(y_{it}-y_{is})=(\mathbf{x}_{it}-\mathbf{x}_{is})'\boldsymbol{\beta}+(\epsilon_{it}-\epsilon_{is}),\tag{8.4.8}$$

which no longer involves the incidental parameters α_i. Since $E[(\epsilon_{it}-\epsilon_{is})\mid\mathbf{x}_{it},\mathbf{x}_{is}]=0$, applying least squares to (8.4.8) will yield a consistent estimator of $\boldsymbol{\beta}$.

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 8.3. Since u_{i1} and u_{i2} are i.i.d. conditional on $(\mathbf{x}_{i1},\mathbf{x}_{i2},\alpha_i)$, the probability density of y_{i1}^* and y_{i2}^* conditional on $(\mathbf{x}_{i1},\mathbf{x}_{i2},\alpha_i)$ should have the same shape except for the location. The top and bottom graphs of Figure 8.4 postulate the probability density of y_{i1}^* and y_{i2}^* conditional on $(\mathbf{x}_{i1},\mathbf{x}_{i2},\alpha_i)$, respectively, assuming that $\Delta\mathbf{x}_i'\boldsymbol{\beta}<0$, where $\Delta\mathbf{x}_i=\Delta\mathbf{x}_{i2}=\mathbf{x}_{i2}-\mathbf{x}_{i1}$. The truncated data correspond to those sample points where y_{it}^* or $y_{it}>0$. Because $\mathbf{x}_{i1}'\boldsymbol{\beta}\neq\mathbf{x}_{i2}'\boldsymbol{\beta}$, the probability density of y_{i1} is different from that of y_{i2}. However, the probability density of y_{i1}^* given $y_{i1}^*>-\Delta\mathbf{x}_i'\boldsymbol{\beta}$ (or y_{i1} given $y_{i1}>-\Delta\mathbf{x}_i'\boldsymbol{\beta}$) is identical to the probability density of y_{i2}^* given $y_{i2}^*>0$ (or y_{i2} given $y_{i2}>0$) as shown in Figure 8.4. Similarly, if $\Delta\mathbf{x}_i'\boldsymbol{\beta}>0$, 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 $y_{i2}^*>\Delta\mathbf{x}_i'\boldsymbol{\beta}$ as shown in Figure 8.5.[7] In other words, in a two-dimensional diagram of (y_{i1}^*,y_{i2}^*) as in Figure 8.6 or 8.7, (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 $(\mathbf{x}_{i1}'\boldsymbol{\beta},\mathbf{x}_{i2}'\boldsymbol{\beta})$ or $(-\Delta\mathbf{x}_i'\boldsymbol{\beta},0)$, e.g., 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 $(\mathbf{x}_{i1},\mathbf{x}_{i2})$. 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 $B_1=\{(y_{i1}^*,y_{i2}^*):y_{i1}^*>-\Delta\mathbf{x}_i'\boldsymbol{\beta},\ 0<y_{i2}^*<y_{i1}^*+\Delta\mathbf{x}_i'\boldsymbol{\beta}\}$

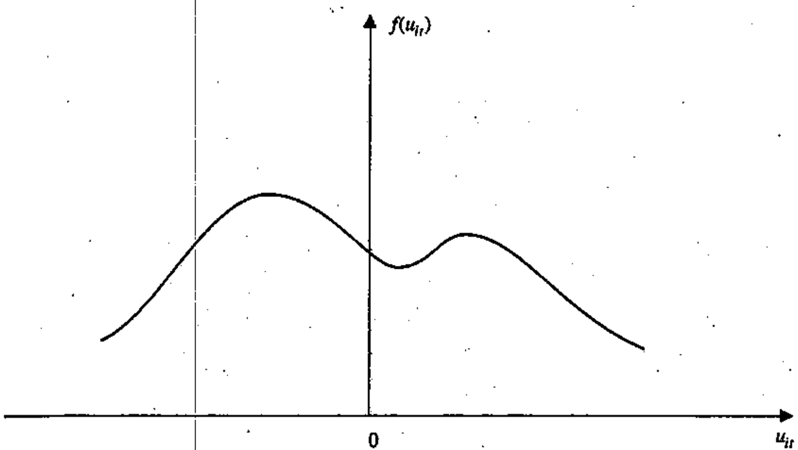

Fig. 8.3 Probability density of u_{it}.

(Figure 8.6). 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 $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0,\ \Delta\mathbf{x}_i'\boldsymbol{\beta} < y_{i2}^* < y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ (Figure 8.7). 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 $(0 < y_{i1}^* < -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2}^* > 0)$ in Figure 8.6 (corresponding to points $(y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2})$) and $(y_{i1}^* > 0,\ 0 < y_{i2}^* < \Delta\mathbf{x}_i'\boldsymbol{\beta})$ in Figure 8.7 (corresponding to points $(y_{i1}, y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta})$) will have to be thrown away to restore symmetry.

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

$$\begin{aligned} E[\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta} \mid y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}] &= E[\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta} \mid y_{i1}^* > 0,\ y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2}^* > 0 \\ &\quad y_{i2}^* > \Delta\mathbf{x}_i'\boldsymbol{\beta}] = 0, \end{aligned} \tag{8.4.9a}$$

and

$$E[(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})\Delta\mathbf{x}_i \mid y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}] = \mathbf{0}, \tag{8.4.9b}$$

where $\Delta y_i = \Delta y_{i2} = y_{i2} - y_{i1}$. Therefore, Honoré (1992) suggests the trimmed LAD and LS estimators $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ that minimize the objective functions

$$\begin{aligned} Q_N(\boldsymbol{\beta}) &= \sum_{i=1}^{N} [|\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta}|\, 1\{y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}\} \\ &\qquad + |y_{i1}|\, 1\{y_{i1} \geq -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta}\} \\ &\qquad + |y_{i2}|\, 1\{y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} \geq \Delta\mathbf{x}_i'\boldsymbol{\beta}\}] \\ &= \sum_{i=1}^{N} \psi(y_{i1}, y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta}), \end{aligned} \tag{8.4.10}$$

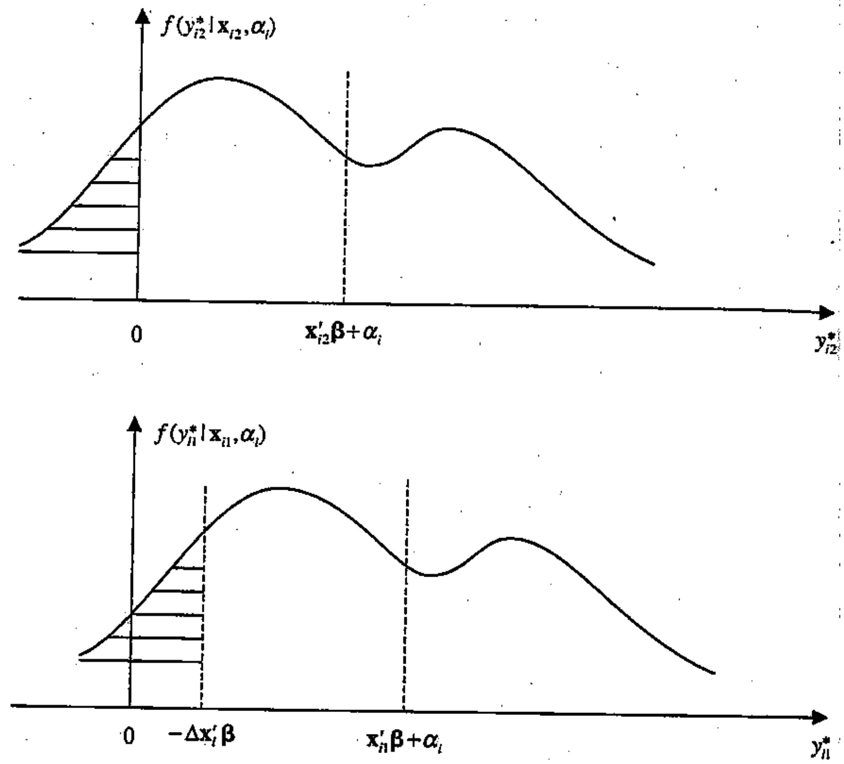

Fig. 8.4 Conditional densities of y_{i1}^* and y_{i2}^* given $(x_{i1}, x_{i2}, \alpha_i)$, assuming $\Delta x_i' \boldsymbol{\beta} < 0$.

and

$$\begin{aligned} R_N(\mathbf{b}) &= \sum_{i=1}^{N} [(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})^2 1\{y_{i1} \geq -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}\} \\ &\quad + y_{i1}^2 1\{y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta}\} \\ &\quad + y_{i2}^2 1\{y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}\}] \\ &= \sum_{i=1}^{N} \psi(y_{i1}, y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta})^2, \end{aligned} \tag{8.4.11}$$

respectively. The function $\psi(w_1, w_2, c)$ is defined for $w_1 > 0$ and $w_2 > 0$ by

$$\psi(w_1, w_2, c) = \begin{cases} w_1 & \text{for } w_2 < c, \\ w_2 - w_1 - c & \text{for } -w_1 < c < w_2, \\ w_2 & \text{for } w_1 < -c. \end{cases}$$

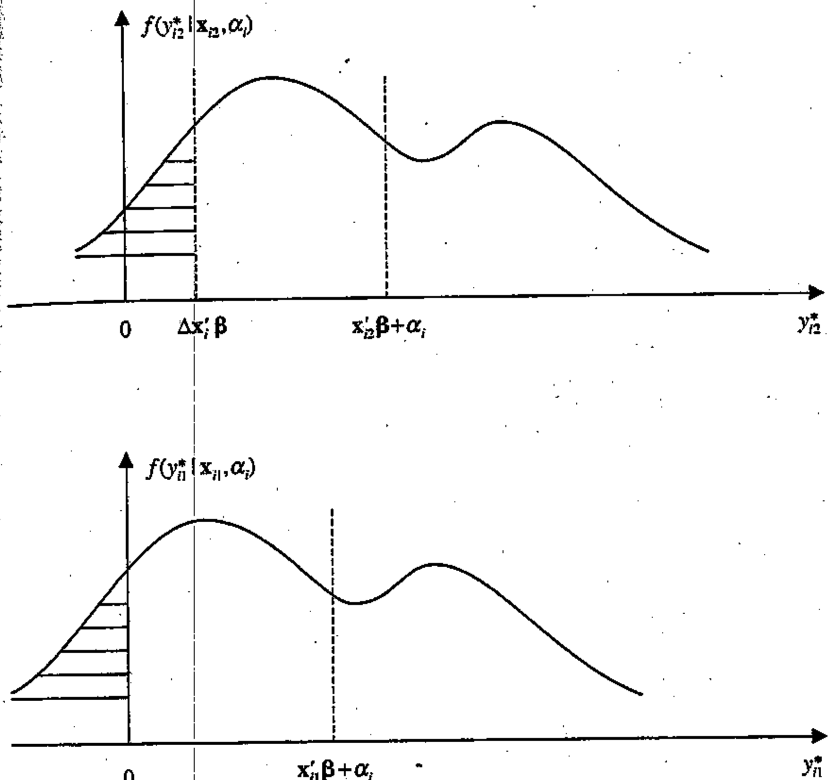

Fig. 8.5 Conditional densities of y_{i1}^* and y_{i2}^* given $(x_{i1}, x_{i2}, \alpha_i)$, assuming $\Delta x_i' \boldsymbol{\beta} > 0$.

The first-order conditions for (8.4.10) and (8.4.11) are the sample analogues of

$$E\{[P(y_{i1} > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ y_{i2} > y_{i1} + \Delta \mathbf{x}_i' \boldsymbol{\beta}) \\ - P(y_{i1} > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ \Delta \mathbf{x}_i' \boldsymbol{\beta} < y_{i2} < y_{i1} + \Delta \mathbf{x}_i' \boldsymbol{\beta})]\Delta \mathbf{x}_i'\} = \mathbf{0}', \tag{8.4.12}$$

and

$$E\{(\Delta y_i - \Delta \mathbf{x}_i' \boldsymbol{\beta})\, \Delta \mathbf{x}_i \mid (y_{i1} > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ y_{i2} > y_{i1} + \Delta \mathbf{x}_i' \boldsymbol{\beta}) \\ \cup\ (y_{i1} > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ \Delta \mathbf{x}_i' \boldsymbol{\beta} < y_{i2} < y_{i1} + \Delta \mathbf{x}_i' \boldsymbol{\beta})\} = \mathbf{0}, \tag{8.4.13}$$

respectively. Honoré (1992) proves that $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent and asymptotically normally distributed if the density of u is strictly log-concave. The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ may be

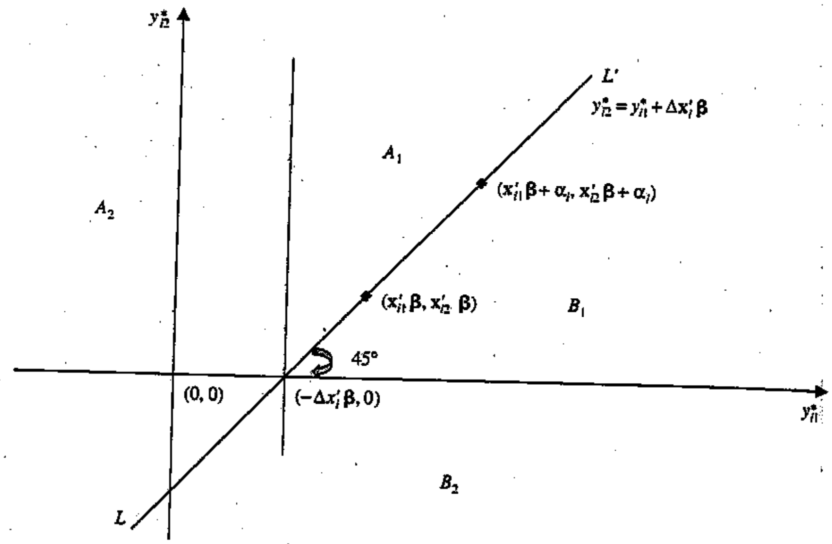

Fig. 8.6 The distribution of (y_{i1}^*, y_{i2}^*) assuming $\Delta \mathbf{x}_i' \beta < 0$.
$A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ y_{i2}^* > y_{i1}^* + \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$, $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ y_{i2}^* > 0\}$,
$B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ 0 < y_{i2}^* < y_{i1}^* + \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$, $B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta \mathbf{x}_i' \boldsymbol{\beta},\ y_{i2}^* \leq 0\}$.

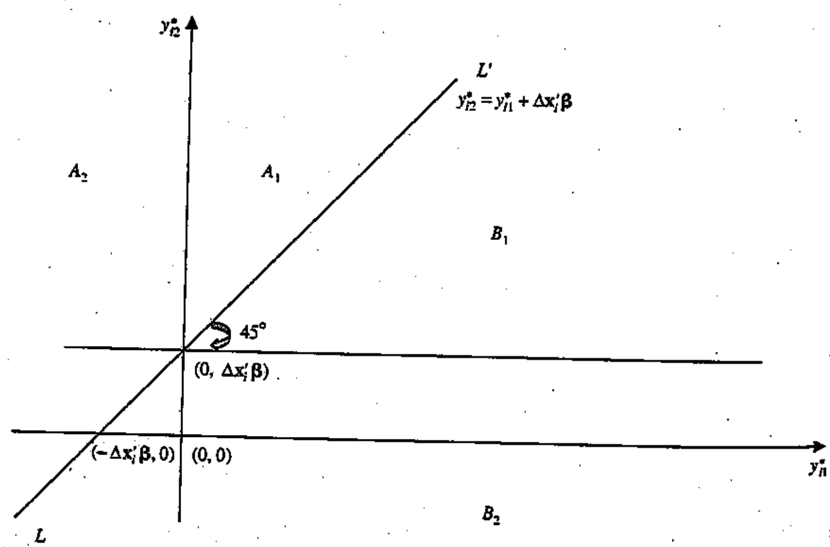

Fig. 8.7 The distribution of (y_{i1}^*, y_{i2}^*) assuming $\Delta \mathbf{x}_i' \boldsymbol{\beta} > 0$.
$A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0,\ y_{i2}^* > y_{i1}^* + \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$, $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq 0,\ y_{i2}^* > \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$,
$B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0,\ \Delta \mathbf{x}_i' \boldsymbol{\beta} < y_{i2}^* < y_{i1}^* + \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$, $B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0,\ y_{i2}^* \leq \Delta \mathbf{x}_i' \boldsymbol{\beta}\}$.

approximated by

$$\text{Asy Cov}(\sqrt{N}(\hat{\boldsymbol\beta} - \boldsymbol\beta)) = \Gamma_1^{-1} V_1 \Gamma_1^{-1}, \tag{8.4.14}$$

and

$$\text{Asy Cov}(\sqrt{N}(\tilde{\boldsymbol\beta} - \boldsymbol\beta)) = \Gamma_2^{-1} V_2 \Gamma_2^{-1}, \tag{8.4.15}$$

where V_1, V_2, Γ_1, and Γ_2 may be approximated by

$$\hat{V}_1 = \frac{1}{N}\sum_{i=1}^{N} 1\{-y_{i1} < \Delta\mathbf{x}_i'\hat{\boldsymbol\beta} < y_{i2}\}\Delta\mathbf{x}_i\,\Delta\mathbf{x}_i', \tag{8.4.16}$$

$$\hat{V}_2 = \frac{1}{N}\sum_{i=1}^{N} 1\{-y_{i1} < \Delta\mathbf{x}_i'\tilde{\boldsymbol\beta} < y_{i2}\}(\Delta y_i - \Delta\mathbf{x}_i'\tilde{\boldsymbol\beta})^2\Delta\mathbf{x}_i\,\Delta\mathbf{x}_i', \tag{8.4.17}$$

$$\begin{aligned}\hat{\Gamma}_1^{(j,k)} = \frac{1}{h_N}\Bigg[&\frac{1}{N}\sum_{i=1}^{N}(1\{\Delta y_i < \Delta\mathbf{x}_i(\hat{\boldsymbol\beta} + h_N\mathbf{i}_k) < y_{i2}\} \\ &\quad - 1\{-y_{i1} < \Delta\mathbf{x}_i(\hat{\boldsymbol\beta} + h_N\mathbf{i}_k) < \Delta y_i\})\,\Delta\mathbf{x}_i^{(j)} \\ &+ \frac{1}{N}\sum_{i=1}^{N}(-1\{\Delta y_i < \Delta\mathbf{x}_i'\hat{\boldsymbol\beta} < y_{i2}\} \\ &\quad + 1\{-y_{i1} < \Delta\mathbf{x}_i'\hat{\boldsymbol\beta} < \Delta y_i\})\,\Delta\mathbf{x}_i^{(j)}\Bigg],\end{aligned} \tag{8.4.18}$$

$$\begin{aligned}\hat{\Gamma}_2^{(j,k)} = \frac{1}{h_N}\Bigg[&\frac{1}{N}\sum_{i=1}^{N} 1\{-y_{i1} < \Delta\mathbf{x}_i'(\tilde{\boldsymbol\beta} + h_N\mathbf{i}_k) < y_{i2}\} \\ &\quad \times (\Delta y_i - \Delta\mathbf{x}_i'(\tilde{\boldsymbol\beta} + h_N\mathbf{i}_k))\,\Delta\mathbf{x}_i^{(j)} \\ &- \frac{1}{N}\sum_{i=1}^{N} 1\{-y_{i1} < \Delta\mathbf{x}_i'\tilde{\boldsymbol\beta} < y_{i2}\}(\Delta y_i - \Delta\mathbf{x}_i'\tilde{\boldsymbol\beta})\,\Delta\mathbf{x}_i^{(j)}\Bigg],\end{aligned} \tag{8.4.19}$$

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 jth coordinate of $\Delta\mathbf{x}_i$, $\mathbf{i}_k$ is a unit vector with 1 in its kth place, and h_N decreases to zero with the speed of $N^{-\frac{1}{2}}$. The bandwidth factor h_N appears in (8.4.18) and (8.4.19) because Γ_ℓ is a function of densities and conditional expectations of y (Honoré (1992)).

8.4.1.b Censored Regressions

When data are censored, observations $\{y_{it}, \mathbf{x}_{it}\}$ are available for $i = 1, \ldots, N$, $t = 1, \ldots, T$, 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 8.4.1.a. But this will lead to a loss of information.

In the case that data are censored, in addition to the relations (8.4.9a,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 8.6, regions 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 8.7. 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 $(-\Delta\mathbf{x}_i'\boldsymbol{\beta}, 0)$ – the line LL' in Figure 8.6 or 8.7. 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 8.6. 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 $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq 0,\ y_i^* > \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ 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 seen in Figure 8.7. 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

$$E\left[(1\{(y_{i1}, y_{i2}) \in A\} - 1\{(y_{i1}, y_{i2}) \in B\})\,\Delta\mathbf{x}_i\right] = \mathbf{0}. \tag{8.4.20}$$

In other words, to restore symmetry of censored observations around their expected values, observations corresponding 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

$$\begin{aligned}
&E[(1\{(y_{i1}, y_{i2}) \in A_1\}(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta}) + 1\{(y_{i1}, y_{i2}) \in A_2)\} \\
&\quad\times (y_{i2} - \max(0, \Delta\mathbf{x}_i'\boldsymbol{\beta})) - 1\{(y_{i1}, y_{i2}) \in B_1\}(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta}) \\
&\quad - 1\{y_{i1}, y_{i2} \in B_2\}(y_{i1} - \max(0, -\Delta\mathbf{x}_i'\boldsymbol{\beta})))\Delta\mathbf{x}_i] = \mathbf{0}. \qquad (8.4.21)
\end{aligned}$$

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

$$\begin{aligned}
Q_N^*(\boldsymbol{\beta}) &= \sum_{i=1}^N [1 - 1\{y_{i1} \le -\Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i2} \le 0\}] \\
&\qquad \times [1 - 1\{y_{i2} \le \Delta\mathbf{x}_i'\boldsymbol{\beta},\ y_{i1} \le 0\}]|\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta}| \\
&= \sum_{i=1}^N \psi^*(y_{i1}, y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta}), \qquad (8.4.22) \\
R_N^*(\boldsymbol{\beta}) &= \sum_{i=1}^N \{[\max\{y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta}\} - \max\{y_{i1}, -\Delta\mathbf{x}_i'\boldsymbol{\beta}\} - \Delta\mathbf{x}_i'\boldsymbol{\beta})]^2 \\
&\qquad - 2 \times 1\{y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta}\}(y_{i1} + \Delta\mathbf{x}_i'\boldsymbol{\beta})y_{i2} \\
&\qquad - 2 \times 1\{y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta}\}(y_{i2} - \Delta\mathbf{x}_i'\boldsymbol{\beta})y_{i1}\} \\
&= \sum_{i=1}^N \chi(y_{i1}, y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta}), \qquad (8.4.23)
\end{aligned}$$

where

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

and

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

The first-order conditions for (8.4.22) and (8.4.23) are the sample analogues of (8.4.20) and (8.4.21), 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 $E\{[1\{(y_{i1}, y_{i2}) \in A_1\}(\Delta y_i - \Delta\mathbf{x}_i'\beta) - 1\{(y_{i1}, y_{i2}) \in B_1\}(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})]\,\Delta\mathbf{x}_i\}$. The partial derivatives 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 $\hat{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$ are consistent and asymptotically normally distributed. The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})$ is equal to

$$\text{Asy Cov}(\sqrt{N}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})) = \Gamma_3^{-1} V_3 \Gamma_3^{-1}, \tag{8.4.24}$$

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

$$\text{Asy Cov}(\sqrt{N}(\tilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta})) = \Gamma_4^{-1} V_4 \Gamma_4^{-1}, \tag{8.4.25}$$

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

$$\begin{aligned} \hat{V}_3 = \frac{1}{N}\sum_{i=1}^{N} 1\{[\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^* < \Delta y_i,\ y_{i2} > \max(0, \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)] \\ \cup [\Delta y_i < \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*,\ y_{i1} > \max(0, -\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)]\}\, \Delta\mathbf{x}_i\, \Delta\mathbf{x}_i', \end{aligned} \tag{8.4.26}$$

$$\begin{aligned} \hat{V}_4 = \frac{1}{N}\sum_{i=1}^{N} \big[y_{i2}^2 1\{\Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* \leq -y_{i1}\} + y_{i1}^2 1\{y_{i2} \leq \Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^*\} \big] \\ + (\Delta y_i - \Delta\mathbf{x}_1'\tilde{\boldsymbol{\beta}}^*)^2 1\{-y_{i1} < \Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* < y_{i2}\}]\, \Delta\mathbf{x}_i\, \Delta\mathbf{x}_i', \end{aligned} \tag{8.4.27}$$

$$\begin{aligned} \hat{\Gamma}_3^{(j,k)} = \frac{-1}{h_N}\Bigg\{ \frac{1}{N}\sum_{i=1}^{N} [1\{y_{i2} > 0,\ y_{i2} > y_{i1} + \Delta\mathbf{x}_i'(\hat{\boldsymbol{\beta}}^* + h_N \mathbf{i}_k)\} \\ - 1\{y_{i1} > 0,\ y_{i1} > y_{i2} - \Delta\mathbf{x}_i'(\hat{\boldsymbol{\beta}}^* + \omega_n \mathbf{i}_k)\}]\, \Delta\mathbf{x}_i^{(j)} \\ - \frac{1}{N}\sum_{i=1}^{N} [1\{y_{i2} > 0,\ y_{i2} > y_{i1} + \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*\} \\ - 1\{y_{i1} > 0,\ y_{i1} > y_{i2} - \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*\}]\, \Delta\mathbf{x}_i^{(j)} \Bigg\}, \end{aligned} \tag{8.4.28}$$

and

$$\hat{\Gamma}_4 = \frac{1}{N}\sum_{i=1}^{N} 1\{-y_{i1} < \Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* < y_{i2}\}\, \Delta\mathbf{x}_i \Delta\mathbf{x}_i'. \tag{8.4.29}$$

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, (8.4.23) can be modified to be the estimator

$$\tilde{\boldsymbol{\beta}}^* = \arg\min \sum_{i=1}^{N}\sum_{t=2}^{T} \chi(y_{i,t-1}, y_{it}, (\mathbf{x}_{it} - \mathbf{x}_{it-1})'\boldsymbol{\beta}), \tag{8.4.30}$$

when $T > 2$.

8.4.2 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 (8.3.4) and (8.1.17) for the linear regression model (8.3.1), where $y_{it}^* = y_{it}$ if $d_{it} = 1$ and y_{it}^* is unknown if $d_{it} = 0$, as in (8.1.18). We assume that both (8.3.1) and (8.3.4) 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 (1976a), two-step estimation procedure for the parametric model, Kyriazidou (1997) proposes a two-step semiparametric method for estimating the main regression of interest, (8.3.1). In the first step, the unknown coefficients of the selection equation (8.3.4), $\mathbf{a}$, are consistently estimated by some semiparametric method. In the second step, these estimates are substituted into the equation of interest, (8.3.1), 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 effect 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 ν_{it} are uncorrelated conditional on α_i and $\mathbf{x}_{it}$, then (8.3.1) and (8.1.18) correspond to the standard variable intercept model for panel data discussed in Chapter 3 with balanced panel or randomly missing data.[8] If u_{it} and ν_{it} are correlated, sample selection will arise because $E(u_{it} \mid \mathbf{x}_{it}, \mathbf{w}_{it}, \alpha_i, d_{it} = 1) \neq 0$. Let $\lambda(\cdot)$ denote the conditional expectation of u conditional on $d = 1, \mathbf{x}, \mathbf{w}, \alpha$, and η; then (8.3.1) and (8.1.19) conditional on $d_{it} = 1$ can be written as

$$y_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \lambda(\eta_i + \mathbf{w}_{it}'\mathbf{a}) + \epsilon_{it}, \tag{8.4.31}$$

where $E(\epsilon_{it} \mid \mathbf{x}_{it}, \mathbf{w}_{it}, d_{it} = 1) = 0$.

The form of the selection function $\lambda(\cdot)$ is derived from the joint distribution of u and ν. For instance, if u and ν are bivariate normal, then we have the

Heckman sample selection correction

$$\lambda(\eta_i + \mathbf{a}'\mathbf{w}_{it}) = \frac{\sigma_{uv}}{\sigma_v}\frac{\phi\left(\frac{\eta_i + \mathbf{w}_{it}'\mathbf{a}}{\sigma_v}\right)}{\Phi\left(\frac{\eta_i + \mathbf{w}_{it}'\mathbf{a}}{\sigma_v}\right)}.$$

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

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

The presence of individual specific effects in (8.4.23) is easily obviated 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. But conditional on given i, if (u_{it}, v_{it}) are stationary and $\mathbf{w}_{it}'\mathbf{a} = \mathbf{w}_{is}'\mathbf{a}$, then $\lambda(\eta_i + \mathbf{w}_{it}\mathbf{a}) = \lambda(\eta_i + \mathbf{w}_{is}'\mathbf{a})$. Then the difference in (8.4.31) between t and s if both y_{it} and y_{is} are observable no longer contains the individual-specific effects α_i or the selection factor $\lambda(\eta_i + \mathbf{w}_{it}'\mathbf{a})$:

$$\Delta y_{its} = y_{it} - y_{is} = (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} + (\epsilon_{it} - \epsilon_{is}) = \Delta\mathbf{x}_{its}'\boldsymbol{\beta} + \Delta\epsilon_{its}. \tag{8.4.32}$$

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 propose to estimate the fixed-effects sample selection models in two steps: In the first step, estimate $\mathbf{a}$ by either Andersen (1970) and Chamberlain's (1980) conditional maximum likelihood approach or Horowitz (1992) and Lee's (1999) smoothed version of the Manski (1975) maximum score method discussed in Chapter 7. In the second step, the estimated $\hat{\mathbf{a}}$ is used to estimate $\boldsymbol{\beta}$ 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

$$\hat{\boldsymbol{\beta}}_K = \left\{\sum_{i=1}^N \frac{1}{T_i - 1}\sum_{1\le s<t\le T_i}(\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{x}_{it} - \mathbf{x}_{is})'K\left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N}\right]d_{it}\,d_{is}\right\}^{-1}$$
$$\times\left\{\sum_{i=1}^N \frac{1}{T_i - 1}\sum_{1\le s<t<T_i}(\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{y}_{it} - \mathbf{y}_{is})'K\left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N}\right]d_{it}\,d_{is}\right\}, \tag{8.4.33}$$

where T_i denotes the number of positively observed y_{it} for the ith individual, K 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 $N \to \infty$. The effect of multiplying the kernel function $K(\cdot)$ is to give more weight to observations with $(1/h_N)(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}} \simeq 0$ and less weight to those with $\mathbf{w}_{it}\hat{\mathbf{a}}$ 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 (8.4.33). Under appropriate regularity conditions (8.4.33) 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 (8.4.33) may be approximate by the Eicker (1963) and White's (1980) formulae for the asymptotic covariance matrix of the least-squares estimator of the linear regression model with heteroscedasticity,

$$\left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}'_i\right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i \, \Delta\hat{e}_i^2 \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i\right)^{-1}, \tag{8.4.34}$$

where $\hat{\mathbf{x}}_i = K(\Delta\mathbf{w}'_i\,\hat{\mathbf{a}}/h_N)^{1/2}\,\Delta\mathbf{x}_i(d_{i2}\,d_{i1})$ and $\Delta\hat{e}_i$ is the estimated residual of (8.4.32).

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 that only observations with $\Delta\mathbf{w}'_{its}\,\mathbf{a} = 0$ are used is to replace $K[\Delta\mathbf{w}_{its}\,\hat{\mathbf{a}}/h_N]$ by a multivariate kernel function $K((\mathbf{w}_{it} - \mathbf{w}_{is})/h_N)$ in (8.4.33). However, the speed of convergence of (8.4.33) to the true $\boldsymbol{\beta}$ will be $\sqrt{Nh_N^k}$, where k denotes the dimension of $\mathbf{w}_{it}$. This is much slower than $\sqrt{Nh_N}$, since h_N converges to zero as $N \to \infty$.

8.5 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:

$$d_{it} = 1(\mathbf{w}'_{it}\mathbf{a} + \eta_i + \nu_{it} > 0), \tag{8.5.1}$$

$$y_{1it} = \boldsymbol{\beta}'_1\mathbf{x}_{it} + \alpha_{1i} + u_{1it} \qquad \text{if} \quad d_{it} = 1, \tag{8.5.2}$$

$$y_{2it} = \boldsymbol{\beta}'_2\mathbf{x}_{it} + \alpha_{2i} + u_{2it} \qquad \text{if} \quad d_{it} = 0, \tag{8.5.3}$$

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; $\eta_i, \alpha_{1i}, \alpha_{2i}$ are unobserved household-specific effects; and $\nu_{it}, u_{1it}, u_{2it}$ 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 percent 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), its square (L2INC), monthly total family expenditure (EXP), logarithm of monthly total family expenditure (LEXP), 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 (8.5.1) are DOP, LEXP, L2EXP, LRP, and LOP. The variables excluded from the budget share equations ((8.5.2) and (8.5.3)) 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, Melenberg, and van Soest (2001) also estimate this model by the instrumental-variable (IV) method. For instance, the Kyriazidou (1997) weighted least-squares estimator is modified as

$$\hat{\boldsymbol{\beta}}_{KN} = \left\{ \sum_{i=1}^{N} \sum_{1\le s<t\le T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{z}_{it} - \mathbf{z}_{is})' K\left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N}\right] d_{it}\, d_{is} \right\}^{-1} \times \left\{ \sum_{i=1}^{N} \sum_{1\le s<t\le T_i} (\mathbf{z}_{it} - \mathbf{z}_{is})(y_{it} - y_{is}) K\left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N}\right] d_{it}\, d_{is} \right\}, \tag{8.5.4}$$

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

Tables 8.2 and 8.3 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 (8.5.1). The random-effects estimator is based on Newey's (1989) series expansion method (Charlier, Melenberg, and van Soest (2000)). The differences among these 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 specification tests for 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 percent level for the chi-square distribution with seven degrees of freedom.

Table 8.2. *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
Owners				
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)
Renters				
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 percent level; ** means significant at the 1 percent 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 8.3. *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
Owners				
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)
Renters				
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 percent level; ** means significant at the 1 percent 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).

This indicates that the model that does not allow for correlation between the error terms in the share equations ((8.5.2) and (8.5.3)) and the error term or fixed effect in the selection equation (8.5.1) is probably misspecified.

The Hausman (1978) specification test of no correlation between the household-specific effects and the xs 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 percent 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 linear panel data models or random-effects linear panel models, which only allow for very specific selection mechanisms (both of which can be estimated with just the cross-sectional data), are probably too restrictive.

8.6 DYNAMIC TOBIT MODELS

8.6.1 Dynamic Censored Models

In this section we consider censored dynamic panel data models of the form[9]

$$y_{it}^* = \gamma y_{i,t-1}^* + \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i + u_{it}, \tag{8.6.1}$$

$$y_{it} = \begin{cases} y_{it}^* & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \leq 0. \end{cases} \tag{8.6.2}$$

where the error u_{it} is assumed to be i.i.d. across i and over t. 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 $y_{i,t-1} > 0$ then $y_{i,t-1} = y_{i,t-1}^*$, (8.6.1) and (8.6.2) with $\alpha_i = 0$ become

$$\begin{aligned} y_{it}^* &= \gamma y_{i,t-1}^* + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it} \\ &= \gamma y_{i,t-1} + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}. \end{aligned} \tag{8.6.3}$$

Thus, by treating $y_{i,t-1}$ and $\mathbf{x}_{it}$ as predetermined variables that are independent of the error u_{it}, the censored estimation techniques for the cross-sectional static model discussed in Section 8.1 can be applied to the subsample where (8.6.3) hold.

When random individual-specific effects α_i are present in (8.6.1), y_{is}^* and α_i are correlated for all s even if α_i can be assumed to be uncorrelated with $\mathbf{x}_i$. 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 $\mathbf{x}_{i1}, \dots, \mathbf{x}_{it}$. The second step estimates $(\gamma, \boldsymbol{\beta}')$ from the reduced-form parameters of the y_{it}^* equation, $\boldsymbol{\pi}_t$, by a minimum-distance estimator of the form (3.8.14). To avoid the censoring problem in the first step, they suggest that for the ith individual, only the string $(y_{is}, y_{i,s-1}, \ldots, y_{i0})$, where $y_{i0} > 0, \ldots, y_{i,s-1} > 0$ be used. However, in order to derive the estimates of $\boldsymbol{\pi}_t$, the conditional distribution of y_{it}^* given $y_{i0}^*, \ldots, y_{i,t-1}^*$ will have to be assumed. Moreover, the reduced-form parameters $\boldsymbol{\pi}_t$ are related to $(\gamma, \boldsymbol{\beta}')$ 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 trimmed estimator due to Honoré (1993) and Hu (1999).

Consider the case where $T = 2$ and y_{i0} are available. In Figures 8.8 and 8.9, let the vertical axis measure the value of $y_{i2}^* - \gamma y_{i1}^* = \tilde{y}_i^*(\gamma)$ and horizontal axis measure y_{i1}^*. If u_{i1} and u_{i2} are i.i.d. conditional on $(y_{i0}^*, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$, then y_{i1}^* and $y_{i2}^* - \gamma y_{i1}^* = \tilde{y}_{i2}^*(\gamma)$ are symmetrically distributed around the line (1), $\tilde{y}_{i2}^*(\gamma) = y_{i1}^* - \gamma y_{i0}^* + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$ (which is the 45-degree line through $(\gamma y_{i0}^* + \boldsymbol{\beta}'\mathbf{x}_{i1} + \alpha_i, \boldsymbol{\beta}'\mathbf{x}_{i2} + \alpha_i)$ and $(\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}, 0)$). However, censoring destroys this symmetry. We only observe

$$\begin{aligned} y_{i1} &= \max(0, y_{i1}^*) \\ &= \max(0, \gamma y_{i0}^* + \boldsymbol{\beta}' x_{i1} + \alpha_i + u_{i1}) \end{aligned}$$

and $y_{i2} = \max(0, \gamma y_{i1}^* + \boldsymbol{\beta}'\mathbf{x}_{i2} + \alpha_i + u_{i2})$ or $\tilde{y}_{i2}(\gamma) = \max(-\gamma y_{i1}^*, y_{i2}^* - \gamma y_{i1}^*)$. That is, observations are censored from the left at the vertical axis, and for any $y_{i1} = y_{i1}^* > 0$, $y_{i2} = y_{i2}^* > 0$ implies $y_{i2}^* - \gamma y_{i1}^* > -\gamma y_{i1}^*$. In other words, observations are also censored from below by $\tilde{y}_{i2}(\gamma) = -\gamma y_{i1}$, which is line (2) in Figures 8.8 and 8.9. As shown in Figure 8.8, the observable range of y_{i1}^* and $y_{i2}^* - \gamma y_{i1}^*$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, y_{i0}^*)$ are not symmetric around line (1), which we have drawn with $\gamma \geq 0$, $\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0$. To restore symmetry, we have to find the mirror images of these two borderlines – the vertical axis and line (2) – around the centerline (1), and then symmetrically truncate observations that fall outside these two new lines.

The mirror image of the vertical axis around line (1) is the horizontal line $\tilde{y}_{i2}^*(\gamma) = -\gamma y_{i0}^* + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$, line (3) in Figure 8.8. The mirror image of line (2) around line (1) has slope equal to the reciprocal of that of line (2), $-\frac{1}{\gamma}$. Therefore, the mirror image of line (2) is the line $\tilde{y}_{i2}^*(\gamma) = -\frac{1}{\gamma} y_{i1}^* + c$ that passes through the intersection of line (1) and line (2). The intersection of line (1) and line (2) is given by $\bar{\tilde{y}}_{i2}^*(\gamma) = \bar{y}_{i1}^* - (\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}) = -\gamma \bar{y}_{i1}^*$. Solving for $(\bar{y}_{i1}^*, \bar{\tilde{y}}_{i2}^*(\gamma))$, we have $\bar{y}_{i1}^* = \frac{1}{1+\gamma}(\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$, $\bar{\tilde{y}}_{i2}^*(\gamma) = -\frac{\gamma}{1+\gamma}(\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$. Substituting $\tilde{y}_{i2}^*(\gamma) = \bar{\tilde{y}}_{i2}^*(\gamma)$ and $y_{i1}^* = \bar{y}_{i1}^*$ into the equation $\tilde{y}_{i2}^*(\gamma) = -\frac{1}{\gamma} y_{i1}^* + c$, we have $c = \frac{1-\gamma}{\gamma}(\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$. Thus the mirror image of line (2) is $\tilde{y}_{i2}^*(\gamma) = -\frac{1}{\gamma}(y_{i1}^* - \gamma y_{i0}^* + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}) - (\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$, line (4) in Figure 8.8.

In Figure 8.9 we show the construction of the symmetrical truncation region for the case when $\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0$. Since observations are truncated at the vertical axis from the left and at line (2) from below, the mirror image of the vertical axis around line (1) is given by line (3). Therefore, if we truncate observations at line (3) from below, then the remaining observations will be symmetrically distributed around line (1).

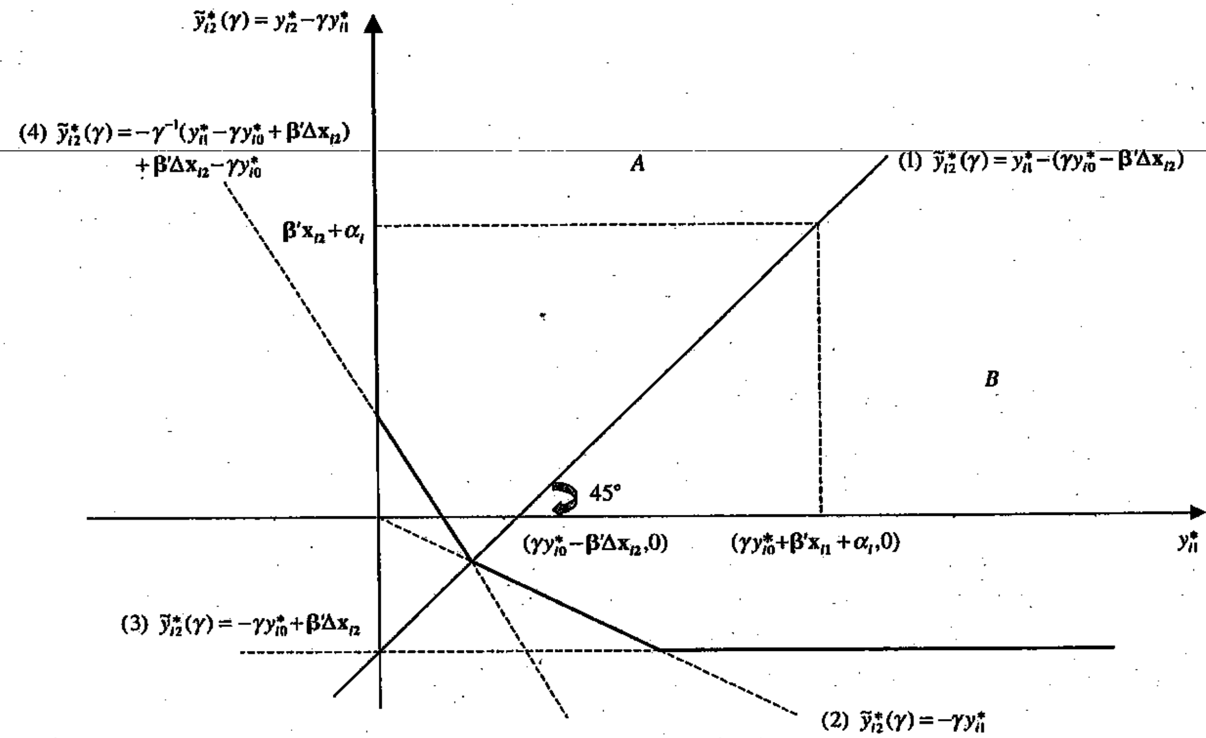

Fig. 8.8 $\gamma > 0, \gamma y^*_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0$.

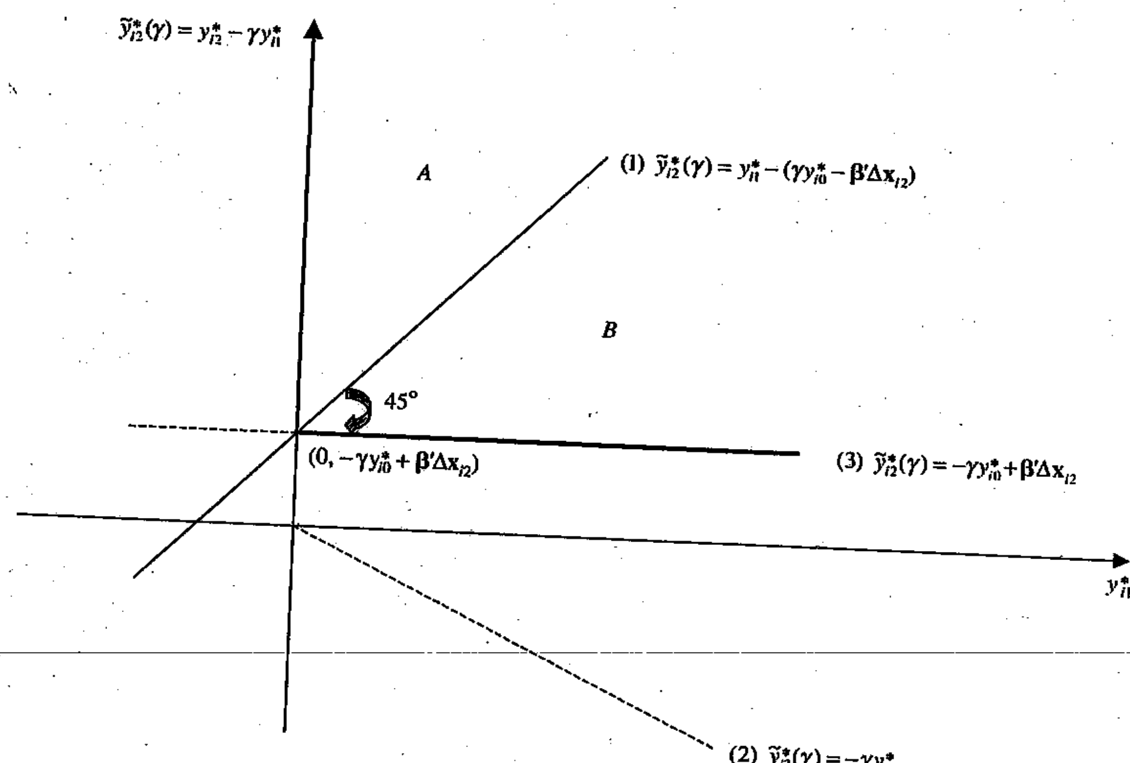

Fig. 8.9 $\gamma > 0$, $\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0$.

The observations of $(y_{i1}, \tilde{y}_{i2}(\gamma))$ falling to the northeast direction of the region bordered by the lines (2), (3), and (4) in Figure 8.8 or by the vertical axis and line (3) in Figure 8.9 are symmetrically distributed around line (1) (the 45-degree line through $(\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}, 0)$). Denote the region above the 45-degree line by A and the region below it by B. Then

$$\begin{aligned}
A \cup B \equiv \{(y_{i1}, \tilde{y}_{i2}(\gamma)) : \; & y_{i1} > 0, \;\; \tilde{y}_{i2}(\gamma) > -\gamma y_{i1}, \\
& y_{i1} > \gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} - \gamma(\tilde{y}_{i2}(\gamma) + \gamma y_{i0}^* - \boldsymbol{\beta}\Delta\mathbf{x}_{i2}), \\
& \tilde{y}_{i2}(\gamma) > -\gamma y_{i0}^* + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}\} \\
= \{(y_{i1}, \tilde{y}_{i2}(\gamma)) : \; & y_{i1} > 0, \;\; y_{i2} > 0, \\
& y_{i1} > \gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} - \gamma(\tilde{y}_{i2}(\gamma) + \gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}), \\
& \tilde{y}_{i2}(\gamma) > -\gamma y_{i0}^* + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}\}.
\end{aligned} \tag{8.6.4}$$

Symmetry implies that conditional on $y_{i0} > 0$, $y_{i1} > 0$, $y_{i2} > 0$, and $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, the probability of an observation falling in region A equals the probability of it falling in region B. That is,

$$\begin{aligned}
E\{(y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B\} \cdot [&1\{y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0\} \\
&-1\{y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0\}] = 0.
\end{aligned} \tag{8.6.5}$$

Another implication of symmetry is that conditional on $y_{i0} > 0$, $y_{i1} > 0$, $y_{i2} > 0$, and $\mathbf{x}_{i1}, \mathbf{x}_{i2}$, the expected vertical distance from a point in region A to the line (1), $\tilde{y}_{i2}(\gamma) - y_{i1} + \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$, equals the expected horizontal distance from a point in region B to that line, $y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} = -(\tilde{y}_{i2}(\gamma) - y_{i1} + \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$. Therefore,

$$E[1\{(y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B\}(y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})] = 0. \tag{8.6.6}$$

More generally, for any function $\xi(\cdot, \cdot)$ satisfying $\xi(e_1, e_2) = -\xi(e_2, e_1)$ for all (e_1, e_2), we have the orthogonality condition

$$\begin{aligned}
E[1\{(y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B\} \cdot \xi(y_{i1} - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}, \tilde{y}_{i2}(\gamma)) \\
\times h(y_{i0}, \mathbf{x}_{i1}, \mathbf{x}_{i2})] = 0,
\end{aligned} \tag{8.6.7}$$

for any function $h(\cdot)$, where

$$\begin{aligned}
1\{(y_{i1}, \tilde{y}_{i2}(\gamma)) \in A \cup B\} \equiv\; & 1\{y_{i0} > 0, \;\; y_{i1} > 0, \;\; y_{i2} > 0\} \\
\times [&1\{\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0\} \cdot 1\{y_{i1} > \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} - \gamma(\tilde{y}_{i2}(\gamma) \\
&+ \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})\} \cdot 1\{\tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}\} \\
&+ 1\{\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0\} \cdot 1\{\tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}\}].
\end{aligned} \tag{8.6.8}$$

If one chooses $h(\cdot)$ to be a constant, the case $\xi(e_1, e_2) = \mathrm{sgn}(e_1 - e_2)$ corresponds to (8.6.5), and $\xi(e_1, e_2) = e_1 - e_2$ corresponds to (8.6.6).

If $T \geq 4$, one can also consider any pair of observations y_{it}, y_{is} with $y_{i,t-1} > 0$, $y_{it} > 0$, $y_{i,s-1} > 0$, and $y_{is} > 0$. Note that conditional on $\mathbf{x}_{it}, \mathbf{x}_{is}$, the variables $(\alpha_i + u_{it})$ and $(\alpha_i + u_{is})$ are identically distributed. Thus, let

$$W_{its}(\boldsymbol{\beta}', \gamma) = \max\{0, (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}, y_{it} - \gamma y_{i,t-1}\} - \mathbf{x}_{it}'\boldsymbol{\beta}$$
$$= \max\{-\mathbf{x}_{it}'\boldsymbol{\beta}, -\mathbf{x}_{is}'\boldsymbol{\beta}, \alpha_i + u_{it}\}, \tag{8.6.9}$$

and

$$W_{ist}(\boldsymbol{\beta}', \gamma) = \max\{0, (\mathbf{x}_{is} - \mathbf{x}_{it})'\boldsymbol{\beta}, y_{is} - \gamma y_{i,s-1}\} - \mathbf{x}_{is}'\boldsymbol{\beta}$$
$$= \max\{-\mathbf{x}_{is}'\boldsymbol{\beta}, -\mathbf{x}_{it}'\boldsymbol{\beta}, \alpha_i + u_{is}\}. \tag{8.6.10}$$

Then $W_{its}(\boldsymbol{\beta}, \gamma)$ and $W_{ist}(\boldsymbol{\beta}, \gamma)$ are distributed symmetrically around the 45-degree line conditional on $(\mathbf{x}_{it}, \mathbf{x}_{is})$. This suggests the orthogonality condition

$$E[1\{y_{it-1} > 0,\ y_{it} > 0,\ y_{i,s-1} > 0,\ y_{is} > 0\}$$
$$\times\, \xi(W_{its}(\boldsymbol{\beta}', \gamma), W_{ist}(\boldsymbol{\beta}', \gamma)) \cdot h(\mathbf{x}_{it}, \mathbf{x}_{is})] = 0 \tag{8.6.11}$$

for any function $h(\cdot)$. When $T \geq 3$, the symmetric trimming procedure (8.6.11) 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 (8.6.7) or (8.6.11), Hu (1999) suggests finding a GMM estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma)'$ by minimizing $\mathbf{m}_N(\boldsymbol{\theta})' A_N \mathbf{m}_N(\boldsymbol{\theta})$, where $\mathbf{m}_N(\boldsymbol{\theta})$ is the sample analogue of (8.6.7) or (8.6.11), and A_N is a positive definite matrix that converges to a constant matrix A as $N \to \infty$. The GMM estimator will have a limiting distribution of the form

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{GMM}} - \boldsymbol{\theta}) \to N(\boldsymbol{\theta}, (\Gamma' A \Gamma)^{-1}[\Gamma' A V A \Gamma](\Gamma' A \Gamma)^{-1}), \tag{8.6.12}$$

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

However, the orthogonality conditions (8.6.5)–(8.6.7) or (8.6.11) 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 $\delta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}$ goes to infinity, the number of observations falling in the (nontruncated) region $A \cup B$ in Figures 8.7 and 8.8 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 (1999) suggests using a subset of the moments that exactly identify $\boldsymbol{\beta}$ for given γ to provide the estimates of $\boldsymbol{\beta}$, then test 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,

Table 8.4. *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	0.399	0.210	0.380
(0.030)	(0.018)	(0.129)	(0.051)

[a]Standard errors in parenthesis.
Source: Hu (1999).

the one which yields the smallest test statistic is chosen as the estimate of γ. Hu (1999) 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 8.4 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. 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 is consistent with the notion that white men had jobs that had better security and were less vulnerable to economic fluctuation than black men in the period 1957–1973.

8.6.2 Dynamic Sample Selection Models

When the selection rule is endogenously determined as given by (8.2.4) and y_{it}^* is given by (8.6.1), 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]

$$y_{it} = d_{it} y_{it}^*, \tag{8.6.13}$$

$$d_{it} = 1\{\mathbf{w}_{it}'\mathbf{a} + \eta_i + \nu_{it}\}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{8.6.14}$$

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 $\boldsymbol{\beta}$ by relying 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 Kyriazidou estimator (8.4.33) breaks down in the presence of the lagged dependent variable in (8.6.1). 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 4). 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 $\boldsymbol{\beta}$.

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

$$E(d_{it}\, d_{i,t-1}\, d_{i,t-2}\, d_{i,t-j}\, \Delta u_{it} \mid \Delta\mathbf{w}_{it}'\mathbf{a} = 0) = \mathbf{0}, \qquad j = 2, \ldots, t, \tag{8.6.15}$$

and

$$E(d_{is}d_{it}d_{i,t-1}d_{i,t-2}\, \mathbf{x}_{is}\, \Delta u_{it} \mid \Delta\mathbf{w}_{it}'\mathbf{a} = 0) = 0 \quad \text{for} \quad t = 2, \ldots, T, \quad s = 1, \ldots, T. \tag{8.6.16}$$

This is because, for an individual i, with 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.

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

$$m_{1it}(\boldsymbol{\theta}) = d_{it}d_{i,t-1}d_{i,t-2}d_{i,t-j}y_{i,t-j}(\Delta y_{it} - \Delta\mathbf{z}_{it}'\boldsymbol{\theta}), \quad t = 2, \ldots, T, \quad j = 2, \ldots, t, \tag{8.6.17}$$

$$m_{2it,k}(\boldsymbol{\theta}) = d_{is}d_{it}d_{i,t-1}d_{i,t-2}x_{is,k}(\Delta y_{it} - \Delta\mathbf{z}_{it}'\boldsymbol{\theta}), \quad t = 2, \ldots, T, \quad s = 1, \ldots, T, \quad k = 1, \ldots, K. \tag{8.6.18}$$

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

$$\hat{G}_N(\boldsymbol{\theta})'A_N\hat{G}_N(\boldsymbol{\theta}), \tag{8.6.19}$$

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

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N}K\left(\frac{\Delta\mathbf{w}_{it}'\hat{\mathbf{a}}}{h_N}\right)m_{\ell it}(\boldsymbol{\theta}), \tag{8.6.20}$$

where $K(\cdot)$ 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 $N \to \infty$. 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 speed $\sqrt{Nh_N}$.