CHAPTER 9

Miscellaneous Topics

In this chapter we consider some topics that are related to treating unobserved individual- and/or time-specific effects additively but do not fit into the previous chapters. Section 9.1 considers quantile regression analysis. Section 9.2 examines simulations methods to obtain extreme estimators. Section 9.3 looks at data with multilevel structures. Section 9.4 focuses on measurement errors. Section 9.5 analyzes the identification and estimation of distributed lag models in short panels. Section 9.6 covers rotating or randomly missing data. Section 9.7 considers repeated cross-sectional data (pseudo panel). Finally, Section 9.8 discusses the general approach for discretizing unobserved heterogeneity.

9.1 QUANTILE REGRESSION ANALYSIS

The focus of quantile analysis is not on the conditional mean but on the conditional distribution of y. The τth quantile of a random variable y, y_τ, for $0 < \tau < 1$ is defined as

$$\text{Prob}\,(y \leq y_\tau) = \int_{-\infty}^{y_\tau} f(y)dy = F(y_\tau) = \tau, \tag{9.1.1}$$

where $f(y)$ denotes the probability density function of y. The sample location quantiles estimator for the τth quantile, $0 < \tau < 1$, for N random sample y_i is the solution to the minimization problem

$$\underset{c}{\text{Min}} \left\{ \sum_{i \in \psi_c} \tau \mid y_i - c \mid + \sum_{i \in \bar{\psi}_c} (1 - \tau) \mid y_i - c \mid \right\}, \tag{9.1.2}$$

where $\psi_c = \{i \mid y_i \geq c\}$ and $\bar{\psi}_c = \{i \mid y_i < c\}$.

As $N \to \infty$, Equation (9.1.2) divided by N converges to

$$\begin{aligned} S(c) = (1 - \tau) & \int_{-\infty}^{c} \mid y - c \mid f(y)dy \\ & + (\tau) \int_{c}^{\infty} \mid y - c \mid f(y)dy. \end{aligned} \tag{9.1.3}$$

Suppose $0 < c < y_\tau$. For $y < c$, $\mid y - c \mid = \mid y - y_\tau \mid - \mid y_\tau - c \mid$. For $c < y < y_\tau$, $\mid y - c \mid = \mid y_\tau - c \mid - \mid y - y_\tau \mid$. For $y > y_\tau$, $\mid y - c \mid = \mid y - y_\tau \mid + \mid y_\tau - c \mid$. Equation (9.1.3) can be written as

$$
\begin{aligned}
S(c) = (1-\tau)\int_{-\infty}^{c} &\mid y - c \mid f(y)dy \\
+\,\tau \int_{c}^{y_\tau} &\mid y - c \mid f(y)dy \\
+\,\tau \int_{y_\tau}^{\infty} &\mid y - c \mid f(y)dy \\
= S(y_\tau) + \mid y_\tau - c \mid (\tau - F(c)) &- \int_{c}^{y_\tau} \mid y - y_\tau \mid f(y)dy \\
\geq S(y_\tau), &
\end{aligned}
\tag{9.1.4}
$$

where $S(y_\tau) = (1-\tau)\int_{-\infty}^{y_\tau} \mid y - y_\tau \mid f(y)dy + \tau\int_{y_\tau}^{\infty} \mid y - y_\tau \mid f(y)dy$. Similarly, one can show that for other values of c where $c \neq y_\tau, S(c) \geq S(y_\tau)$. Therefore, as $N \to \infty$, the solution (9.1.2) yields a consistent estimator of y_τ.

Koenker and Bassett (1978) generalize the ordinary notion of sample quantiles based on an ordering of sample observations to the regression framework

$$
\min_{\boldsymbol{b}} \left\{ \sum_{i \in \psi_b} \tau \mid y_i - \boldsymbol{x}_i'\boldsymbol{b}(\tau) \mid + \sum_{i \in \bar{\psi}_b} (1-\tau) \mid y_i - \boldsymbol{x}_i'\boldsymbol{b}(\tau) \mid \right\}, \tag{9.1.5}
$$

where $\psi_b = \{i \mid y_i \geq \boldsymbol{x}_i'\boldsymbol{b}(\tau)\}$ and $\bar{\psi}_b = \{i \mid y_i < \boldsymbol{x}_i'\boldsymbol{b}(\tau)\}$. When $\tau = \frac{1}{2}$, the quantile estimator (9.1.5) is the least absolute deviation estimator. Minimizing (9.1.5) can also be written in the form

$$
\min_{\boldsymbol{b}} \sum_{i=1}^{N} \rho_\tau(y_i - \boldsymbol{x}_i\boldsymbol{b}(\tau)), \tag{9.1.6}
$$

where $\rho_\tau(u) := [\tau - 1(u \leq 0)]u$. Equation (9.1.6) is equivalent to the linear programming form,

$$
\text{Min } [\tau \boldsymbol{e}'\boldsymbol{u}^{+} + (1-\tau)\boldsymbol{e}'\boldsymbol{u}^{-}] \tag{9.1.7}
$$

subject to

$$
\boldsymbol{y} = X\boldsymbol{b}(\tau) + \boldsymbol{u}^{+} - \boldsymbol{u}^{-}, \tag{9.1.8}
$$

$$
(\boldsymbol{u}^{+}, \boldsymbol{u}^{-}) \in R_{+}^{2N}, \tag{9.1.9}
$$

where $\boldsymbol{e}$ is an $N \times 1$ vector of $(1, \ldots, 1)$, R_{+}^{2N} denotes the positive quadrant of the $2N$ dimensional real space such that if $u_i^{+} > 0, u_i^{-} = 0$ and if $u_i^{-} > 0, u_i^{+} = 0$. Sparse linear algebra and interior point methods for solving large linear programs are essential computational tools.

The quantile estimator for the panel data model,

$$
y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{9.1.10}
$$

is the solution of

$$
\underset{\boldsymbol{b}(\tau), \alpha_i(\tau)}{\text{Min}} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_\tau(y_{it} - \boldsymbol{x}_{it}'\boldsymbol{b}(\tau) - \alpha_i(\tau)), \tag{9.1.11}
$$

where

$$Q_\tau(y_{it} \mid \boldsymbol{x}_{it}, \alpha_i) = \boldsymbol{x}'_{it}\boldsymbol{b}(\tau) + \alpha_i(\tau) \qquad (9.1.12)$$

is the τth conditional quantile.

The main idea of regression quantile is to break up the common assumption that u_{it} are independently identically distributed. The conditional quantile (9.1.12) provides information on how $\boldsymbol{x}$ influences the location, scale, and shape of the conditional distribution of the response. For instance, suppose

$$u_{it} = (1 + \boldsymbol{x}'_{it}\boldsymbol{\gamma})\epsilon_{it}, \qquad (9.1.13)$$

where $\boldsymbol{x}'_{it}\boldsymbol{\gamma} > 0$ and ϵ_{it} has the distribution function $F_\epsilon(\cdot)$. Then

$$\begin{aligned} Q_\tau(y_{it} \mid \boldsymbol{x}_{it}, \alpha_i) &= \boldsymbol{x}'_{it}(\boldsymbol{\beta} + \boldsymbol{\gamma} F_\epsilon^{-1}(\tau)) + (\alpha_i + F_\epsilon^{-1}(\tau)) \\ &= \boldsymbol{x}'_{it}\boldsymbol{\beta}(\tau) + \alpha_i(\tau). \end{aligned} \qquad (9.1.14)$$

In other words, (9.1.14) is just a straight line describing the τth quantile of y_{it} given $\boldsymbol{x}_{it}$. In this sense, one should not confuse (9.1.14) with the traditional meaning of $E(y_{it} \mid \boldsymbol{x}_{it}, \alpha_i)$ in which $\boldsymbol{x}_{it}$ and α_i are causal factors that drive the outcomes of y_{it}.

Kato, Galvao, and Montes-Rojas (2012) show that the quantile estimator of $(\boldsymbol{b}(\tau), \alpha_i(\tau))$ of (9.1.11) is consistent and asymptotically normally distributed provided $\frac{N^2\ (\log N)^3}{T} \longrightarrow 0$ as $N \longrightarrow \infty$. The requirement that the time dimension of a panel, T, grow much faster than the cross-sectional dimension, N, as N increases is because directly estimating the individual- specific effects significantly increases the variability of the estimates of $\boldsymbol{b}(\tau)$. Standard linear transformation procedures such as first differencing or mean differencing are not applicable in quantile regression. Koenker (2004) noted that shrinking the individual-specific effects towards a common mean can reduce the variability of the estimates arising from directly estimating the large number of individual-specific effects. He suggested a penalized version of (9.1.11),

$$\underset{\boldsymbol{b}(\tau), \alpha_i(\tau)}{\text{Min}} \sum_{i=1}^{N}\sum_{t=1}^{T} \rho_\tau(y_{it} - \boldsymbol{x}'_{it}\boldsymbol{b}(\tau) - \alpha_i(\tau)) + d\sum_{i=1}^{N} \mid \alpha_i(\tau) \mid . \qquad (9.1.15)$$

The penalty $d\sum_{i=1}^{N} |\alpha_i(\tau)\ |$ serves to shrink the individual-effects estimates towards zero. When $d \longrightarrow 0$, (9.1.15) yields the quantile fixed-effects estimator (9.1.11). When $d \longrightarrow \infty, \hat{\alpha}_i(\tau) \longrightarrow 0$ for all $i = 1, \ldots, N$. Minimizing (9.1.15) leads to improved performance for the estimates of the slope parameter $\boldsymbol{\beta}(\tau)$.

One trouble with either (9.1.11) or (9.1.15) is that the individual-specific effects could change because the realized value of y_{it} at different time periods could fall into different quantiles. One way to get around this problem is to view the individual-specific effect summarizing the impact of some time-invariant latent variables while the error, u_{it}, bounces the responses y_{it} around from quantile to quantile. Under this assumption, one can first take the covariance transformation to get rid of individual- and/or time-specific effects, then apply the quantile regression method to the transformed model because the covariance transformation changes only the location, not the spread of a probability distribution. Thus, for instance, instead of considering model (9.1.10), we can consider the model

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\,\boldsymbol{\beta} + (u_{it} - \bar{u}_i), \qquad (9.1.16)$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^T y_{it}$, $\bar{\boldsymbol{x}}_i = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_{it}$, and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^T u_{it}$. Then, we can estimate $\boldsymbol{b}(\tau)$ by

$$\underset{\boldsymbol{b}(\tau),c(\tau)}{\text{Min}} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau\left((y_{it} - \bar{y}_i) - (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{b}(\tau) - c(\tau)\right). \tag{9.1.17}$$

Since $c(\tau)$ is the same for all i and t, the estimator (9.1.17) reduces the variability of $\boldsymbol{b}(\tau)$ due to the variability of having to estimate $\alpha_i(\tau)$. Alternatively, one can condition not only on the observed covariates, $\boldsymbol{x}_{it}$, but also on the individual fixed effects, α_i, and replace the objective function (9.1.15) by pooling the estimates of the individual quantile through

$$\text{Min} \sum_{j=1}^J \sum_{i=1}^N \sum_{t=1}^T \omega_j \rho_{\tau_j}(y_{it} - \boldsymbol{x}_{it}'\boldsymbol{b}(\tau_j) - \alpha_i) + d\sum_{i=1}^N \mid \alpha_i \mid, \tag{9.1.18}$$

where ω_j is a relative weight given to the τ_jth quantile. Monte Carlo studies show that shrinking the unconstrained individual-specific effects toward a common value helps to achieve improved performance for the estimates of the individual-specific effects and $\boldsymbol{b}(\tau_j)$.

Although introducing the penalty factor $d\sum_{i=1}^N \mid \alpha_i(\tau) \mid$ achieves the improved performance of panel quantile estimates, deciding d is a challenging question. Lamarche (2010) shows that when the individual-specific effects α_i are independent of $\boldsymbol{x}_{it}$, the penalized quantile estimator is asymptotically unbiased and normally distributed if the individual-specific effects, α_i, are drawn from a class of zero-median distribution functions. The regularization parameter, d, can thus be selected accordingly to minimize the estimated asymptotic variance.

9.2 SIMULATION METHODS

Panel data contains at least two dimensions – a cross-sectional dimension and a time dimension. Models using panel data also often contain unobserved heterogeneity factors. To transform a latent variable model involving missing data, random coefficients, heterogeneity, etc., into an observable model often requires the integration of latent variables over multiple dimensions (e.g., Hsiao 1989, 1991a, 1991b, 1992a). The resulting panel data model estimators can be quite difficult to compute. Simulation methods have been suggested to get around the complex computational issues involving multiple integrations (e.g., Geweke 1991; Gourieroux and Monfort 1996; Hajivassiliou 1990; Hsiao and Sun 2000; Keane 1994; McFadden 1989; Pakes and Pollard 1989, Richard and Zhang 2007).

The basic idea of the simulation approach is to rely on the law of large numbers to obtain the approximation of the integrals through taking the averages of random drawings from a known probability distribution function. For instance, consider the problem of computing the conditional density function of $\boldsymbol{y}_i$ given $\boldsymbol{x}_i$, $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i;\boldsymbol{\theta})$ or some conditional moments $\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{x}_i;\boldsymbol{\theta})$, say $E(\boldsymbol{y}_i \mid \boldsymbol{x}_i;\boldsymbol{\theta})$ or $E(\boldsymbol{y}_i\boldsymbol{y}_i' \mid \boldsymbol{x}_i;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters characterizing these functions. In many cases, it is difficult to compute these functions because they do not have closed forms. However, if the conditional density or moments conditional on $\boldsymbol{x}$ and another vector $\boldsymbol{\eta}$, $f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i,\boldsymbol{\eta};\boldsymbol{\theta})$ or $\boldsymbol{m}^*(\boldsymbol{y},\boldsymbol{x} \mid \boldsymbol{\eta};\boldsymbol{\theta})$ have closed forms, and the probability distribution of $\boldsymbol{\eta}$, $P(\boldsymbol{\eta})$, is known, then from

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i;\boldsymbol{\theta}) = \int f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i,\boldsymbol{\eta};\boldsymbol{\theta})dP(\boldsymbol{\eta}), \tag{9.2.1}$$

and

$$\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \int \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}; \boldsymbol{\theta}) dP(\boldsymbol{\eta}), \tag{9.2.2}$$

we may approximate (9.2.1) and (9.2.2) by

$$\tilde{f}_H(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{H} \sum_{h=1}^{H} f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\eta}_{ih}; \boldsymbol{\theta}), \tag{9.2.3}$$

and

$$\tilde{\boldsymbol{m}}_H(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{H} \sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta}), \tag{9.2.4}$$

where $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH})$ are H random draws from $P(\boldsymbol{\eta})$.

For example, consider the random effects panel Probit and Tobit models defined by the latent response function

$$y_{it}^* = \boldsymbol{\beta}' \boldsymbol{x}_{it} + \alpha_i + u_{it}, \tag{9.2.5}$$

where α_i and u_{it} are assumed to be independently normally distributed with mean 0 and variance σ_α^2 and 1, respectively, and are mutually independent. The Probit model assumes that the observed y_{it} takes the form

$$y_{it} = \begin{cases} 1 & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \le 0. \end{cases} \tag{9.2.6}$$

The Tobit model assumes that

$$y_{it} = \begin{cases} y_{it}^* & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \le 0. \end{cases} \tag{9.2.7}$$

We note that the density function of α_i and u_{it} can be expressed as transformations of some standard distributions, here standard normal, so that the density function of $\boldsymbol{y}_i' = (y_{i1}, \ldots, y_{iT})$ becomes an integral of a conditional function over the range of these standard distributions:

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \int f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta; \boldsymbol{\theta}) dP(\eta) \tag{9.2.8}$$

with $p(\eta) \sim N(0,1)$. For instance, in the case of the Probit model,

$$f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta; \boldsymbol{\theta}) = \sum_{t=1}^{T} \Phi(\boldsymbol{x}_{it}' \boldsymbol{\beta} + \sigma_\alpha \eta_i)^{y_{it}} [1 - \Phi(\boldsymbol{x}_{it}' \boldsymbol{\beta} + \sigma_\alpha \eta_i)]^{1 - y_{it}}, \tag{9.2.9}$$

and in the case of the Tobit model,

$$\begin{aligned} f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta; \boldsymbol{\theta}) = \prod_{t \in \Psi_1} & \phi(y_{it} - \boldsymbol{x}_{it}' \boldsymbol{\beta} - \sigma_\alpha \eta_i) \\ \cdot \prod_{t \in \Psi_0} & \Phi(-\boldsymbol{x}_{it}' \boldsymbol{\beta} - \sigma_\alpha \eta_i), \end{aligned} \tag{9.2.10}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denotes the standard normal density and integrated normal, respectively, and $\Psi_1 = \{t \mid y_{it} > 0\}$ and $\Psi_0 = \{t \mid y_{it} = 0\}$. Since conditional on $\boldsymbol{x}_{it}$ and each of the H random draws of η from a standard normal distribution, $\eta_{ih}, h = 1, \ldots, H,$

the conditional density function (9.2.9) on (9.2.10) is well defined in terms of $\boldsymbol{\beta}, \sigma_\alpha^2$, the approximation of $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}, \sigma_\alpha^2)$ can be obtained by taking their averages as in (9.2.3).

Random draws of η_h from $P(\eta)$ can be obtained through the *inversion* technique from a sequence of independent uniform [0, 1] pseudo random draws

$$\eta_h = P^{-1}(\epsilon_h),$$

where $P^{-1}(\cdot)$ denote the inverse of P. For instance, if ϵ is normally distributed with mean μ and variance σ_ϵ^2, then $\eta_h = \Phi^{-1}\big(\frac{\epsilon_h - \mu}{\sigma_\epsilon}\big)$. If η is a Weibull random variable with parameters a and b, $P(\eta_h) = 1 - \exp(-b\eta_h^a)$, then $\eta_h = \big[-\frac{1}{b} ln \epsilon_h\big]^{\frac{1}{a}}$.

The generation of a multivariate $\boldsymbol{\eta}_h$ can be obtained through recursive factorization of its density into lower-dimensional density (e.g., Liesenfeld and Richard 2008). The basic idea of factorization of a k-dimensional $\boldsymbol{\eta}_h = (\eta_{1h}, \ldots, \eta_{kh})$ is to write

$$P(\boldsymbol{\eta}_h) = P(\eta_{kh} \mid \boldsymbol{\eta}^*_{k-1,h}) P(\eta_{k-1,h} \mid \boldsymbol{\eta}^*_{k-2,h}) \ldots P(\eta_{2h} \mid \eta_{1h}) P(\eta_{1h}), \qquad (9.2.11)$$

where $\boldsymbol{\eta}^*_{j,h} = (\eta_{1h}, \ldots, \eta_{jh})$. For example, random draws from an n-dimensional multivariate normal density are typically obtained based on Cholesky decomposition of its covariance matrix $\sum = \bigwedge \bigwedge'$, $\boldsymbol{\eta}_h = \bigwedge \boldsymbol{\eta}_h^*$, where $\bigwedge$ is a lower triangular matrix

$$\begin{bmatrix} a_{11} & 0 & \cdot & \cdot & 0 & 0 \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & a_{n-1,n-1} & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & a_{n,n-1} & a_{nn} \end{bmatrix}$$

and η_h^* is a random draw from a standard normal density, $N(0, 1)$.

A particularly useful technique for evaluating high-dimensional integrals is known as *importance sampling*. The idea of importance sampling is to replace $P(\boldsymbol{\eta}_i)$ by an alternative simulator with density $\mu(\cdot)$. Substituting $\mu(\cdot)$ into (9.2.11),

$$f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \int f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\eta}; \boldsymbol{\theta}) \omega(\boldsymbol{\eta}_i) \mu(\boldsymbol{\eta}_i) d\boldsymbol{\eta}_i, \qquad (9.2.12)$$

where $dP(\boldsymbol{\eta}_i) = p(\boldsymbol{\eta}_i) d\boldsymbol{\eta}_i$,

$$\omega(\boldsymbol{\eta}_i) = \frac{p(\boldsymbol{\eta}_i)}{\mu(\boldsymbol{\eta}_i)}. \qquad (9.2.13)$$

Then the corresponding Monte Carlo simulator of (9.2.13), known as the importance sampling estimator, is given by

$$\tilde{f}_H(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{H} \sum_{h=1}^{H} \omega(\boldsymbol{\eta}^*_{ih}) \mu(\boldsymbol{\eta}^*_{ih}) f^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\eta}^*_{ih}; \boldsymbol{\theta}), \qquad (9.2.14)$$

where $\boldsymbol{\eta}^*_{ih}$ are random draws from $\mu(\boldsymbol{\eta}^*_i)$.

If u_{it} in the above example follows a first-order autoregressive process,

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}, \qquad \mid \rho \mid < 1, \qquad (9.2.15)$$

then we can rewrite (9.2.5) as

$$y^*_{it} = \boldsymbol{\beta}' \boldsymbol{x}_{it} + \sigma_\alpha \eta_i + \sum_{\tau=1}^{t} a_{t\tau} \eta^*_{i\tau}, \qquad (9.2.16)$$

where $\eta^*_{i\tau}, \tau = 1, \ldots, T$ are random draws from independent $N(0,1)$, and $a_{t\tau}$ are the entries of the lower triangular matrix Λ. It turns out that here $a_{t\tau} = (1-\rho^2)^{-\frac{1}{2}}\rho^{t-\tau}$ if $t \geq \tau$ and $a_{t\tau} = 0$ if $t < \tau$.

Using the approach described above, we can obtain an unbiased, differentiable, and positive simulator of $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}), \boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_\alpha, \rho)'$, in the Probit case by considering the following drawings:

η_{ih} is drawn from N(0,1).
η^*_{i1h} is drawn from N(0,1) restricted to
$[-(\boldsymbol{\beta}'\boldsymbol{x}_{i1}+\sigma_\alpha\eta_{ih})/a_{11}, \infty]$ if $y_{i1} = 1$ and $[-\infty, -(\boldsymbol{\beta}'\boldsymbol{x}_{i1}+\sigma_\alpha\eta_{ih})/a_{11}]$ if $y_{i1} = 0$,
η^*_{i2h} is drawn from N(0,1) restricted to

$$[-(\boldsymbol{\beta}'\boldsymbol{x}_{i2} + \sigma_\alpha\eta_{ih} + a_{21}\eta^*_{i1h})/a_{22}, \infty] \text{ if } y_{i2} = 1$$

and

$$[-\infty, -(\boldsymbol{\beta}'\boldsymbol{x}_{i2} + \sigma_\alpha\eta_{ih} + a_{21}\eta^*_{i1h})/a_{22}] \text{ if } y_{i2} = 0,$$

and so on. The simulator of $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$ is

$$\tilde{f}_H(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}) = \frac{1}{H}\sum_{h=1}^{H}\prod_{t=1}^{T}\Phi\left[(-1)^{1-y_{it}}\left(\boldsymbol{\beta}'\boldsymbol{x}_{it} + \sigma_\alpha\eta_{ih} + \sum_{\tau=1}^{t-1}a_{t\tau}\eta^*_{i\tau h}\right)/a_{tt}\right], \tag{9.2.17}$$

where for $t = 1$, the sum over τ term disappears.

In the Tobit case, the same kind of method can be used. The only difference is that the simulator of $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$ becomes

$$\begin{aligned}&\tilde{f}_H(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})\\ &= \frac{1}{H}\sum_{h=1}^{H}\left\{\left[\prod_{t\in\Psi_1}\frac{1}{a_{tt}}\phi\left(\left[y_{it} - \left(\boldsymbol{\beta}'\boldsymbol{x}_{it} + \sigma_\alpha\eta_{ih} + \sum_{\tau=1}^{t-1}a_{t\tau}\eta^*_{i\tau h}\right)\right]\Big/ a_{tt}\right)\right]\right.\\ &\quad\left.\times\prod_{t\in\Psi_0}\Phi\left[-\left(\boldsymbol{\beta}'\boldsymbol{x}_{it} + \sigma_\alpha\eta_{ih} + \sum_{\tau=1}^{t-1}a_{t\tau}\eta^*_{i\tau h}\right)\Big/ a_{tt}\right]\right\}.\end{aligned} \tag{9.2.18}$$

The simulated maximum likelihood estimator (SMLE) is obtained from maximizing the simulated log-likelihood function. The simulated method of moments estimator (SMM) is obtained from the simulated moments. The simulated least squares estimator (SLS) is obtained if we let $\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = E(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$ and minimize $\sum_{i=1}^{N}[\boldsymbol{y}_i - E(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})]^2$.

Although we need $H \to \infty$ to obtain consistent simulator of $f(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})$ and $\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})$, McFadden (1989) showed that when finite H vectors $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH})$ are drawn by simple random sampling and independently for different i from the marginal density $P(\boldsymbol{\eta})$, the simulation errors are independent across observations; hence, the variance introduced by simulation will be controlled by the law of large numbers operating across observations, making it unnecessary to consistently estimate each theoretical $\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ for the consistency of SMM, $\hat{\boldsymbol{\theta}}_{SGMM}$, as $N \to \infty$.

The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\theta}}_{SMM} - \boldsymbol{\theta})$ obtained from minimizing $[\hat{\boldsymbol{m}}(\boldsymbol{\theta}) - \boldsymbol{m}(\boldsymbol{\theta})]'A[\hat{\boldsymbol{m}}(\boldsymbol{\theta}) - \boldsymbol{m}(\boldsymbol{\theta})]$ where A is a positive definite matrix such as moments of the form (3.3.14) can be approximated by

$$(R'AR)^{-1}R'AG_{NH}AR(R'AR)^{-1}, \tag{9.2.19}$$

where

$$\begin{aligned}
R &= \frac{1}{N}\sum_{i=1}^{N} W_i' \frac{\partial \tilde{\boldsymbol{m}}_H(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'},\\
G_{NH} &= \frac{1}{N}\sum_{i=1}^{N} W_i \left(\Omega + \frac{1}{H}\Delta_H\right) W_i', \qquad (9.2.20)\\
\Omega &= \text{Cov } (\boldsymbol{m}_i(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}))\\
\Delta_H &= \text{Cov } [\tilde{\boldsymbol{m}}_H(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) - \boldsymbol{m}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})],
\end{aligned}$$

and W_i is the matrix of the instruments that satisfy the orthogonality conditions, for instance in the dynamic model of the form (3.1.2), W_i could take the form (3.3.10). When $A = [\text{plim Cov } (\hat{\boldsymbol{m}}_i(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})]^{-1}$, the SMM is the simulated generalized method of moments estimator (SGMM). It is clear that as $H \to \infty$, the SGMM has the same asymptotic efficiency as the GMM. However, even with finite H, the relative efficiency of SGMM is quite high. For instance, for the simple frequency simulator, $\Delta_H = \Omega$, one draw per observation gives 50% of the asymptotic efficiency of the corresponding GMM estimator, and nine draws per observation gives 90% relative efficiency.

However, for the consistency of SMLE or SLS, we will need $H \to \infty$ as $N \to \infty$. With a finite H, the approximation error of the conditional density or moments is of order H^{-1}. This will lead to the asymptotic bias of $O(1/H)$ (e.g., Gourieroux and Monfort 1996; Hsiao, Wang and Wang 1997). Nevertheless, with a finite H it is still possible to propose an SLS estimator which is consistent and asymptotically normally distributed as $N \to \infty$ by noting that for the sequence of 2H random draws $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH}, \boldsymbol{\eta}_{i,H+1}, \ldots, \boldsymbol{\eta}_{i,2H})$ for each i,

$$\begin{aligned}
E\left[\frac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta})\right] &= E\left[\frac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{i,H+h}; \boldsymbol{\theta})\right]\\
&= \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}), \qquad (9.2.21)
\end{aligned}$$

and

$$\begin{aligned}
E&\left[\boldsymbol{y}_i - \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta})\right]'\left[\boldsymbol{y}_i - \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{i,H+h}; \boldsymbol{\theta})\right]\\
&= E\left[\boldsymbol{y}_i - \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})\right]'\left[\boldsymbol{y}_i - \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})\right], \qquad (9.2.22)
\end{aligned}$$

because of the independence between $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH})$ and $(\boldsymbol{\eta}_{i,H+1}, \ldots, \boldsymbol{\eta}_{i,2H})$. Then the SLS estimator that minimizes

$$\sum_{i=1}^{N}\left[\boldsymbol{y}_i - \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta})\right]'\left[\boldsymbol{y}_i - \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{m}^*(\boldsymbol{y}_i, \boldsymbol{x}_i \mid \boldsymbol{\eta}_{i,H+h}; \boldsymbol{\theta})\right] \qquad (9.2.23)$$

is consistent, as $N \to \infty$ even when H is fixed (e.g., Gourieroux and Monfort 1996; Hsiao and Wang 2000).

9.3 DATA WITH MULTILEVEL STRUCTURES

We have illustrated panel data methodology by assuming the presence of individual and/or time effects only. However, panel data need not be restricted to two dimensions. We can have a more complicated "clustering" or "hierarchical" structure. For example, Antweiler (2001), Baltagi, Song, and Jung (2001), and Davis (2002), following the methodology developed by Wansbeek (1982) and Wansbeek and Kapteyn (1978), consider the multi-way error components model of the form

$$y_{ij\ell t} = \boldsymbol{x}'_{ij\ell t}\boldsymbol{\beta} + v_{ij\ell t}, \tag{9.3.1}$$

for $i = 1, \dots, N, j = 1, \dots, M_i, \ell = 1, \dots, L_{ij}$, and $t = 1, \dots, T_{ij\ell}$. For example, the dependent variable $y_{ij\ell t}$ could denote the air pollution measured at station ℓ in city j of country i in time period t. This means that there are N countries, and each country i has M_i cities in which L_{ij} observation stations are located. At each station, air pollution is observed for $T_{ij\ell}$ periods. The $\boldsymbol{x}_{ij\ell t}$ denotes a vector of K explanatory variables, and the disturbance is assumed to have a multi-way error components structure,

$$v_{ij\ell t} = \alpha_i + \lambda_{ij} + \nu_{ij\ell} + \epsilon_{ij\ell t} \tag{9.3.2}$$

where $\alpha_i, \lambda_{ij}, \nu_{ij\ell}$, and $\epsilon_{ij\ell t}$ are assumed to be independently identically distributed and are mutually independent with mean zero and variances $\sigma^2_\alpha, \sigma^2_\lambda, \sigma^2_\nu$, and σ^2_ϵ, respectively.

In the case that the data are balanced, the variance–covariance matrix of $\boldsymbol{v}$, has the form

$$\Omega = \sigma^2_\alpha(I_N \otimes J_{MLT}) + \sigma^2_\lambda(I_{NM} \otimes J_{LT}) + \sigma^2_\nu(I_{NML} \otimes J_T) + \sigma^2_\epsilon I_{LMNT}, \tag{9.3.3}$$

where J_s is a square matrix of dimension s with all elements equal to 1. Rewriting (9.3.3) in the form representing the spectral decomposition Ω (e.g., as in Appendix 2B), we have

$$\begin{aligned} \Omega &= MLT\sigma^2_\alpha(I_N \otimes P_{MLT}) + LT\sigma^2_\lambda(I_{NM} \otimes P_{LT}) \\ &\quad + T\sigma^2_\nu(I_{NML} \otimes P_T) + \sigma^2_\epsilon I_{LMNT} \\ &= \sigma^2_\epsilon(I_{NML} \otimes Q_T) + \sigma^2_1(I_{NM} \otimes Q_L \otimes P_T) \\ &\quad + \sigma^2_2(I_N \otimes Q_M \otimes P_{LT}) + \sigma^2_3(I_N \otimes P_{MLT}) \end{aligned} \tag{9.3.4}$$

where $P_s = \frac{1}{s}J_s, Q_s = I_s - P_s$, and

$$\sigma^2_1 = T\sigma^2_\nu + \sigma^2_\epsilon, \tag{9.3.5}$$

$$\sigma^2_2 = LT\sigma^2_\lambda + T\sigma^2_\nu + \sigma^2_\epsilon, \tag{9.3.6}$$

$$\sigma^2_3 = MLT\sigma^2_\alpha + LT\sigma^2_\lambda + T\sigma^2_\nu + \sigma^2_\epsilon, \tag{9.3.7}$$

σ^2_ϵ are the characteristic roots of Ω. As each of the terms of (9.3.4) is orthogonal to each other and sum to I_{NMLT}, it follows that

$$\begin{aligned} \Omega^{-1/2} &= \sigma^{-1}_\epsilon(I_{NML} \otimes Q_T) + \sigma^{-1}_1(I_{NM} \otimes Q_L \otimes P_T) \\ &\qquad + \sigma^{-1}_2(I_N \otimes Q_M \otimes P_{LT}) + \sigma^{-1}_3(I_N \otimes P_{MLT}) \end{aligned} \tag{9.3.8}$$

Expanding all the Q matrices as the difference of I and P, multiplying both sides of the equation by σ_ϵ, and collecting terms yield

$$\begin{aligned}\sigma_\epsilon \Omega^{-1/2} = I_{NMLT} &- \left(1 - \frac{\sigma_\epsilon}{\sigma_1}\right)(I_{NML} \otimes P_T) \\ &- \left(\frac{\sigma_\epsilon}{\sigma_1} - \frac{\sigma_\epsilon}{\sigma_2}\right)(I_{NM} \otimes P_{LT}) \\ &- \left(\frac{\sigma_\epsilon}{\sigma_2} - \frac{\sigma_\epsilon}{\sigma_3}\right)(I_N \otimes P_{MLT}).\end{aligned} \tag{9.3.9}$$

The generalized least squares estimator (GLS) of (9.3.1) is equivalent to the least squares estimator of regressing

$$y^*_{ij\ell t} = y_{ij\ell t} - \left(1 - \frac{\sigma_\epsilon}{\sigma_1}\right)\bar{y}_{ij\ell.} - \left(\frac{\sigma_\epsilon}{\sigma_1} - \frac{\sigma_\epsilon}{\sigma_2}\right)\bar{y}_{ij..} - \left(\frac{\sigma_\epsilon}{\sigma_2} - \frac{\sigma_\epsilon}{\sigma_3}\right)\bar{y}_{i...}, \tag{9.3.10}$$

on

$$\boldsymbol{x}^*_{ij\ell t} = x_{ij\ell t} - \left(1 - \frac{\sigma_\epsilon}{\sigma_1}\right)\bar{\boldsymbol{x}}_{ij\ell.} - \left(\frac{\sigma_\epsilon}{\sigma_1} - \frac{\sigma_\epsilon}{\sigma_2}\right)\bar{\boldsymbol{x}}_{ij..} - \left(\frac{\sigma_\epsilon}{\sigma_2} - \frac{\sigma_\epsilon}{\sigma_3}\right)\bar{\boldsymbol{x}}_{i...}, \tag{9.3.11}$$

where $\bar{y}_{ij\ell.}(\bar{\boldsymbol{x}}_{ij\ell.}), \bar{y}_{ij..}(\bar{\boldsymbol{x}}_{ij..})$, and $\bar{y}_{i...}(\bar{\boldsymbol{x}}_{i...})$ indicate group averages. The application of feasible GLS can be carried out by replacing the variances in (9.3.10) and (9.3.11) by their estimates obtained from the three group-wise between estimates and the within estimate of the innermost group.

The pattern exhibited in (9.3.10) and (9.3.11) is suggestive of solutions for higher-order hierarchy with a balanced structure. If the hierarchical structure is unbalanced, the Kronecker product operation can no longer be applied. It introduces quite a bit of notational inconvenience into the algebra (e.g., Baltagi 1995, chapter 9; Wansbeek and Koning 1989). Neither can the GLS estimator be molded into a simple transformation for a least squares estimator. However, an unbalanced panel is made up of N top-level groups, each containing M_i second-level groups, the second-level groups containing the innermost L_{ij} subgroups, which in turn contain $T_{ij\ell}$ observations; the number of observations in the higher-level groups is thus $T_{ij} = \sum_{\ell=1}^{L_{ij}} T_{ij\ell}$ and $T_i = \sum_{j=1}^{M_i} T_{ij}$, and the total number of observations is $H = \sum_{i=1}^{N} T_i$. The number of top-level groups is N, the number of second-level groups is $F = \sum_{i=1}^{N} M_i$, and the number of bottom-level groups is $G = \sum_{i=1}^{N}\sum_{j=1}^{M_i} L_{ij}$. We can redefine J matrices to be block diagonal of size $H \times H$, corresponding in structure to the groups or subgroups they represent. They can be constructed explicitly by using “group membership” matrices consisting of ones and zeros that uniquely assign each of the H observations to one of the G (or F or N) groups. Antweiler (2001) derived the maximum likelihood estimator for the panels with unbalanced hierarchy.

When data contain a multilevel hierarchical structure, the application of a simple error component estimation, although inefficient, remains consistent under the assumption that the error component is independent of the regressors. However, the estimated standard errors of the slope coefficients are usually biased downward.

The component in (9.3.2), $\alpha_i, \lambda_{ij}, \nu_{ijl}$, need not be random and uncorrelated with $\boldsymbol{x}_{ijlt}$. They could be correlated with $\boldsymbol{x}_{ijlt}$ and treated as fixed constants, Balazsi et al. (2017); Matyas (2017) consider fixed effects models with various possible specifications. Moreover, the impacts of omitted variables need not be restricted to the error terms. They could also affect the coefficients of the included explanatory variables (e.g., Durlauf et al. 2001).

The estimation of a large number of parameters can produce noise and create confusion. Chapter 13 discusses some of the varying parameter formulations.

9.4 ERRORS OF MEASUREMENT

Thus far we have assumed that variables are observed without errors. Economic quantities, however, are frequently measured with errors, particularly if longitudinal information is collected through one-time retrospective surveys, which are notoriously susceptible to recall errors. If variables are indeed subject to measurement errors, exploiting panel data to control for the effects of unobserved individual characteristics using standard difference estimators (deviations from means, etc.) may result in even more biased estimates than simple least squares estimators using cross-sectional data alone.

Consider, for example, the following single-equation model (Solon 1985):

$$y_{it} = \alpha_i^* + \beta x_{it} + u_{it}, i = 1, \ldots, N, t = 1, \ldots, T, \tag{9.4.1}$$

where u_{it} is independently identically distributed, with mean zero and variance σ_u^2, and $\text{Cov}(x_{it}, u_{is}) = \text{Cov}(\alpha_i^*, u_{it}) = 0$ for any t and s, but $\text{Cov}(x_{it}, \alpha_i^*) \neq 0$. Suppose further that we observe not x_{it} itself, but rather the error-ridden measure

$$x_{it}^* = x_{it} + \tau_{it}, \tag{9.4.2}$$

where $\text{Cov}(x_{is}, \tau_{it}) = \text{Cov}\ (\alpha_i^*, \tau_{it}) = \text{Cov}(u_{it}, \tau_{is}) = 0$, and $\text{Var}(\tau_{it}) = \sigma_\tau^2$, $\text{Cov}(\tau_{it}, \tau_{i,t-1}) = \gamma_\tau \sigma_\tau^2$.

If we estimate (9.4.1) by OLS with cross-sectional data for period t, the estimator converges to (as $N \to \infty$)

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_{LS} = \beta + \frac{\text{Cov}(x_{it}, \alpha_i^*)}{\sigma_x^2 + \sigma_\tau^2} - \frac{\beta\sigma_\tau^2}{\sigma_x^2 + \sigma_\tau^2}, \tag{9.4.3}$$

where $\sigma_x^2 = \text{Var}(x_{it})$. The inconsistency of the least squares estimator involves two terms, the first due to the failure to control for the individual effects α_i^* and the second due to measurement error.

If we have panel data, say $T = 2$, we can alternatively first difference the data to eliminate the individual effects, α_i^*,

$$y_{it} - y_{i,t-1} = \beta(x_{it}^* - x_{i,t-1}^*) + [(u_{it} - \beta\tau_{it}) - (u_{i,t-1} - \beta\tau_{i,t-1})], \tag{9.4.4}$$

and then apply least squares. The probability limit of the differenced estimator as $N \to \infty$ becomes

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_d = \beta\left[1 - \frac{2(1-\gamma_\tau)\sigma_\tau^2}{\text{Var}(x_{it}^* - x_{i,t-1}^*)}\right] = \beta - \frac{\beta\sigma_\tau^2}{[(1-\gamma_x)/(1-\gamma_\tau)]\sigma_x^2 + \sigma_\tau^2}, \tag{9.4.5}$$

where γ_x is the first-order serial-correlation coefficient of x_{it}. The estimator $\hat{\beta}_d$ eliminates the first source of inconsistency but may aggravate the second. If $\gamma_x > \gamma_\tau$, the inconsistency due to measurement error is larger for $\hat{\beta}_d$ than for $\hat{\beta}_{LS}$. This occurs because if the serial correlation of the measurement error is less than that of the true x (as seems often likely to be the case), first differencing increases the noise-to-signal ratio for the measured explanatory variable.

The standard treatment for the errors-in-variables models requires extraneous information in the form of either additional data (replication and/or instrumental variables) or additional assumptions to identify the parameters of interest (e.g., Aigner et al. 1984). The repeated measurement property of panel data allows a researcher to use different transformations of the data to induce different and deducible changes in the biases in the estimated parameters that can then be used to identify the importance of measurement errors and recover the "true" parameters (Ashenfelter et al. 1984; Griliches and Hausman 1986). For instance, if the measurement error, τ_{it}, is independently identically distributed across i and t and x is serially correlated, then in the foregoing example we can use $x^*_{i,t-2}$ or $(x^*_{i,t-2} - x^*_{i,t-3})$ as instruments for $(x^*_{it} - x^*_{i,t-1})$ as long as $T > 3$. Thus, even though T may be finite, the resulting IV estimator is consistent when N tends to infinity.

Alternatively, we can obtain consistent estimates through a comparison of magnitudes of the bias arrived at by subjecting a model to different transformations (Griliches and Hausman 1986). For instance, if we use a covariance transformation to eliminate the contributions of unobserved individual components, we have

$$(y_{it} - \bar{y}_i) = \beta(x^*_{it} - \bar{x}^*_i) + [(u_{it} - \bar{u}_i) - \beta(\tau_{it} - \bar{\tau}_i)], \tag{9.4.6}$$

where $\bar{y}_i, \bar{x}^*_i, \bar{u}_i$, and $\bar{\tau}_i$ are individual time means of respective variables. Under the assumption that the measurement errors are independently identically distributed, the LS regression of (9.4.6) converges to

$$\operatorname*{plim}_{N\to\infty} \beta_w = \beta \left[1 - \frac{T-1}{T} \frac{\sigma^2_\tau}{\mathrm{Var}(x^*_{it} - \bar{x}^*_i)} \right]. \tag{9.4.7}$$

Then consistent estimators of β and σ^2_τ can be solved from (9.4.5) and (9.4.7),

$$\hat{\beta} = \left[\frac{2\hat{\beta}_w}{\mathrm{Var}\ (x^*_{it} - x^*_{i,t-1})} - \frac{(T-1)\hat{\beta}_d}{T\ \mathrm{Var}(x^*_{it} - \bar{x}^*_i)} \right] \left[\frac{2}{\mathrm{Var}(x^*_{it} - x^*_{i,t-1})} - \frac{T-1}{T\ \mathrm{Var}(x^*_{it} - \bar{x}^*_i)} \right]^{-1}, \tag{9.4.8}$$

$$\sigma^2_\tau = \frac{\hat{\beta} - \hat{\beta}_d}{\hat{\beta}} \cdot \frac{\mathrm{Var}(x^*_{it} - x^*_{i,t-1})}{2}. \tag{9.4.9}$$

In general, if the measurement errors are known to possess certain structures, consistent estimators may be available from a method of moments and/or from an IV approach by utilizing the panel structure of the data. Moreover, the first difference and the within estimators are not the only ones that will give us an implicit estimate of the bias. In fact, there are $T/2$ such independent estimates. For a six-period cross section with τ_{it} independently identically distributed, we can compute estimates of β and σ^2_τ from $y_6 - y_1, y_5 - y_2$, and $y_4 - y_3$ using the relationships

$$\begin{aligned} \operatorname*{plim}_{N\to\infty} \hat{\beta}_{61} &= \beta - 2\beta\sigma^2_\tau/\mathrm{Var}(x^*_{i6} - x^*_{i1}), \\ \operatorname*{plim}_{N\to\infty} \hat{\beta}_{52} &= \beta - 2\beta\sigma^2_\tau/\mathrm{Var}(x^*_{i5} - x^*_{i2}), \\ \operatorname*{plim}_{N\to\infty} \hat{\beta}_{43} &= \beta - 2\beta\sigma^2_\tau/\mathrm{Var}(x^*_{i4} - x^*_{i3}). \end{aligned} \tag{9.4.10}$$

Thus, there are alternative consistent estimators. This fact can be exploited to test the assumption with regard to measurement errors, which provide the rationale for the validity of the instruments, by comparing whether or not the alternative estimates of β are mutually coherent (e.g., Griliches and Hausman 1986). The moment conditions (9.4.5), (9.4.7), and (9.4.10) can also be combined together to obtain efficient estimates of β and σ_τ^2 by the use of the Chamberlain π method (Section 2.9) or the generalized method of moments estimator.

For instance, transforming $\boldsymbol{y}$ and $\boldsymbol{x}$ by the transformation matrix P_s such that $P_s \boldsymbol{e}_T = \boldsymbol{0}$ eliminates the individual effects from the model (9.4.1). Regressing the transformed $\boldsymbol{y}$ on transformed $\boldsymbol{x}$ yields an estimator that is a function of $\beta, \sigma_x^2, \sigma_\tau$ and the serial correlations of x and τ. Wansbeek and Koning (1989) provided a general formula for the estimates based on various transformation of the data by letting

$$Y^* = \boldsymbol{e}_{NT}\mu + X^*\boldsymbol{\beta} + \boldsymbol{v}^* \tag{9.4.11}$$

where $Y^* = (\boldsymbol{y}_1^{*'}, \ldots, \boldsymbol{y}_T^{*'})', \boldsymbol{y}_t^* = (y_{1t}, \ldots, y_{Nt})', X^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_T^*)'$, an $NT \times K$ matrix, $\boldsymbol{x}_t^* = (\boldsymbol{x}_{1t}', \ldots, \boldsymbol{x}_{Nt}'), \boldsymbol{v}^* = (\boldsymbol{v}_1^{*'}, \ldots, \boldsymbol{v}_T^{*'})'$, and $\boldsymbol{v}_t^* = (v_{1t} \ldots, v_{Nt})'$. Then,

$$\begin{aligned} \hat{\boldsymbol{b}}_s &= [X^{*'}(Q_s \otimes I_N)X^*]^{-1}[X^{*'}(Q_s \otimes I_N)Y^*] \\ &= \boldsymbol{\beta} + [X^{*'}(Q_s \otimes I_N)X^*]^{-1}[X^{*'}(Q_s \otimes I_N)(\boldsymbol{u}^* - \boldsymbol{\tau}^*\boldsymbol{\beta})], \end{aligned} \tag{9.4.12}$$

where $Q_s = P_s'P_s, \boldsymbol{u}^* = (\boldsymbol{u}_1^{*'}, \ldots, \boldsymbol{u}_T^{*'})', \boldsymbol{u}_t^* = (u_{1t}, \ldots, u_{Nt})', \boldsymbol{\tau}^* = (\boldsymbol{\tau}_1^*, \ldots, \boldsymbol{\tau}_T^*)'$, and $\boldsymbol{\tau}_t^* = (\boldsymbol{\tau}_{1t}, \ldots, \boldsymbol{\tau}_{Nt})$. In case the dimension of x_{it}, $K = 1$, and measurement errors are serially uncorrelated, Wansbeek and Koning (1989) showed that the m different transformed estimators $\boldsymbol{b} = (b_1, \ldots, b_m)'$

$$\sqrt{N}([\boldsymbol{b} - \beta(\boldsymbol{e}_m - \sigma_\tau^2\boldsymbol{\phi})] \sim N(\boldsymbol{0}, V), \tag{9.4.13}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)', \phi_s = (tr Q_s / tr Q_s \Sigma_{x^*})$,

$$\Sigma_{x^*} = \text{Cov}\,(\boldsymbol{x}_i^*), \boldsymbol{x}_i^* = (\boldsymbol{x}_{i1}^*, \ldots, \boldsymbol{x}_{iT}^*)',$$

$$V = F'\left\{\sigma_u^2 \Sigma_{x^*} \otimes I_T + \beta^2\sigma_\tau^2\left(\Sigma_{x^*} + \sigma_\tau^2 I_T\right) \otimes I_T\right\}F,$$

and F is the $T^2 \times m$ matrix with the sth column $\boldsymbol{f}_s = \text{vec}\,(Q_s)/(tr Q_s \Sigma_{x^*})$, where vec (A) denotes the operation of transforming an $m \times n$ matrix A into the $mn \times 1$ vector by stacking the columns of A one underneath the other (Magnus and Neudecker 1999, p. 30). Then one can obtain an efficient estimator by minimizing

$$[\boldsymbol{b} - \beta(\boldsymbol{e}_m - \sigma_\tau^2\boldsymbol{\phi})]'V^{-1}[\boldsymbol{b} - \beta(\boldsymbol{e}_m - \sigma_\tau^2\boldsymbol{\phi})], \tag{9.4.14}$$

with respect to β and σ_τ^2, which yields

$$\hat{\beta} = \left\{\frac{\boldsymbol{\phi}'V^{-1}\boldsymbol{b}}{\boldsymbol{\phi}'V^{-1}\boldsymbol{\phi}} - \frac{\boldsymbol{e}_m'V^{-1}\boldsymbol{b}}{\boldsymbol{e}_m'V^{-1}\boldsymbol{\phi}}\right\} \Big/ \left\{\frac{\boldsymbol{\phi}'V^{-1}\boldsymbol{e}_m}{\boldsymbol{\phi}'V^{-1}\boldsymbol{\phi}} - \frac{\boldsymbol{e}_m'V^{-1}\boldsymbol{e}_m}{\boldsymbol{e}_m'V^{-1}\boldsymbol{\phi}}\right\} \tag{9.4.15}$$

and

$$\hat{\sigma}_\tau^2 = \left\{\frac{\boldsymbol{\phi}'V^{-1}\boldsymbol{e}_m}{\boldsymbol{\phi}'V^{-1}\boldsymbol{b}} - \frac{\boldsymbol{e}_m'V^{-1}\boldsymbol{e}_m}{\boldsymbol{e}_m'V^{-1}\boldsymbol{b}}\right\} \Big/ \left\{\frac{\boldsymbol{\phi}'V^{-1}\boldsymbol{\phi}}{\boldsymbol{\phi}'V^{-1}\boldsymbol{b}} - \frac{\boldsymbol{e}_m'V^{-1}\boldsymbol{\phi}}{\boldsymbol{e}_m'V^{-1}\boldsymbol{b}}\right\}. \tag{9.4.16}$$

Extensions of this simple model to the serially correlated measurement errors are given by Biørn (1992, 2000) and Hsiao and Taylor (1991). Wansbeek (1978) considered simple estimators for dynamic panel data models with measurement errors. In the case of only one regressor for a linear panel data model, Wansbeek (2001) has provided a neat framework to derive the moment conditions under a variety of measurement errors assumptions by stacking the matrix of covariances between the vector of dependent variables over time and the regressors, then projecting out nuisance parameters. To illustrate the basic idea, consider a linear model,

$$\begin{aligned} y_{it} = \alpha_i^* + \beta x_{it} + \boldsymbol{\gamma}'\boldsymbol{w}_{it} + u_{it},\ i &= 1,\ldots,N,\\ t &= 1,\ldots,T, \end{aligned} \tag{9.4.17}$$

where x_{it} is not observed. Instead, one observes x_{it}^* which is related to x_{it} by (9.4.2). Suppose that the $T \times 1$ measurement error vector $\boldsymbol{\tau}_i = (\tau_{i1},\ldots,\tau_{iT})'$ is i.i.d. with mean zero and covariance matrix $\Omega = E(\boldsymbol{\tau}_i\boldsymbol{\tau}_i')$.

Suppose Ω has a structure of the form

$$\text{vec}\,\Omega = R_0\boldsymbol{\lambda}, \tag{9.4.18}$$

where vec denotes the operation that stacks the rows of a matrix one after another in a column vector form, R is a matrix of order $T^2 \times m$ with known elements, and $\boldsymbol{\lambda}$ is an $m \times 1$ vector of unknown constants. Using the covariance transformation matrix $Q = I_T - \frac{1}{T}\boldsymbol{e}_T\boldsymbol{e}_T'$ to eliminate the individual effects, α_i^*, yields

$$Q\boldsymbol{y}_i = Q\boldsymbol{x}_i + QW_i\boldsymbol{\gamma} + Q\boldsymbol{u}_i, \tag{9.4.19}$$

$$Q\boldsymbol{x}_i^* = Q\boldsymbol{x}_i + Q\boldsymbol{\tau}_i, \tag{9.4.20}$$

where $\boldsymbol{x}_i = (x_{i1},\ldots,x_{iT})', W_i = (\boldsymbol{w}_{it}')$. Let

$$R = (I_T \otimes Q)R_0. \tag{9.4.21}$$

From (9.4.2), we have

$$\begin{aligned} E(\boldsymbol{\tau}_i \otimes Q\boldsymbol{\tau}_i) &= (I_T \otimes Q)E(\boldsymbol{\tau}_i \otimes \boldsymbol{\tau}_i)\\ &= (I_T \otimes Q)R_0\boldsymbol{\lambda}\\ &= R\boldsymbol{\lambda}. \end{aligned} \tag{9.4.22}$$

It follows that

$$\begin{aligned} E(\boldsymbol{x}_i^* \otimes Q\boldsymbol{x}_i) &= E(\boldsymbol{x}_i^* \otimes Q\boldsymbol{x}_i^*) - E[(\boldsymbol{x}_i + \boldsymbol{\tau}_i) \otimes Q\boldsymbol{\tau}_i]\\ &= E(\boldsymbol{x}_i^* \otimes Q\boldsymbol{x}_i^*) - R\boldsymbol{\lambda}. \end{aligned} \tag{9.4.23}$$

Therefore,

$$E(\boldsymbol{x}_i^* \otimes Q\boldsymbol{y}_i) = E(\boldsymbol{x}_i^* \otimes Q\boldsymbol{x}_i^*)\beta + E(\boldsymbol{x}_i^* \otimes QW_i)\boldsymbol{\gamma} - \boldsymbol{R\lambda\beta}. \tag{9.4.24}$$

Equation (9.4.24) contains the nuisance parameter $\boldsymbol{\lambda}$. To eliminate $\boldsymbol{\lambda}$ from (9.4.24), multiplying $M_R = I_{T^2} - R(R'R)^{-1}R'$ to both sides of (9.4.24), we have the orthogonality conditions:

$$M_R E\{\boldsymbol{x}_i^* \otimes Q(\boldsymbol{y}_i - \boldsymbol{x}_i^*\beta - W_i\boldsymbol{\gamma})\} = \mathbf{0} \tag{9.4.25}$$

Combining (9.4.25) with the moment conditions $E(W_i'Q\boldsymbol{u}_i) = \mathbf{0}$, we have the moment conditions for the measurement error model (9.4.17)

$$E\left[M(\boldsymbol{d}_i - C_i\boldsymbol{\theta})\right] = \mathbf{0}, \tag{9.4.26}$$

where

$$M = \begin{bmatrix} M_R & \mathbf{0}\\ \mathbf{0} & I_K \end{bmatrix}, \boldsymbol{d}_i = \begin{bmatrix} \boldsymbol{x}_i^* \otimes I_T\\ W_i' \end{bmatrix} Q\boldsymbol{y}_i,$$

$$C_i = \begin{bmatrix} \boldsymbol{x}_i^* \otimes I_T\\ W_i' \end{bmatrix} Q[\boldsymbol{x}_i^*, W_i], \quad \boldsymbol{\theta}' = (\beta, \boldsymbol{\gamma}').$$

A GMM estimator is obtained by minimizing

$$\frac{1}{N}\left[\sum_{i=1}^{N} M(\boldsymbol{d}_i - C_i\boldsymbol{\theta})\right]' A_N \left[\sum_{i=1}^{N} M(\boldsymbol{d}_i - C_i\boldsymbol{\theta})\right]. \tag{9.4.27}$$

An optimal GMM estimator is to let

$$A_N^{-1} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{d}_i - C_i\hat{\boldsymbol{\theta}})(\boldsymbol{d}_i - C_i\hat{\boldsymbol{\theta}})', \tag{9.4.28}$$

where $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ such as

$$\hat{\boldsymbol{\theta}} = \left[\left(\sum_{i=1}^{N} C_i'\right) M \left(\sum_{i=1}^{N} C_i\right)\right]^{-1} \left[\left(\sum_{i=1}^{N} C_i\right)' M \left(\sum_{i=1}^{N} \boldsymbol{d}_i\right)\right]. \tag{9.4.29}$$

In the case when τ_{it} is i.i.d. across i and over t, Ω is diagonal with an equal diagonal element. Then $m = 1$ and $R_0 =$ vec I_T, $R = (I_T \otimes Q)$ vec $I_T =$ vec Q, $R'R = tr\, Q = T-1$, and $M_R = I_{T^2} - \frac{1}{T-1}$ (vec Q) (vec Q)$'$. When Ω is diagonal with distinct diagonal elements, $m = T$ and $R_0 = \boldsymbol{i}_t\boldsymbol{i}_t' \otimes \boldsymbol{i}_t$, where $\boldsymbol{i}_t$ is the tth unit vector of order T. When τ_{it} is a first-order moving average process and $T = 4$,

$$\Omega = \begin{bmatrix} a & c & 0 & 0 \\ c & b & c & 0 \\ 0 & c & b & c \\ 0 & 0 & c & a \end{bmatrix},$$

then

$$R_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix},$$

and $\boldsymbol{\lambda} = (a, b, c)'$.

In general, the identifiability of the slope parameters $\boldsymbol{\beta}$ for a linear regression model depends on whether the moment equations involving observables in levels and differences for different order of lags are sufficient to obtain a unique solution for $\boldsymbol{\beta}$ given the assumption about the autocorrelation patterns of measurement errors. For additional references, see Biørn (2000), Biørn and Klette (1998), Biørn and Krishnakumar (2008), Wansbeek (2001), and Wansbeek and Meijer (2000, section 6.6).

Panel data can also provide possibilities to identify measurement errors for nonlinear models without requiring external instruments. For instance, Kao and Schnell (1987a, 1987b) and Hsiao (1991b) discuss the identification for binary choice models and Wilhelm (2015) for the discussion of the general framework for the identification of nonlinear models. In general, the measurement errors for nonlinear models are much more difficult to handle than linear models. (e.g., Hsiao 1991b).

9.5 ESTIMATING DISTRIBUTED LAGS IN SHORT PANELS[1]

9.5.1 Introduction

Because of technical, institutional, and psychological rigidities, often behavior is not adapted immediately to changes in the variables that condition it. In most cases this adaptation is progressive. The progressive nature of adaptations in behavior can be expressed in various ways. Depending on the rationale behind it, we can set up an autoregressive model with the current value of y being a function of lagged dependent variables and exogenous variables, or we can set up a distributed-lag model, with the current value of y being a function of current and previous values of exogenous variables. Although usually a linear distributed-lag model can be expressed in an autoregressive form, and, similarly, as a rule any stable linear autoregressive model can be transformed into a distributed-lag model,[2] the empirical determination of time lags is very important in applied economics. The roles of many economic measures can be correctly understood only if we know when they will begin to take effect and when their effects will be fully worked out. Therefore, we would like to distinguish these two types of dynamic models when a precise specification (or reasoning) is possible. In Chapter 3 we discussed the issues of estimating autoregressive models with panel data. In this section we discuss estimation of distributed-lag models (Pakes and Griliches 1984).

A general distributed-lag model for a single time series of observations is usually written as

$$y_t = \mu + \sum_{\tau=0}^{\infty} \beta_\tau x_{t-\tau} + u_t, \qquad t = 1, \ldots, T, \tag{9.5.1}$$

where, for simplicity, we assume that there is only one exogenous variable, x, and, conditional on $\{x_t\}$, the u_t are independent draws from a common distribution function. When no restrictions are imposed on the lag coefficients, one cannot obtain consistent estimates of β_τ even when $T \to \infty$, because the number of unknown parameters increases with the number of observations. Moreover, the available samples often consist of fairly short time series on variables that are highly correlated over time. There is not sufficient information to obtain precise estimates of any of the lag coefficients without specifying, a priori, that all of them are functions of only a very small number of parameters (Koyck lag, Almon lag, etc.) (Dhrymes 1971; Malinvaud 1970).

On the other hand, when there are N time series, we can use cross-sectional information to identify and estimate (at least some of the) lag coefficients without having to specify a priori that the sequence of lag coefficients progresses in a particular way. For instance, consider the problem of using panel data to estimate the model (9.5.1), which for a given t we rewrite as

$$y_{it} = \mu + \sum_{\tau=0}^{t-1} \beta_\tau x_{i,t-\tau} + b_{it} + u_{it}, \qquad i = 1, \ldots, N, \tag{9.5.2}$$

[1] The material in this section is adapted from Pakes and Griliches (1984) with permission.

[2] We must point out that the errors are also transformed when we go from one form to the other (e.g., Malinvaud 1970, chapter 15).

where

$$b_{it} = \sum_{\tau=0}^{\infty} \beta_{t+\tau} x_{i,-\tau} \tag{9.5.3}$$

is the contribution of the unobserved presample x values to the current values of y, to which we shall refer as the truncation remainder. Under certain assumptions about the relationships between the unobserved b_{it} and the observed x_{it}, it is possible to obtain consistent estimates of $\beta_\tau, \tau = 0, \ldots, t-1$, by regressing (9.5.2) cross-sectionally. Furthermore, the problem of collinearity among $x_t, x_{t-1}, \ldots,$ in a single time series can be reduced or avoided by use of the cross-sectional differences in individual characteristics.

9.5.2 Common Assumptions

To see under what conditions the addition of a cross-sectional dimension can provide information that cannot be obtained in a single time series, first we note that if the lag coefficients vary across individuals $\{\beta_{i\tau}\}_{\tau=0}^{\infty}$, for $i = 1, \ldots, N$, and if there is no restriction on the distribution of these sequences over members of the population, each time series contains information on only a single sequence of coefficients. The problem of lack of information remains for panel data. Second, even if the lag coefficients do not vary across individuals ($\beta_{i\tau} = \beta_\tau$ for $i = 1, \ldots, N$ and $\tau = 0, 1, 2, \ldots$), the (often very significant) increase in sample size that accompanies the availability of panel data is entirely an increase in cross-sectional dimension. Panel data sets, in fact, usually track their observations over only a relatively short time interval. As a result, the contributions of the unobserved presample x values to the current values of y (the truncation remainder, b_{it}) are likely to be particularly important if we do not wish to impose the same type of restrictions on the lag coefficients as we often do when a single time series data set is used to estimate a distributed-lag model. Regression analysis, ignoring the unobserved truncation-remainder term, will suffer from the usual omitted-variable bias.

Thus, in order to combine N time series to estimate a distributed-lag model, we have to impose restrictions on the distribution of lag coefficients across cross-sectional units and/or on the way the unobserved presample terms affect current behavior. Pakes and Griliches (1984) considered a distributed-lag model of the form

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{\infty} \beta_{i\tau} x_{i,t-\tau} + u_{it}, \qquad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{9.5.4}$$

where u_{it} is independent of x_{is} and is independently identically distributed, with mean zero and variance σ_u^2. The coefficients of α_i^* and $\beta_{i\tau}$ are assumed to satisfy the following assumptions.

Assumption 9.5.1 $E(\beta_{i\tau}) = \beta_\tau$.

Assumption 9.5.2 *Let* $\bar{\beta}_{i\tau} = \beta_{i\tau} - \beta_\tau, \xi_{it} = \sum_{\tau=0}^{\infty} \bar{\beta}_{i\tau} x_{i,t-\tau}$, *and* $\boldsymbol{\xi}_i' = (\xi_{i1}, \ldots, \xi_{iT})$; *then* $E^*[\boldsymbol{\xi}_i \mid \boldsymbol{x}_i] = \mathbf{0}$.

Assumption 9.5.3 $E^*(\alpha_i^* \mid \boldsymbol{x}_i) = \mu + \boldsymbol{a}'\boldsymbol{x}_i$

Here $E^*(Z_1 \mid Z_2)$ refers to the minimum mean-square-error linear predictor (or the projection) of Z_1 onto Z_2; $\boldsymbol{x}_i$ denotes the vector of all observed $\boldsymbol{x}_{it}$. We assume that there are $\ell + 1$ observations on x before the first observation on y, and the $1 \times (\ell + 1 + T)$ vector $\boldsymbol{x}_i' = [x_{i,-\ell}, \dots, x_{iT}]$ is an independent draw from a common distribution with $E(\boldsymbol{x}_i \boldsymbol{x}_i') = \sum_{xx}$ positive definite.[3]

A sufficient condition for Assumption 9.5.2 to hold is that differences in lag coefficients across individuals are uncorrelated with the $\boldsymbol{x}_i$; i.e., $\beta_{i\tau}$ is a random variable defined in the sense of Swamy (1970), or see Chapter 13. However, Assumption 9.5.3 does allow for individual-specific constant terms (the α_i^*) to be correlated with $\boldsymbol{x}_i$. The combination of Assumptions 9.5.1–9.5.3 is sufficient to allow us to identify the expected value of the lag-coefficient sequence $\{\beta_\tau\}$ if both N and T tend to infinity.

If T is fixed, substituting Assumptions 9.5.1 and 9.5.2 into Equation (9.5.4), we rewrite the distributed-lag model as

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{t+\ell} \beta_\tau x_{i,t-\tau} + b_{it} + \tilde{u}_{it}, \; i = 1, \dots, N, \qquad t = 1, \dots, T, \tag{9.5.5}$$

where $b_{it} = \sum_{\tau=\ell+1}^{\infty} \beta_{t+\tau} x_{i,-\tau}$ is the truncation remainder for individual i in period t, and $\tilde{u}_{it} = \xi_{it} + u_{it}$ is the amalgamated error term satisfying $E^*[\tilde{u}_{it} \mid \boldsymbol{x}_i] = 0$. The unobserved truncation remainders are usually correlated with the included explanatory variables. Therefore, without additional restrictions, we still cannot get consistent estimates of any of the lag coefficients β_τ by regressing y_{it} on $x_{i,t-\tau}$, even when $N \to \infty$.

Because the values of the truncation remainders b_{it} are determined by the lag coefficients and the presample x values, identification requires constraints either on the lag coefficients or on the stochastic process generating these x values. Because there usually are many more degrees of freedom available in panel data, this allows us to use prior restrictions of a different kind from that in the usual approach of constraining lag coefficients to identify truncation remainders (e.g., Dhrymes 1971). In the next two subsections, we illustrate how various restrictions can be used to identify the lag coefficients.

9.5.3 Identification Using Prior Structure on the Process of the Exogenous Variable

In this subsection we consider the identification of a distributed-lag model using a kind of restriction different from that in the usual approach of constraining lag coefficients. Our interest is focused on estimating at least some of the population parameters $\beta_\tau = E(\beta_{i\tau})$ for $\tau = 0, 1, \dots,$ without restricting β_τ to be a function of a small number of parameters. We consider a lag coefficient identified if it can be calculated from the matrix of coefficients obtained from the projection of $\boldsymbol{y}_i$ onto $\boldsymbol{x}_i$, a $T \times (T + \ell + 1)$ matrix labeled Π, where $E^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i) = \boldsymbol{\mu}^* + \Pi \boldsymbol{x}_i$, $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_T^*)'$ and $\boldsymbol{y}_i' = (y_{i1}, \dots, y_{iT})$ is a $1 \times T$ vector.

Equation (9.5.5) makes it clear that each row of Π will contain a combination of the lag coefficients of interest and the coefficients from the projections of the two unobserved

[3] Note that assuming that there exist $\ell + 1$ observations on x before the first observation on y is not restrictive. If x_{it} does not exist before time period 0, we can always let $\ell = -1$. If ℓ has to be fixed, we can throw away the first $\ell + 1$ observations of y.

components, α_i^* and b_{it}, on $\boldsymbol{x}_i$. Therefore, the problem is to separate out the lag coefficients from the coefficients defining these two projections.

Using Equation (9.5.5), the projection of $\boldsymbol{y}_i$ onto $\boldsymbol{x}_i$ and α_i^* is given by[4]

$$E^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \alpha_i^*) = [B + W]\boldsymbol{x}_i + [\boldsymbol{e} + \boldsymbol{c}]\alpha_i^* \tag{9.5.6}$$

where B is the $T \times (T + \ell + 1)$ matrix of the lag coefficients

$$B = \begin{bmatrix} \beta_{\ell+1} & \beta_\ell & . & \beta_1 & \beta_0 & 0 & . & . & . & 0 \\ \beta_{\ell+2} & \beta_{\ell+1} & . & \beta_2 & \beta_1 & \beta_0 & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ \beta_{T+\ell-1} & \beta_{t+\ell-2} & . & \beta_{T+1} & \beta_T & . & . & . & \beta_0 & 0 \\ \beta_{T+\ell} & \beta_{T+\ell-1} & . & \beta_T & \beta_{T-1} & . & . & . & \beta_1 & \beta_0 \end{bmatrix}$$

W and $\mathbf{c}$ are defined by the unconstrained projection of $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iT})'$ onto $\boldsymbol{x}_i$ and α_i^*,

$$E^*[\boldsymbol{b}_i \mid \boldsymbol{x}_i, \alpha_i^*] = W\boldsymbol{x}_i + \boldsymbol{c}\alpha_i^*. \tag{9.5.7}$$

Equation (9.5.6) and the fact that $E^*\{E^*(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \alpha_i^*) \mid \boldsymbol{x}_i\} = E^*[\boldsymbol{y}_i \mid \boldsymbol{x}_i] = (\boldsymbol{e}+\boldsymbol{c})\mu + \Pi\boldsymbol{x}_i$ imply that

$$\Pi = B + [W + (\boldsymbol{e} + \boldsymbol{c})\boldsymbol{a}']. \tag{9.5.8}$$

where $\mathbf{a}$ is defined by the unconstrained projection of α_i^* onto $\boldsymbol{x}_i$, $[E^*(\alpha_i^* \mid \boldsymbol{x}_i) = \mu + \boldsymbol{a}'\boldsymbol{x}_i]$.

Clearly, if the $T \times (T + \ell + 1)$ matrix W is unrestricted, we cannot separate out the lag coefficients, B, and the impact of the truncation-remainder term from the Π matrix. But given that $\boldsymbol{ca}'$ is a matrix of rank 1, we may be able to identify some elements of B if there are restrictions on W. Thus, in order to identify some of the lag coefficients from Π, we shall have to restrict W. W will be restricted if it is reasonable to assume that the stochastic process generating $\{x_{it}\}_{t=-\infty}^T$ restricts the coefficients on $\boldsymbol{x}_i$ in the projection of the presample $x_{i,-j}$ values onto the in-sample $\boldsymbol{x}_i$ and α_i^*. The particular case analyzed by Pakes and Griliches (1984) is given by the following assumption.[5]

Assumption 9.5.4 *For $q \geq 1$, $E^*[x_{i,-\ell-q} \mid \boldsymbol{x}_i, \alpha_i^*] = c_q\alpha_i^* + \Sigma_{j=1}^p \rho_j^{(q)} x_{i,-\ell+j-1}$. That is, in the projection of the unseen presample x values onto $\boldsymbol{x}_i$ and α_i^*, only $[x_{i,-\ell}, x_{i,-\ell+1}, \ldots, x_{i,-\ell+p-1}]$ have nonzero coefficients.*

If $c_q = 0$, a sufficient condition for Assumption 9.5.4 to hold is that x is generated by a pth-order autoregressive process.[6]

Because each element of $\boldsymbol{b}_i$ is just a different linear combination of the same presample x values, the addition of Assumption 9.5.4 implies that

$$E^*[b_{it} \mid \boldsymbol{x}_i, \alpha_i^*] = c_t\alpha_i^* + \sum_{j=1}^{p} w_{t,j-\ell-1}x_{i,j-\ell-1}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{9.5.9}$$

[4] Note that we allow the projection of presample $x_{i,-\tau}$ on in-sample $\boldsymbol{x}_i$ and α_i^* to depend freely on the α_i^* by permitting each element of the vector $\boldsymbol{c}$ to be different.

[5] One can use various model-selection criteria to determine p (e.g., Amemiya 1980a).

[6] We note that $c_q = 0$ implies that α_i^* is uncorrelated with presample x_i.

where $w_{t,j-\ell-1} = \Sigma_{q=1}^{\infty}\beta_{t+\ell+q}\rho_j^{(q)}$, $j = 1, \ldots, p$, and $c_t = \Sigma_{q=1}^{\infty}\beta_{t+l+q}c_q$. This determines the vector **c** and the matrix W in (9.5.7). In particular, it implies that W can be partitioned into a $T \times (T+\ell-p+1)$ matrix of zeros and $T \times p$ matrix of free coefficients,

$$W = \begin{bmatrix} \tilde{W} & \vdots & \mathbf{0} \\ T \times p & \vdots & T \times (T+\ell-p+1). \end{bmatrix}. \tag{9.5.10}$$

Substituting (9.5.10) into (9.5.8) and taking partial derivatives of Π with respect to the leading $(T + \ell - p + 1)$ lag coefficients, we can show that the resulting Jacobian matrix satisfies the rank condition for identification of these coefficients (e.g., Hsiao 1983, Theorem 5.1.2). A simple way to check that the leading $(T + \ell - p + 1)$ lag coefficients are indeed identified is to show that consistent estimators for them exist. We note that by construction, cross-sectional regression of $\boldsymbol{y}_i$ on $\boldsymbol{x}_i$ yields consistent estimates of Π. For the special case in which $c_q = 0$, the projections of each period's value of y_{it} on all in-sample values of $\boldsymbol{x}_i' = (x_{i,-\ell}, x_{i,-\ell+1}, \ldots, x_{iT})$ are[7]

$$\begin{aligned}
E^*(y_{i1} \mid \boldsymbol{x}_i) &= \mu + \sum_{j=1}^{p}\phi_{1,j-\ell-1}x_{i,j-\ell-1}, \\
E^*(y_{i2} \mid \boldsymbol{x}_i) &= \mu + \beta_0 x_{i2} + \sum_{j=1}^{p}\phi_{2,j-\ell-1}x_{i,j-\ell-1}, \\
E^*(y_{i3} \mid \boldsymbol{x}_i) &= \mu + \beta_0 x_{i3} + \beta_1 x_{i2} + \sum_{j=1}^{p}\phi_{3,j-\ell-1}x_{i,j-\ell-1} \\
&\vdots \\
E^*(y_{iT} \mid \boldsymbol{x}_i) &= \mu + \beta_0 x_{iT} + \cdots + \beta_{T+\ell-p}x_{i,p-\ell} + \sum_{j=1}^{p}\phi_{T,j-\ell-1}x_{i,j-\ell-1},
\end{aligned} \tag{9.5.11}$$

where $\phi_{t,j-\ell-1} = \beta_{t+\ell+1-j} + w_{t,j-\ell-1}$ for $t = 1, \ldots, T$, and $j = 1, \ldots, p$, and for simplicity we have let $p = \ell + 2$. The first p values of $\boldsymbol{x}_i$ in each projection have nonzero partial correlations with the truncation remainders (the b_{it}). Hence, their coefficients do not identify the parameters of the lag distribution. Only when $(t + \ell - p + 1) > 0$, the leading coefficients in each equation are, in fact, estimates of the leading lag coefficients. As t increases, we gradually uncover the lag structure.

When $c_q \neq 0$, the finding of consistent estimators (hence identification) for the leading $(T + \ell - p + 1)$ lag coefficients is slightly more complicated. Substituting (9.5.9) into (9.5.5), we have

$$\begin{aligned}
E^*\left(y_{it} \mid \boldsymbol{x}_i, \alpha_i^*\right) = (1 + c_t)\alpha_i^* + \sum_{\tau=0}^{t+\ell-p}\beta_\tau x_{i,t-\tau} \\
+ \sum_{j=1}^{p}\phi_{t,j-\ell-1}x_{i,j-\ell-1}, \qquad t = 1, \ldots, T,
\end{aligned} \tag{9.5.12}$$

[7] The coefficient of (9.5.11) is another way of writing Π (9.5.8).

where again (for simplicity) we have assumed $p = \ell + 2$. Conditioning this equation on $\boldsymbol{x}_i$, and passing through the projection operator once more, we obtain

$$\begin{aligned}
E^*(y_{i1} \mid \boldsymbol{x}_i) &= \mu(1+c_1) + (1+c_1)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p}[(1+c_1)a_{j-\ell-1} + \phi_{1,j-\ell-1}]x_{i,j-\ell-1}, \\
E^*(y_{i2} \mid \boldsymbol{x}_i) &= \mu(1+c_2) + \beta_0 x_2 + (1+c_2)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p}[(1+c_2)a_{j-\ell-1} + \phi_{2,j-\ell-1}]x_{i,j-\ell-1}, \qquad (9.5.13) \\
&\ \ \vdots \\
E^*(y_{iT} \mid \boldsymbol{x}_i) &= \mu(1+c_T) + \sum_{\tau=0}^{T+\ell-p}\beta_\tau x_{i,T-\tau} + (1+c_T)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p}[(1+c_T)a_{j-\ell-1} + \phi_{T,j-\ell-1}]x_{i,j-\ell-1}.
\end{aligned}$$

Multiplying y_{i1} by $\tilde{c}_t$ and subtracting it from y_{it}, we produce the system of equations

$$y_{it} = \tilde{c}_t y_{i1} + \sum_{\tau=0}^{t+\ell-p}\beta_\tau x_{i,t-\tau} + \sum_{j=1}^{p}\tilde{\phi}_{t,j-\ell-1}x_{i,j-\ell-1} + \nu_{it}, \qquad (9.5.14)$$

for $t = 2, \ldots, T$, where

$$\tilde{c}_t = \frac{(1+c_t)}{1+c_1}, \quad \tilde{\phi}_{t,j-\ell-1} = \phi_{t,j-\ell-1} - \tilde{c}_t\phi_{1,j-\ell-1},$$

and

$$\nu_{it} = y_{it} - \tilde{c}_t y_{i1} - E^*(y_{it} - \tilde{c}_t y_{i1} \mid \boldsymbol{x}_i).$$

By construction, $E^*(\nu_{it} \mid \boldsymbol{x}_i) = 0$.

For a given t, the only variable on the right-hand side of (9.5.14) that is correlated with ν_{it} is y_{i1}. If we know the values of $\{\tilde{c}_t\}_{t=2}^{T}$, the system (9.5.14) will allow us to estimate the leading $(T + \ell - p + 1)$ lag coefficients consistently by first forming $\tilde{y}_{it} = y_{it} - \tilde{c}_t y_{i1}$ (for $t = 2, \ldots, T$), then regressing this sequence on in-sample x_{it} values cross-sectionally. In the case in which all c_t values are identical, we know that the sequence $\{\tilde{c}_t\}_{t=2}^{T}$ is just a sequence of ones. In the case in which α_i^* have a free coefficient in each period of the sample, we have unknown $(1 + c_t)$. However, we can consistently estimate $\tilde{c}_t$, β_τ, and $\tilde{\phi}_{t,j}$ by the instrumental-variable method, provided there is at least one x_{is} that is excluded from the determinants of $y_{it} - \tilde{c}_t y_{i1}$ and that is correlated with y_{i1}. If $T \geq 3, x_{i3}, \ldots, x_{iT}$ are excluded from the equation determining $(y_{i2} - \tilde{c}_2 y_{i1})$, and provided that not all of a_3 to a_T are zero, at least one of them will have the required correlation with y_{i1}.

We have shown that under Assumptions 9.5.1–9.5.4, the use of panel data allows us to identify the leading $T + \ell - p + 1$ lag coefficients without imposing any restrictions on the sequence $\{\beta_\tau\}_{\tau=0}^{\infty}$. Of course, if $T + \ell$ is small relative to p, we will not be able to build up much information on the tail of the lag distribution. This simply reflects the fact that short panels, by their very nature, do not contain unconstrained information on that tail. However, the early coefficients are often of significant interest in themselves. Moreover, they may provide a basis for restricting the lag structure (to be a function of a small number of parameters) in further work.

9.5.4 Identification Using Prior Structure on the Lag Coefficients

In many situations we may know that all β_τ are positive. We may also know that the first few coefficients β_0, β_1, and β_2 are the largest and that β_τ decreases with τ at least after a certain value of τ. In this subsection we show how the conventional approach of constraining the lag coefficients to be a function of a finite number of parameters can be used and generalized for identification of a distributed-lag model in the panel-data context. Therefore, we drop Assumption 9.5.4. Instead, we assume that we have prior knowledge of the structure of lag coefficients. The particular example we use here is the one assumed by Pakes and Griliches (1984), where the sequence of lag coefficients, after the first few free lags, has an autoregressive structure. This restriction is formalized as follows

Assumption 9.5.5

$$\beta_\tau = \begin{cases} \beta_\tau, & \text{for } \tau \leq k_1, \\ \sum_{j=1}^{J} \delta_j \beta_{\tau-j}, & \text{otherwise,} \end{cases}$$

where the roots of the characteristic equation $1 - \sum_{j=1}^{J} \delta_j L^j = 0$, *say,* $\lambda_1^{-1}, \ldots, \lambda_J^{-1}$, *lie outside the unit circle.*[8] *For simplicity, we assume that* $k_1 = \ell + 1$, *and that* $\lambda_1, \ldots, \lambda_J$ *are real and distinct.*

Assumption 9.5.5 implies that β_τ declines geometrically after the first k_1 lags. Solving the Jth-order difference equation

$$\beta_\tau - \delta_1 \beta_{\tau-1} - \cdots - \delta_J \beta_{\tau-J} = 0, \tag{9.5.15}$$

we obtain the general solution (e.g., Box and Jenkins 1970, chapter 3)

$$\beta_\tau = \sum_{j=1}^{J} A_j \lambda_j^\tau, \tag{9.5.16}$$

where A_j are constants to be determined by the initial conditions of the difference equation.

Substituting (9.5.16) into (9.5.5), we write the truncation-remainder term b_{it} as

$$\begin{aligned} b_{it} &= \sum_{\tau=\ell+1}^{\infty} \left(\sum_{j=1}^{J} A_j \lambda_j^{t+\tau} \right) x_{i,-\tau} \\ &= \sum_{j=1}^{J} \lambda_j^t \left(A_j \sum_{\tau=\ell+1}^{\infty} \lambda_j^\tau x_{i,-\tau} \right) \\ &= \sum_{j=1}^{J} \lambda_j^t b_{ij}, \end{aligned} \tag{9.5.17}$$

[8] The condition for the roots of the characteristics equation to lie outside the unit circle is to ensure that $\boldsymbol{\beta}_\tau$ declines geometrically as $\tau \to \infty$ (e.g., Anderson 1971, chapter 5), so that the truncation remember term will stay finite for any reasonable assumption on the x sequence.

where $b_{ij} = A_j \sum_{\tau=\ell+1}^{\infty} \lambda_j^{\tau} x_{i,-\tau}$. That is, we can represent the truncation remainder b_{it} in terms of J unobserved initial conditions $(b_{i1}, \ldots, b_{iJ})$. Thus, under Assumptions 9.5.1–9.5.3 and 9.5.5, the distributed-lag model becomes a system of T regressions with $J + 1$ freely correlated unobserved factors $(\alpha_i^*, b_{i1}, \ldots, b_{iJ})$ with J of them decaying geometrically over time.

Because the conditions for identification of a model in which there are $J+1$ unobserved factors is a straightforward generalization from a model with two unobserved factors, we deal first with the case $J = 1$ and then point out the extensions required for $J > 1$.

When $J = 1$, it is the familiar case of a modified Koyck (or geometric) lag model. The truncation remainder becomes an unobserved factor that follows an exact first-order autoregression (i.e., $b_{it} = \delta b_{i,t-1}$). Substituting this result into (9.5.5), we have

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{\ell+1} \beta_\tau x_{i,t-\tau} + \beta_{\ell+1} \sum_{\tau=\ell+2}^{t+\ell} \delta^{\tau-(\ell+1)} x_{i,t-\tau} + \delta^{t-1} b_i + \tilde{u}_{it}, \qquad (9.5.18)$$

where $b_i = \beta_{\ell+1} \sum_{\tau=1}^{\infty} \delta^{\tau} x_{i,-\tau-\ell}$.

Recall from the discussion in Section 9.5.3 that to identify the lag parameters, we require a set of restrictions on the projection matrix $E^*(\boldsymbol{b}_i \mid \boldsymbol{x}_i) = [W + \boldsymbol{c}\boldsymbol{a}']\boldsymbol{x}_i$ (Equation 9.5.7). The Koyck lag model implies that $b_{it} = \delta b_{i,t-1}$, which implies that $E^*(b_{it} \mid \boldsymbol{x}_i) = \delta E^*(b_{i,t-1} \mid \boldsymbol{x}_i)$; that is, $w_{tr} = \delta w_{t-1,r}$ for $r = 1, \ldots, T + \ell + 1$ and $t = 2, \ldots, T$. It follows that the Π matrix has the form

$$\Pi = B^* + \boldsymbol{\delta}^* \boldsymbol{w}^{*'} + \boldsymbol{e}\boldsymbol{a}', \qquad (9.5.19)$$

where $\boldsymbol{\delta}^{*'} = [1, \delta, \ldots, \delta^{T-1}]$, $\boldsymbol{w}^*$ is the vector of coefficients from the projection of b_i on $\boldsymbol{x}_i$ [i.e., $E^*(b_i \mid \boldsymbol{x}_i) = \sum_{t=-\ell}^{T} w_t^* x_{it}$], and

$$B^* = \begin{bmatrix} \beta_{\ell+1} & \cdot & \cdot & \beta_1 & \beta_0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 & 0 \\ \delta\beta_{\ell+1} & \cdot & \cdot & \beta_2 & \beta_1 & \beta_0 & \cdot & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \delta^{T-1}\beta_{\ell+1} & \cdot & \cdot & \cdot & \delta^{T-\ell-1}\beta_{\ell+1} & \delta^{T-\ell-2}\beta_{\ell+1} & \cdot & \delta\beta_{\ell+1} & \beta_{\ell+1} & \cdot & \beta_1 & \beta_0 \end{bmatrix}.$$

Taking partial derivatives of (9.5.19) with respect to unknown parameters, it can be shown that the resulting Jacobian matrix satisfies the rank condition for identification of the lag coefficients, provided $T \geq 3$ (e.g., Hsiao 1983, Theorem 5.1.2). In fact, an easy way to see that the lag coefficients are identified is to note that (9.5.18) implies that

$$\begin{aligned}(y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) &= \beta_0 x_{it} + [\beta_1 - \beta_0(1+\delta)]x_{i,t-1} \\ &\quad + \sum_{\tau=2}^{\ell}[\beta_\tau - (1+\delta)\beta_{\tau-1} + \delta\beta_{\tau-2}]x_{i,t-\tau} + v_{it}, \\ &\quad i = 1, \ldots, N, \\ &\quad t = 1, \ldots, T,\end{aligned} \tag{9.5.20}$$

where $v_{it} = \tilde{u}_{it} - (1+\delta)\tilde{u}_{i,t-1} + \delta\tilde{u}_{i,t-2}$ and $E^*[\boldsymbol{v}_i \mid \boldsymbol{x}_i] = \boldsymbol{0}$. Provided $T \geq 3, x_{i3}, \ldots, x_{iT}$ can serve as instruments for cross-sectional regression of the equation determining $y_{i2} - y_{i1}$.

In the more general case, with $J > 1$, $\boldsymbol{\delta}^*\boldsymbol{w}^{*\prime}$ in (9.5.19) will be replaced by $\sum_{j=1}^{J}\boldsymbol{\lambda}_j^*\boldsymbol{w}_j^{*\prime}$, where $\boldsymbol{\lambda}_j^{*\prime} = [1, \lambda_j, \ldots, \lambda_j^{T-1}]$, and $\boldsymbol{w}_j^*$ is the vector of coefficients from the projection of b_{ij} on $\boldsymbol{x}_i$. Using a similar procedure, we can show that the Π matrix will identify the lag coefficients if $T \geq J + 2$.

Of course, if in addition to Assumption 9.5.5 we also have information on the structure of x process, there will be more restrictions on the Π matrices than in the models in this subsection. Identification conditions can consequently be relaxed.

9.5.5 Estimation and Testing

We can estimate the unknown parameters of a distributed-lag model using short panels by first stacking all T period equations as a system of reduced-form equations:

$$\underset{T\times 1}{\boldsymbol{y}_i} = \boldsymbol{\mu}^* + [I_T \otimes \boldsymbol{x}_i']\boldsymbol{\pi} + \boldsymbol{v}_i, \qquad i = 1, \ldots, N, \tag{9.5.21}$$

where $\boldsymbol{v}_i = \boldsymbol{y}_i - E^*[\boldsymbol{y}_i \mid \boldsymbol{x}_i]$, and $\boldsymbol{\pi}' = [\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_T']$, where $\boldsymbol{\pi}_j'$ is the jth row of the matrix Π. By construction, $E(\boldsymbol{v}_i \otimes \boldsymbol{x}_i) = \boldsymbol{0}$. Under the assumption that the N vectors $(\boldsymbol{y}_i', \boldsymbol{x}_i')$ are independent draws from a common distribution, with finite fourth-order moments and with $E\boldsymbol{x}_i\boldsymbol{x}_i' = \Sigma_{xx}$ positive definite, the least squares estimator of $\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}$, is consistent, and $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ is asymptotically normally distributed, with mean zero and variance–covariance matrix Ω, which is given by (2.9.11).

The models of Sections 9.5.3 and 9.5.4 imply that $\boldsymbol{\pi} = \boldsymbol{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of the model's parameters of dimensions $m \leq (T + \ell + 1)$. We can impose these restrictions by a minimum-distance estimator that chooses $\hat{\boldsymbol{\theta}}$ to minimize

$$[\hat{\boldsymbol{\pi}} - \boldsymbol{f}(\boldsymbol{\theta})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \boldsymbol{f}(\boldsymbol{\theta})], \tag{9.5.22}$$

where $\hat{\Omega}$ is a consistent estimator of (2.9.11). Under fairly general conditions, the estimator $\hat{\boldsymbol{\theta}}$ is consistent, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed, with asymptotic variance–covariance matrix

$$(F'\Omega^{-1}F)^{-1}, \tag{9.5.23}$$

where $F = \partial\boldsymbol{f}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$. The identification condition ensures that F has rank m. The quadratic form

$$N[\hat{\boldsymbol{\pi}} - \boldsymbol{f}(\hat{\boldsymbol{\theta}})]'\Omega^{-1}[\hat{\boldsymbol{\pi}} - \boldsymbol{f}(\hat{\boldsymbol{\theta}})] \tag{9.5.24}$$

is asymptotically chi-square distributed with $T(T + \ell + 1) - m$ degrees of freedom.

Equation (9.5.24) provides us with a test of the $T(T + \ell + 1) - m$ constraints $\boldsymbol{f}(\boldsymbol{\theta})$ placed on $\boldsymbol{\pi}$. To test nested restrictions, consider the null hypothesis $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\omega})$, where $\boldsymbol{\omega}$

is a k-dimensional vector ($k \leq m$) of the parameters of the restricted model. Let $\boldsymbol{h}(\boldsymbol{\omega}) = \boldsymbol{f}[\boldsymbol{g}(\boldsymbol{\omega})]$; that is, $\boldsymbol{h}$ embodies the restrictions of the constrained model. Then, under the null hypothesis,

$$N[\hat{\boldsymbol{\pi}} - \boldsymbol{h}(\hat{\boldsymbol{\omega}})]'\Omega^{-1}[\hat{\boldsymbol{\pi}} - \boldsymbol{h}(\hat{\boldsymbol{\omega}})] \tag{9.5.25}$$

is asymptotically chi-square distributed with $T(T + \ell + 1) - k$ degrees of freedom, where $\hat{\boldsymbol{\omega}}$ minimizes (9.5.25). Hence, to test the null hypothesis, we can use the statistic[9]

$$N[\hat{\boldsymbol{\pi}} - \boldsymbol{h}(\hat{\boldsymbol{\omega}})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \boldsymbol{h}(\hat{\boldsymbol{\omega}})] - N[\hat{\boldsymbol{\pi}} - \boldsymbol{f}(\hat{\boldsymbol{\theta}})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \boldsymbol{f}(\hat{\boldsymbol{\theta}})], \tag{9.5.26}$$

which is asymptotically chi-square-distributed, with $m - k$ degrees of freedom.

To illustrate the method of estimating unconstrained distributed-lag models using panel data, Pakes and Griliches (1984) investigated empirically the issues of how to construct the "stock of capital" (G) for analysis of rates of return. The basic assumption of their model is that there exists a stable relationship between earnings (gross or net profits) (y) and past investments (x), and firms or industries differ only in terms of the level of the yield on their past investments, with the time shapes of these yields being the same across firms and implicit in the assumed depreciation formula. Namely,

$$E^*[y_{it} \mid G_{it}, \alpha_i^*] = \alpha_i^* + \gamma G_{it}, \tag{9.5.27}$$

and

$$G_{it} = \sum_{\tau=1}^{\infty} \beta_{i\tau}\, x_{it-\tau}. \tag{9.5.28}$$

Substituting (9.5.28) into (9.5.27), we have a model that consists of regressing the operating profits of firms on a distributed lag of their past investment expenditures.

Using a sample of 258 manufacturing firms' annual profit data for the years 1964-1972 and investment data for the years 1961–1971, and assuming that p in Assumption 9.5.4 equals three,[10] they found that the estimated lag coefficient rose over the first three periods and remained fairly constant over the next four or five. This pattern implies that the contribution of past investment to the capital stock first "appreciate" in the early years as investments are completed, shaken down, or adjusted to. This is distinctly different from the pattern implied by the commonly used straight-line or declining-balance depreciation formula to construct the "stock of capital." Both formulas imply that the lag coefficients decline monotonically in τ, with the decline being the greatest in earlier periods for the second case.

9.6 ROTATING OR RANDOMLY MISSING DATA

In many situations we do not have complete time series observations on cross-sectional units. Instead, individuals are selected according to a "rotating" scheme that can be briefly stated as follows: Let all individuals in the population be numbered consecutively. Suppose the sample in period 1 consists of individuals $1, 2, \ldots, N$. In period 2, individuals $1, \ldots, m_1$ ($0 \leq m_1 \leq N$) are replaced by individuals $N+1, \ldots, N+m_1$. In period 3, individuals $m_1 + 1, \ldots, m_1 + m_2$ ($0 \leq m_2 \leq N$) are replaced by individuals $N + m_1 + 1, \ldots, N + m_1 + m_2$, and so on. This procedure of dropping the first m_{t-1} individuals from the sample

[9] See Neyman (1949) or Hsiao (1985b).

[10] Thus, they assume that this year's investment does not affect this year's profits and that there are two presample observations ($\ell = 1$) on investment.

selected in the previous period and augmenting the sample by drawing m_{t-1} individuals from the population so that the sample size remains the same continues through all periods. Hence, for T periods, although the total number of observations remains at NT, we have observed $N + \sum_{t=1}^{T-1} m_t$ individuals.

"Rotation" of a sample of micro units over time is quite common. It can be caused by deliberate policy of the data-collecting agency (e.g., the Bureau of the Census) because of the worry that if the number of times respondents have been exposed to a survey gets long, the data may be affected, and even behavioral changes may be induced. Or, it can arise because of the consideration of optimal sample design so as to gain as much information as possible from a given budget (e.g., Aigner and Balestra 1988; Nijman, Verbeek, and van Soest 1991). It can also arise because the data-collecting agency can neither force nor persuade randomly selected individuals to report more than once or twice, particularly if detailed and time-consuming reporting is required. For example, the Survey of Income and Program Participation, which began fieldwork in October 1983, has been designed as an ongoing series of national panels, each consisting of about 20,000 interviewed households and having a duration of 2.5 years. Every four months the Census Bureau will interview each individual of age 15 years or older in the panel. Information will be collected on a monthly basis for most sources of money and nonmoney income, participation in various governmental transfer programs, labor-force status, and household composition.

Statistical methods developed for analyzing complete panel data can be extended in a straightforward manner to analyze rotating samples if rotation is by design (i.e., randomly dropping and addition of individuals) and if a model is static and the error terms are assumed to be independently distributed across cross-sectional units. The likelihood function for the observed samples in this case is simply the product of the $N + \sum_{t=1}^{T-1} m_t$ joint density of $(y_{it_i}, y_{i,t_i+1}, \ldots, y_{iT_i})$,

$$L = \prod_{i=1}^{N+\sum_{t=1}^{T-1} m_t} f(y_{it_i}, \ldots, y_{iT_i}), \tag{9.6.1}$$

where t_i and T_i denote the first and the last periods during which the ith individual was observed. Apart from the minor modifications of t_i for 1 and T_i for T, (9.6.1) is basically of the same form as the likelihood functions for the complete panel data.

As an illustration, we consider a single-equation error-components model (Biørn 1981). Let

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{9.6.2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{x}_{it}$ are $k \times 1$ vectors of parameters and explanatory variables, respectively, and

$$v_{it} = \alpha_i + u_{it}. \tag{9.6.3}$$

The error terms α_i and u_{it} are independent of each other and are independently distributed, with zero means and constant variances σ_α^2 and σ_u^2, respectively. For ease of exposition, we assume that α_i and u_{it} are uncorrelated with $\boldsymbol{x}_{it}$.[11] We also assume that in each period a fixed number of individuals are dropped out of the sample and the same number of

[11] If α_i are correlated with $\boldsymbol{x}_{it}$, we can eliminate the linear dependence between α_i and $\boldsymbol{x}_{it}$ by assuming $\alpha_i = \Sigma_t \boldsymbol{a}_t' \boldsymbol{x}_{it} + \epsilon_i$. For details, see Chapter 2 or Mundlak (1978a).

individuals from the population are added back to the sample (namely, $m_t = m$ for all t). Thus, the total number of individuals observed is

$$H = (T-1)m + N. \tag{9.6.4}$$

Denote the number of times the ith individual is observed by q_i, then $q_i = T_i - t_i + 1$. Stacking the time series observations for the ith individual in vector form, we have

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{v}_i, \tag{9.6.5}$$

where

$$\underset{q_i\times 1}{\boldsymbol{y}_i} = (y_{it_i}, \ldots, y_{iT_i})', \ \underset{q_i\times k}{X_i} = (\boldsymbol{x}_{it}'),$$

$$\boldsymbol{v}_i = (\alpha_i + u_{it_i}, \ldots, \alpha_i + u_{iT_i})'.$$

The variance–covariance matrix of v_i is

$$V_i = \sigma_u^2 + \sigma_\alpha^2 \quad \text{if } q_i = 1 \tag{9.6.6a}$$

and is

$$V_i = E\boldsymbol{v}_i\boldsymbol{v}_i' = \sigma_u^2 I_{q_i} + \sigma_\alpha^2 J_i \quad \text{if } q_i > 1, \tag{9.6.6b}$$

where J_i is a $q_i \times q_i$ matrix with all elements equal to 1. Then, for $q_i = 1$,

$$V_i^{-1} = (\sigma_u^2 + \sigma_\alpha^2)^{-1}, \tag{9.6.7a}$$

and for $q_i > 1$,

$$V_i^{-1} = \frac{1}{\sigma_u^2}\left[I_{q_i} - \frac{\sigma_\alpha^2}{\sigma_u^2 + q_i\sigma_\alpha^2}J_i\right]. \tag{9.6.7b}$$

Because $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are uncorrelated, the variance–covariance matrix of the stacked equations $(y_1', \ldots, y_{N+(T-1)m}')'$ is block diagonal. Therefore, the GLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\sum_{i=1}^{N+(T-1)m} X_i'V_i^{-1}X_i\right]^{-1}\left[\sum_{i=1}^{N+(T-1)m} X_i'V_i^{-1}\boldsymbol{y}_i\right]. \tag{9.6.8}$$

The GLS estimator of $\boldsymbol{\beta}$ is equivalent to first premultiplying the observation matrix $[\boldsymbol{y}_i, X_i]$ by P_i, where $P_i'P_i = V_i^{-1}$ then regressing $P_i\boldsymbol{y}_i$ on P_iX_i (Theil 1971, chapter 6). In other words, the least squares method is applied to the data transformed by the following procedure: For individuals who are observed only once, multiply the corresponding y's and $\boldsymbol{x}$'s by $(\sigma_u^2 + \sigma_\alpha^2)^{-1/2}$. For individuals who are observed q_i times, subtract from the corresponding y's and $\boldsymbol{x}$'s a fraction $1 - [\sigma_u/(\sigma_u^2 + q_i\sigma_\alpha^2)^{1/2}]$ of their group means, $\bar{y}_i$ and $\bar{\boldsymbol{x}}_i$, where $\bar{y}_i = (1/q_i)\sum_t y_{it}$ and $\bar{\boldsymbol{x}}_i = (1/q_i)\sum_t \boldsymbol{x}_{it}$, then divide them by σ_u.

To obtain separate estimates σ_u^2 and σ_α^2, we need at least one group for which $q_i > 1$. Let $\ominus$ denote the set of those individuals with $q_i > 1$, $\ominus = \{i \mid q_i > 1\}$, and $H^* = \sum_{i\in\ominus} q_i$. Then σ_u^2 and σ_α^2 can be consistently estimated by

$$\hat{\sigma}_u^2 = \frac{1}{H^*}\sum_{i\in\ominus}\sum_{t=t_i}^{Ti}[(y_{it} - \bar{y}_i) - \hat{\boldsymbol{\beta}}'(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)]^2, \tag{9.6.9}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{N+(T-1)m}\sum_{i=1}^{N+(T-1)m}\left[(\bar{y}_i - \hat{\boldsymbol{\beta}}'\bar{\boldsymbol{x}}_i)^2 - \frac{1}{q_i}\hat{\sigma}_u^2\right]. \tag{9.6.10}$$

Similarly, we can apply the MLE by maximizing the logarithm of the likelihood function (9.6.1):

$$\begin{aligned}\log L = &-\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{N+(T-1)m}\log \mid V_i \mid \\ &-\frac{1}{2}\sum_{i=1}^{N+(T-1)m}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'V_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta}) \\ = &-\frac{NT}{2}\log 2\pi - \frac{1}{2}\left[\sum_{i=1}^{N+(T-1)m}(q_i-1)\right]\log\sigma_u^2 \\ &-\frac{1}{2}\sum_{i=1}^{N+(T-1)m}\log(\sigma_u^2 + q_i\sigma_\alpha^2) \\ &-\frac{1}{2}\sum_{i=1}^{N+(T-1)m}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'V_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta}).\end{aligned} \tag{9.6.11}$$

Conditioning on σ_u^2 and σ_α^2, the MLE is the GLS (9.6.8). Conditioning on $\boldsymbol{\beta}$, the MLEs of σ_u^2 and σ_α^2 are the simultaneous solutions of the following equations:

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma_u^2} = &-\frac{1}{2\sigma_u^2}\left[\sum_{i=1}^{N+(T-1)m}(q_i-1)\right] \\ &-\frac{1}{2}\left[\sum_{i=1}^{N+(T-1)m}\frac{1}{(\sigma_u^2+q_i\sigma_\alpha^2)}\right] \\ &+\frac{1}{2\sigma_u^4}\sum_{i=1}^{N+(T-1)m}(y_i - X_i\boldsymbol{\beta})'Q_i(y_i - X_i\boldsymbol{\beta}) \\ &+\frac{1}{2}\sum_{i=1}^{N+(T-1)m}\frac{q_i}{(\sigma_u^2+q_i\sigma_\alpha^2)^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})^2 = 0\end{aligned} \tag{9.6.12}$$

and

$$\frac{\partial \log L}{\partial \sigma_\alpha^2} = -\frac{1}{2}\sum_{i=1}^{N+(T-1)m}\left[\frac{q_i}{\sigma_u^2+q_i\sigma_\alpha^2} - \frac{q_i^2}{(\sigma_u^2+q_i\sigma_\alpha^2)^2}(\bar{y}_i - \bar{\boldsymbol{x}}_i'\boldsymbol{\beta})^2\right] = 0 \tag{9.6.13}$$

where $Q_i = I_{q_i} - (1/q_i)\boldsymbol{e}_{q_i}\boldsymbol{e}_{q_i}'$, and $\boldsymbol{e}_{q_i}$ is a $q_i \times 1$ vector of ones. Unfortunately, because q_i are different for different i, (9.6.12) and (9.6.13) cannot be put in the simple form of (2.3.25) and (2.3.26). Numerical methods will have to be used to obtain a solution. However, computation of the MLEs of $\boldsymbol{\beta}$, σ_u^2, and σ_α^2 can be simplified by iteratively switching between (9.6.8) and (9.6.12)–(9.6.13).

If α_i are treated as fixed constants, $\boldsymbol{\beta}$ of (9.6.2) can be consistently estimated through the within transformation,

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=t_i}^{Ti}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=t_i}^{Ti}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(y_{it} - \bar{y}_i)\right]. \tag{9.6.14}$$

If the model is dynamic, similar modification of the GMM (e.g., (3.3.17); Collado 1997; Moffitt 1993) can be applied to obtain consistent estimators of the coefficients. The likelihood approach for dynamic models has the issue of initial conditions.[12] Different assumptions about initial conditions will suggest different ways of incorporating new observations with those already in the sample (e.g. (3.4.12) or Assumption 9.5.5). If α_i are treated as random, it would appear a reasonable approximation in this case is to modify the methods based on the assumption that initial observations are correlated with individual effects and have stationary variances. However, the assumption imposed on the model will have to be even more restrictive. If α_i are treated as fixed, similar modification can be applied to the transform MLE (e.g., (3.5.7)).

When data are randomly missing, a common procedure is to focus on the subset of individuals for which complete time series observations are available. However, the subset of incompletely observed individuals also contains some information about unknown parameters. A more efficient and computationally somewhat more complicated way is to treat randomly missing samples in the same way as rotating samples. For instance, the likelihood function (9.6.1), with the modification that $t_i = 1$ for all i, can also be viewed the likelihood function for this situation: In time period 1, there are $N + \sum_{t=1}^{T-1} m_t$ individuals; in period 2, m_1 of them randomly drop out, and so on, such that at the end of T periods, only N individuals are remaining in the sample. Thus, the procedure for obtaining the GLS or MLE for unknown parameters with all the observations utilized is similar to the situation of rotating samples.

To test if attrition is indeed random, we note that either the complete sample unbalanced panel estimators discussed above or the estimators based on the balanced panel subsample estimators converge to the true value under the null. Under the alternative that attrition is behaviorally related, neither set of estimators is consistent. However, if the individual-specific effects α_i and the error u_{it} are independent of the regressors $\boldsymbol{x}_{it}$ and are independently normally distributed, a test of random attrition versus behaviorally related attrition is a student t-test of the significance of sample selection effect (e.g., (7.2.7)). If α_i are correlated with $\boldsymbol{x}_{it}$, one can construct a Hausman (1978)-type test statistic for the significance of the difference between the Kyriazidou (1997) fixed-effects sample selection estimator (e.g., (7.5.4)) and the complete sample unbalanced panel data with estimator (9.6.14). Further, if all initial samples are observed for at least two periods before attrition occurs, then the within estimator based on initial complete samples within estimator and the within estimator based on all observed samples (unbalanced panel) converge to the true value under the null and converge to different values under the alternative. A straightforward Hausman (1978) test statistic,

$$(\hat{\boldsymbol{\beta}}_{cv} - \tilde{\boldsymbol{\beta}}_{cvs})' \left[\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_{cvs}) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{cv})\right]^{-1} (\hat{\boldsymbol{\beta}}_{cv} - \tilde{\boldsymbol{\beta}}_{cvs}) \tag{9.6.15}$$

[12] For details, see Chapters 3 and 5.

can be used to test the null of attrition being random, where $\tilde{\boldsymbol{\beta}}_{cvs}$ and $\text{Cov}(\tilde{\boldsymbol{\beta}}_{cvs})$ denote the within estimator of $\boldsymbol{\beta}$ and its covariance matrix based on the initial sample from period 1 to t^*, where t^* denotes the last time period before any attrition (at period t^*+1) occurs.

9.7 PSEUDO PANELS (OR REPEATED CROSS-SECTIONAL DATA)

In many situations there could be no genuine panel where specific individuals or firms are followed over time. However, repeated cross-sectional surveys may be available, where random samples are taken from the population at consecutive points in time. The major limitation of repeated cross-sectional data is that individual histories are not available, so it is not possible to control the impact of unobserved individual characteristics in a linear model of the form

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha_i + u_{it}, \tag{9.7.1}$$

if α_i and $\boldsymbol{x}_{it}$ are correlated through the fixed-effects estimator discussed in Chapter 2.[13] However, several authors have argued that with some additional assumptions, $\boldsymbol{\beta}$ may be identifiable from a single cross-section or a series of independent cross-sections (e.g., Blundell, Browning, and Meghir 1994; Deaton 1985; Heckman and Robb 1985; Moffitt 1993).

Deaton (1985) suggests to use a *cohort approach* to obtain consistent estimators of $\boldsymbol{\beta}$ of (9.7.1) if repeated cross-sections of data are available. In this approach individuals sharing common observed characteristics, such as age, sex, education, or socioeconomic background are grouped into *cohorts*. For instance, suppose that one can divide the sample into C cohorts in terms of an $L \times 1$ vector of individual characteristics, $z_c, c = 1, \dots, C$. Let z_{it} be the corresponding L-dimensional vector of individual-specific variables for the ith individual of the tth cross-sectional data. Then $(y_{it}, \boldsymbol{x}_{it})$ belong to the cth cohort if $z_{it} \in z_c$. Let $\psi_{ct} = \{i \mid z_{it} \in z_c$ for the tth cross-sectional data$\}$ be the set of individuals that belong to the cohort c at time t, $c = 1, \dots, C, t = 1, \dots, T$. Let N_{ct} be the number of individuals in ψ_{ct}. Deaton (1985) assumes individuals belonging to the same cohort have the same specific effects,

$$\alpha_i = \sum_{c=1}^{C} \alpha_c d_{itc}, \tag{9.7.2}$$

where $d_{itc} = 1$ if the ith individual of the tth cross-sectional data belongs to cohort c and 0 otherwise. Let $\bar{y}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} y_{it}$ and $\bar{\boldsymbol{x}}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} \boldsymbol{x}_{it}$, then the data $(\bar{y}_{ct}, \bar{\boldsymbol{x}}'_{ct})$ becomes a pseudo panel with repeated observations on C cohorts over T time periods. Aggregation of observations to cohort level for the model (9.7.1) leads to

$$\bar{y}_{ct} = \bar{\boldsymbol{x}}'_{ct}\boldsymbol{\beta} + \alpha_c + \bar{u}_{ct}, \quad c = 1, \dots, C, t = 1, \dots, T, \tag{9.7.3}$$

where $\bar{u}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} u_{it}$.

If $\boldsymbol{x}_{it}$ are uncorrelated with u_{it}, the within estimator (2.2.10) can be applied to the pseudo panel

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_c)(\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_c)' \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_c)(\bar{y}_{ct} - \bar{y}_c) \right), \tag{9.7.4}$$

[13] If α_i and $\boldsymbol{x}_{it}$ are uncorrelated, there is no problem of consistently estimating $\boldsymbol{\beta}$ with repeated cross-sectional data because $E(\alpha_i + u_{it} | \boldsymbol{x}_{it}) = 0$.

where $\bar{\boldsymbol{x}}_c = \frac{1}{T}\sum_{t=1}^{T}\bar{\boldsymbol{x}}_{ct}$, and $\bar{y}_c = \frac{1}{T}\sum_{t=1}^{T}\bar{y}_{ct}$. When $T \rightarrow \infty$ or if T is fixed but $N \rightarrow \infty, C \rightarrow \infty$, and $\frac{C}{N} \longrightarrow 0$, (9.7.4) is consistent.

Although the cohort approach offers a useful framework to make use of independent cross-sectional information, there are problems with some of its features. First, the assertion that intra-cohort homogeneity (9.7.2) appears very strong, in particular, in view of the cohort classification is often arbitrary. Second, the practice of establishing the large sample properties of econometric estimators and test statistics by assuming that the number of cohorts, C, tends to infinity is not satisfactory. There is often a physical limit beyond which one cannot increase the number of cohorts. The oft-cited example of date of birth cohorts is a case in point. Third, grouping or aggregating individuals may result in the loss of information. Moreover, in general, the number of individuals at different cohorts or different times are different, $N_{ct} \neq N_{c's}$. Even u_{it} is homoscedastic and independently distributed Var $(\bar{u}_{ct}) = \frac{\sigma_u^2}{N_{ct}} \neq$ Var $(\bar{u}_{c's}) = \frac{\sigma_u^2}{N_{c's}}$. Therefore, the t-statistic based on the conventional within estimator formula is not asymptotically standard normally distributed unless $N_{ct} = N_{c's}$ for all $c, c', t, s,$ and var (u_{it}) is a constant across i. Hence, the resulting inference can be misleading (Inoue 2008).

Suppose (9.7.2) indeed holds and, if u_{it} is independently identically distributed, the problem of heterocesdasticity of $\bar{u}_{ct}$ can be corrected by applying the weighted within estimator,

$$\hat{\boldsymbol{\beta}}_{ww} = \left\{\sum_{c=1}^{C}\sum_{t=1}^{T}\left[N_{ct}(\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_i)(\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_c)'\right]\right\}^{-1} \cdot \left\{\sum_{c=1}^{C}\sum_{t=1}^{T}\left[N_{ct}(\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_c)(\bar{y}_{ct} - \bar{y}_c)\right]\right\}. \tag{9.7.5}$$

The variance covariance matrix of $\hat{\boldsymbol{\beta}}_{ww}$ is

$$\text{Cov }(\hat{\boldsymbol{\beta}}_{ww}) = \sigma^2\left\{\sum_{c=1}^{C}\sum_{t=1}^{T}\left[N_{ct}(\bar{\boldsymbol{x}}_{ct} - \bar{\boldsymbol{x}}_c)(\bar{x}_{ct} - \bar{\boldsymbol{x}}_c)'\right]\right\}^{-1}. \tag{9.7.6}$$

A cohort approach also raises a complicated issue for the estimation of a dynamic model of the form,

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}, \tag{9.7.7}$$

because $y_{i,t-1}$ is unavailable. The cohort approach will have to use the y-values of other individuals observed at $t-1$ to predict the missing $y_{i,t-1}$, $\hat{y}_{i,t-1}$. Suppose there exists a set of instruments z_{it} such that the orthogonal projection of y_{it} on z_{it} are available,

$$E^*(y_{it} \mid z_{it}) = z_{it}'\boldsymbol{\delta}_t, \tag{9.7.8}$$

where $E^*(y \mid z)$ denotes the minimum mean-square-error linear predictor of y by z. Let $\hat{y}_{i,t-1} = z_{i,t-1}'\hat{\boldsymbol{\delta}}_{t-1}$, then (9.7.7) becomes

$$y_{it} = \gamma\hat{y}_{i,t-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{9.7.9}$$

where

$$v_{it} = \alpha_i + u_{it} + \gamma(y_{i,t-1} - \hat{y}_{i,t-1}). \tag{9.7.10}$$

Girma (2000), Moffitt (1993), and McKenzie (2004) assume that $z_{it} = z_c$, are a set of cohort dummies for all i belonging to cohort c. This is equivalent to simply using the dummy variable d_{itc}, as instruments for y_{it} where $d_{itc} = 1$ if y_{it} belongs to cohort c and zero otherwise, for $c = 1, \ldots, C$. Taking the average of y_{it} or $\hat{y}_{it}$ for i belonging to cohort c leads to the following pseudo panel dynamic model:

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \alpha_c + \bar{\boldsymbol{x}}'_{ct}\boldsymbol{\beta} + v_{ct}, \; c = 1, \ldots, C, \qquad t = 1, \ldots, T, \tag{9.7.11}$$

where all variables denote period-by-period averages within each cohort. The covariance estimator of (9.7.11) would be consistent estimators of γ and $\boldsymbol{\beta}$ provided

$$\text{Cov}\,(v_{ct}, \bar{y}_{c,t-1}) = 0, \tag{9.7.12}$$

and

$$\text{Cov}\,(v_{ct}, \boldsymbol{x}_{ct}) = \mathbf{0}. \tag{9.7.13}$$

However, even under the assumption (9.7.2),

$$E[(\alpha_i + u_{it})z_{i,t-1} \mid i \in \psi_{c,t-1}] = \mathbf{0}, \tag{9.7.14}$$

in general,

$$E[(y_{i,t-1} - \hat{y}_{i,t-1})\hat{y}_{i,t-1}] \neq 0. \tag{9.7.15}$$

Moreover, as pointed out by Verbeek (2007) and Verbeek and Vella (2005), although under the exogeneity assumption

$$\text{Cov}\,[(\alpha_i + u_{it})\boldsymbol{x}_{it}] = \mathbf{0}, \tag{9.7.16}$$

(9.7.13) is unlikely to hold because $\boldsymbol{x}_{i,t-1}$ drives $y_{i,t-1}$ and $\boldsymbol{x}_{it}$ is likely to be serially correlated. To overcome the problem of correlations between the regressors and errors in (9.7.11), one will have to also find instruments for $\boldsymbol{x}_{it}$ as well. Unfortunately, the availability of such instruments in addition to z_i in many applications may be questionable (e.g., Verbeek and Vella 2005). It remains to be seen whether in empirical applications of cohort approach suitable instruments can be found that have time-varying relationships with $\boldsymbol{x}_{it}$ and $y_{i,t-1}$, while in the meantime they should not have any time-varying relationship with the error term (9.7.10) (e.g., Verbeek 2007; Verbeek and Vella 2005).

The application of cohort approach to nonlinear model is also complicated. For instance, consider a nonlinear model of the form

$$y_{it} = Q(\boldsymbol{x}_{it}) + \alpha_i + u_{it}, \quad i = 1, \ldots, N_t. \tag{9.7.17}$$

Even though the assumption that all individuals in a given cohort, say the gth cohort, have $\alpha_i = \alpha_g$, for all the $(y_{it}, \boldsymbol{x}_{it}) \in g$,

$$\begin{aligned} y_{gt} &= \frac{1}{N_{gt}} \sum_{i=1}^{N_t} y_{it}\; 1(y_{it} \in g) \\ &= \frac{1}{N_{gt}} \sum_{i=1}^{N_t} Q(\boldsymbol{x}_{it})\; 1(\boldsymbol{x}_{it} \in g)\; + \alpha_g + u_{gt}, \end{aligned} \tag{9.7.18}$$

where $1(\cdot)$ denotes the indicator function, $N_{gt} = \sum_{i=1}^{N_t} 1(y_{it} \in g)$. The term

$$\begin{aligned} \frac{1}{N_{gt}} \sum_{i=1}^{N_t} Q(\boldsymbol{x}_{it}) 1(\boldsymbol{x}_{it} \in g) &\longrightarrow \int Q(\boldsymbol{x}_{it}) f(\boldsymbol{x}_{it} \mid \boldsymbol{x}_{it} \in g) d\boldsymbol{x}_{it} \\ &\neq Q(\boldsymbol{x}_{gt}), \end{aligned} \tag{9.7.19}$$

where $\boldsymbol{x}_{gt} = \frac{1}{N_{gt}} \sum_{i=1}^{N_t} \boldsymbol{x}_{it} 1(\boldsymbol{x}_{it} \in g)$. To implement the cohort approach to the nonlinear model, one has to find a way to transform a nonlinear model into a linear model, as in Section 6.4, under fairly strong conditions.

9.8 DISCRETIZING UNOBSERVED HETEROGENEITY

The cohort approach discussed in Section 9.7 can be viewed as a special form of the general approach of *discretizing unobserved heterogeneity* to reduce the dimension about population unobservables, as suggested by Bonhomme and Manresa (2015) and Bonhomme et al. (2019). They consider the conditional density of y_{it} taking the form,

$$\log f_i(\boldsymbol{\alpha}_i, \boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(y_{it} | y_{i,t-1}, \boldsymbol{x}_{it}, \alpha_{it}, \boldsymbol{\theta}) \tag{9.8.1}$$

where α_{it} denotes the unobserved heterogeneity for agent i at time t, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iT})$, and $\boldsymbol{\theta}$ is common over i and t. The densities of exogenous variables, $\boldsymbol{x}_{it}$, conditional on $\boldsymbol{\mu}_i = (\mu_{01}, \ldots, \mu_{iT})$, take the form,

$$\log g_i(\boldsymbol{\mu}_i) = \sum_{t=1}^{T} \log g(\boldsymbol{x}_{it} | \boldsymbol{x}_{i,t-1}, \mu_{it}). \tag{9.8.2}$$

They assume the unobserved heterogeneity in the data of $(y_{it}, \boldsymbol{x}_{it})$, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\mu}_i$, are generated from a low dimensional (latent) vector $\boldsymbol{\xi}_i$. Specifically,

1. There exist (latent) vector $\boldsymbol{\xi}_i$ of dimension d, vector $\boldsymbol{\lambda}_i$ of dimension d_λ and two functions $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ such that $\alpha_{it} = \boldsymbol{\alpha}(\boldsymbol{\xi}_i, \boldsymbol{\lambda}_i)$ and $\boldsymbol{\mu}_i = \boldsymbol{\mu}(\boldsymbol{\xi}_i, \boldsymbol{\lambda}_i)$.
2. There exist vectors $\boldsymbol{h}_i$, and a function $\boldsymbol{\psi}$ such that $\underset{S\to\infty}{\operatorname{plim}}\, \boldsymbol{h}_i = \boldsymbol{\psi}(\boldsymbol{\xi}_i)$, $\frac{1}{N}\sum_{i=1}^{N} \| \boldsymbol{h}_i - \boldsymbol{\psi}(\boldsymbol{\xi}_i) \|^2 = O_p(\frac{1}{S})$ as (N, S) tend to infinity. Moreover, there exists a function $\boldsymbol{\phi}$ such that $\boldsymbol{\xi}_i = \boldsymbol{\phi}(\boldsymbol{\psi}(\boldsymbol{\xi}_i))$.

Assumption 1 assumes the individual-time varying unobserved heterogeneity $(\alpha_{i1}, \ldots, \alpha_{iT})$ is generated from a smooth low-dimensional probability density function. When $\boldsymbol{\alpha}_i$ is time-varying, they assume that $\boldsymbol{\alpha}_i$ follow a factor structure as discussed in Chapter 10, among many others. Assumption 2 requires the individual moment $\boldsymbol{h}_i$ to be informative about $\boldsymbol{\xi}_i$ in the sense that for large $S, \boldsymbol{\xi}_i$ can be *uniquely* recovered from $\boldsymbol{h}_i$ so the different types of $\boldsymbol{\xi}_i$, $\boldsymbol{\xi}_i^*$, or $\boldsymbol{\xi}_i^{**}$ can be separated by their moments.

Under both Assumptions 1 and 2, Bonhomme et al. (2019) suggest a two-step grouped fixed-effects estimator (GFE) for $\boldsymbol{\theta}$. The first step is to partition the individual units of $\boldsymbol{\alpha}_i$ into K groups, corresponding to group indicators $\hat{k}_i \in \{1, \ldots, K\}$ that approximate the moments $\boldsymbol{h}_i$ in the following sense:

$$(\hat{\boldsymbol{h}}, \hat{k}_1, \ldots, \hat{k}_N) = \arg \min_{(\tilde{\boldsymbol{h}}, k_1, \ldots, k_N)} \sum_{i=1}^{N} \| \boldsymbol{h}_i - \tilde{\boldsymbol{h}}_i(k_i) \|^2 , \tag{9.8.3}$$

when $\{k_i\}$ are partitions of $\{1, \ldots, N\}$ into at most K groups, $\tilde{\boldsymbol{h}} = (\tilde{h}(1)', \ldots, \tilde{h}(K)')'$ where $\tilde{\boldsymbol{h}}(k)$ are vectors, and $\| \cdot \|$ denotes the Euclidean norm. Note that $\hat{\boldsymbol{h}}(k)$ is simply the mean of $\boldsymbol{h}_i$ in group $\hat{k}_i = k$. The solution of (9.8.3) is through the machine learning *kmeans* algorithm (see Chapter 14 for further discussion of algorithm for *kmeans*). The partition of N $\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_{i1}, \ldots, \boldsymbol{\alpha}_{iT})'$ into K groups is to reduce the dimension of fixed NT

unobserved heterogeneity into KT unobserved heterogeneity. The second step of the GFE is to estimate (9.8.1) conditional on $\hat{k}_i$.

The second step of the GFE is to estimate (9.8.1) conditional on $\hat{k}_i$. The maximization is now with respect to K vectors of $\boldsymbol{\alpha}_i$ instead of N. Bonhomme et al. (2019) show that when d is small, a moderate number of groups K is suffice to guarantee a low approximation error. However, the approximation errors increase with the dimension of the latent type d. To reduce the bias of group membership indicators, they use the half-panel jackknife method of Dhaene and Jochmans (2015) (see Chapter 6 for a discussion). To control the size of the approximation error of heterogeneity, they propose the number of groups K to grow with the sample size using a data-driven rule for K,

$$\hat{K} = \min_{K \geq 1} \left\{ K : \hat{Q}(K) \leq \gamma \hat{V}_h \right\}, \tag{9.8.4}$$

where

$$\hat{Q}(K) = \frac{1}{N} \sum_{i=1}^{N} \parallel \boldsymbol{h}_i - \hat{\boldsymbol{h}}_i(\hat{k}_i) \parallel^2 , \tag{9.8.5}$$

$$\hat{V}_h = E \left\{ \parallel \boldsymbol{h}_i - \boldsymbol{\psi}(\boldsymbol{\xi}_i) \parallel^2 \ + o_p \left(\frac{1}{S} \right) \right\} \tag{9.8.6}$$

and $\gamma > 0$ is a user-specific tuning parameter. Bonhomme et al. (2019) recommend setting $\gamma = 1$ as a default and checking how GFE estimates vary when taking $\gamma < 1$; that is, when using a larger K.