CHAPTER 3

Dynamic Models with Additive Specific Effects

3.1 INTRODUCTION

In the last chapter we discussed the implications of treating the specific effects as fixed or random and the associated estimation methods for the linear static model

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha_i + \lambda_t + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{3.1.1}$$

where $\boldsymbol{x}_{it}$ is a $K \times 1$ vector of strictly exogenous explanatory variables (with respect to u_{it}), including the constant term; $\boldsymbol{\beta}$ is a $K \times 1$ vector of constants; α_i and λ_t are the (unobserved) individual and time-specific effects, which are assumed to stay constant for given i over t and for given t over i, respectively; and u_{it} represents the effects of those unobserved variables that vary over i and t. The consistency and asymptotic distribution of the least square dummy variable (covariance) estimator or the generalized least squares estimator is independent of the way N or T goes to infinity as long as $NT \to \infty$. Moreover, when the effects are correlated with the regressors, although the generalized least squares estimater under the assumption of the independence between the random effects and explanatory variables is biased, a properly modified generalized least squares estimater remains consistent and asymptotically normally distributed whether N or T or both go to infinity (e.g. Mundlak 1978a; also see Section 2.4).

The inertia in human behavior and institutional and technological rigidities have led many economists to consider "all interesting economic behavior is inherently dynamic, dynamic panel models are the only relevant models; what might superficially appear to be a static model only conceals underlying dynamics, since any state variables presumed to influence present behavior is likely to depend in some way on past behavior" (M. Nerlove 2002, p. 46). When models containing lagged dependent variables such as[1]

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + v_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{3.1.2}$$

$$v_{it} = \alpha_i + \lambda_t + u_{it}, \tag{3.1.3}$$

where $Eu_{it} = 0$, and $Eu_{it}u_{js} = \sigma_u^2$, if $i = j$ and $t = s$, and $Eu_{it}u_{js} = 0$ otherwise, strict exogeneity of the regressors no longer holds. The presence of time-invariant, individual-specific effects creates correlations between v_{is} and y_{it} for all t. The consistency and

[1] We defer the discussion of estimating distributed-lag models to Chapter 9.

asymptotic properties of various estimators depend on the way N or $T \rightarrow \infty$ or, if both N and $T, (N, T) \rightarrow \infty$, their relative speed, $\frac{N}{T}$. The assumption with regard to the initial observations, y_{io}, could also play a role in the asymptotic distribution. We discuss the properties of the least squares or least squares dummy variable estimator in section 3.2. The method of moments approach is discussed in Section 3.3. The likelihood approach when the individual effects are treated as random or fixed is discussed in Sections 3.4 and 3.5 respectively. Section 3.6 comments on the likelihood approach vs the method of moments approach. In Section 3.7 we relax the assumption on the specific serial-correlation structure of the error term and discuss a system approach to estimating dynamic models. Models with both individual- and time-specific effects are discussed in Section 3.8.

3.2 THE LEAST SQUARES AND THE LEAST SQUARES DUMMY VARIABLE (COVARIANCE) ESTIMATOR

Consider the model,

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}' \boldsymbol{x}_{it} + \boldsymbol{\rho}' \boldsymbol{z}_i + v_{it}, |\gamma| < 1, \tag{3.2.1}$$

$$v_{it} = \alpha_i + u_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{3.2.2}$$

and y_{io} observable.[2] We assume $\boldsymbol{x}_{it}$ are $K_1 \times 1$ individual-time varying variables and $\boldsymbol{z}_i$ are $K_2 \times 1$ individual varying but time-invariant strictly exogenous variables with respect to u_{it}, and that u_{it} represents the impact of omitted variables that are independently, identically distributed over i and t with mean 0 and constant variance σ_u^2.

We note that even α_i are uncorrelated with $(\boldsymbol{x}_{it}, \boldsymbol{z}_i)$, v_{is} are correlated with all y_{it} for all t due to the presence of time-invariant individual specific effects, α_i. When $E(v_{it} | \boldsymbol{x}_{it}, \boldsymbol{z}_i, y_{i,t-1}) \neq 0$, regressing y_{it} on $\boldsymbol{x}_{it}, \boldsymbol{z}_i$ and $y_{i,t-1}$ is not consistent.

3.2.1 Bias of the Least Squares Estimater (LS)

To get an idea of the bias of the LS, we assume $\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\rho} = \mathbf{0}$, then model (3.2.1) becomes

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + u_{it}, \qquad |\, \gamma \,| < 1, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{3.2.3}$$

where u_{it} is independently, identically distributed over i and t. The OLS estimator of γ is

$$\hat{\gamma}_{LS} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \cdot y_{i,t-1}}{\sum_{i=1}^{N} \sum_{t=1}^{T} y_{i,t-1}^2} = \gamma + \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (\alpha_i + u_{it}) y_{i,t-1}}{\sum_{i=1}^{N} \sum_{t=1}^{T} y_{i,t-1}^2}. \tag{3.2.4}$$

[2] The assumption that $|\gamma| < 1$ is made to establish the (weak) stationarity of an autoregressive process (Anderson 1971, chapters 5 and 7). A stochastic process $\{\xi_t\}$ is stationary if its probability structure does not change with time. A stochastic process is weakly stationary if its mean $E\xi_t = m$ is a constant, independent of its time, and if the covariance of any two variables $E(\xi_t - E\xi_t)(\xi_s - E\xi_s) = \sigma_\xi(t-s)$ depends only on their distance apart in time. The statistical properties of a least-squares estimator for the dynamic model vary with whether or not $|\, \gamma \,| < 1$ when $T \rightarrow \infty$ (Anderson 1959). When T is fixed and $N \rightarrow \infty$, it is not necessary to assume that $|\, \gamma \,| < 1$ to establish the asymptotic normality of the least squares estimator (Anderson 1978; Goodrich and Caines 1979). We keep this conventional assumption for simplicity of exposition and also because it allows us to provide a unified approach toward various assumptions about the initial conditions discussed in Section 3.4.

The asymptotic bias of the OLS estimator is given by the probability limit of the second term on the right-hand side of (3.2.4). By continuous substitution,

$$y_{it} = \frac{1-\gamma^t}{1-\gamma}\alpha_i + \sum_{j=0}^{t-1} u_{i,t-j}\gamma^j + \gamma^t y_{io}. \tag{3.2.5}$$

Under the assumption that α_i are independent of u_{it}, the numerator of the second term of (3.2.4) divided by NT converges to

$$\begin{aligned}
&\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\alpha_i + u_{it})y_{i,t-1} \\
&\quad = \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\mathrm{Cov}(y_{i0},\alpha_i) + \frac{1}{T}\frac{\sigma_\alpha^2}{(1-\gamma)^2}\left[T(1-\gamma)-(1-\gamma^T)\right],
\end{aligned} \tag{3.2.6}$$

where $\sigma_\alpha^2 = \text{plim}\frac{1}{N}\sum_{i=1}^N \alpha_i^2$ whether α_i is treated as fixed constants or random variables. The denominator converges to

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2 &= \frac{1-\gamma^{2T}}{T(1-\gamma^2)} \cdot \text{plim}\frac{\sum_{i=1}^N y_{i0}^2}{N} \\
&+ \frac{\sigma_\alpha^2}{(1-\gamma)^2}\cdot\frac{1}{T}\left(T - 2\frac{1-\gamma^T}{1-\gamma} + \frac{1-\gamma^{2T}}{1-\gamma^2}\right) \\
&+ \frac{2}{T(1-\gamma)}\left(\frac{1-\gamma^T}{1-\gamma} - \frac{1-\gamma^{2T}}{1-\gamma^2}\right)\mathrm{Cov}(\alpha_i, y_{i0}) \\
&+ \frac{\sigma_u^2}{T(1-\gamma^2)^2}\left[T(1-\gamma^2)-(1-\gamma^{2T})\right].
\end{aligned} \tag{3.2.7}$$

Usually, y_{i0} are assumed either to be arbitrary constants or to be generated by the same process as any other y_{it}, so that $\mathrm{Cov}(y_{i0},\alpha_i)$ is either zero or positive.[3] Under the assumption that the initial values are bounded, namely, that $\text{plim}_{N\to\infty}\sum_{i=1}^N y_{i0}^2/N$ is finite, the OLS method overestimates the true autocorrelation coefficient γ when N or T or both tend to infinity whether α_i is a fixed constant or a random variable. The overestimation is more pronounced the greater the variation of the individual effects, σ_α^2. This asymptotic result also tends to hold in finite samples according to the Monte Carlo studies conducted by Nerlove (1967) ($N = 25$, $T = 10$).

The addition of exogenous variables to a first-order autoregressive process does not alter the direction of bias of the estimator of the coefficient of the lagged dependent variable, although its magnitude is somewhat reduced. The estimator of the coefficient of the lag dependent variable remains biased upward, and the estimated coefficients of the exogenous variables are biased downward. Formulas for the asymptotic bias of the OLS estimator for a pth-order autoregressive process and for a model also containing exogenous variables were given by Trognon (1978). The direction of the asymptotic bias for a higher-order autoregressive process is difficult to identify a priori.

[3] For details, see Section 2.4 or Sevestre and Trognon (1982).

3.2.2 The Least Squares Dummy Variable (Covariance) Estimator

Treating α_i as fixed constants, the least squares estimator of γ and α_i is to minimize

$$S = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \gamma y_{i,t-1} - \alpha_i\right)^2. \tag{3.2.8}$$

Minimizing (3.2.8) with respect to γ and α_i can be considered as the maximum likelihood estimator under the assumption that u_{it} is independently, identically normally distributed and y_{i0} are fixed constants. Taking partial derivatives of S with respect to α_i yields

$$\hat{\alpha}_i = \bar{y}_i - \lambda \bar{y}_{i,-1}, \quad i = 1, \ldots, N, \tag{3.2.9}$$

where $\bar{y}_i = \sum_{t=1}^{T} y_{it}/T$ and $\bar{y}_{i,-1} = \sum_{t=1}^{T} y_{i,t-1}/T$. Substituting (3.2.9) into (3.2.8) and minimizing the concentrated S leads to the least squares dummy variable (Covariance) estimator (cv) of γ,

$$\begin{aligned}\hat{\gamma}_{\text{cv}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)(y_{i,t-1} - \bar{y}_{i,-1})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,-1})^2} \\ &= \gamma + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_i)/NT}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,-1})^2/NT}.\end{aligned} \tag{3.2.10}$$

The CV exists if the denominator of the second term of (3.2.10) is nonzero. It is consistent if the numerator of the second term of (3.2.10) converges to zero. Making use of (3.2.5) in summing $y_{i,t-1}$ over t, we have

$$\begin{aligned}\sum_{t=1}^{T} y_{i,t-1} &= \frac{1-\gamma^T}{1-\gamma} y_{i0} + \frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2}\alpha_i \\ &\quad + \frac{1-\gamma^{T-1}}{1-\gamma}u_{i1} + \frac{1-\gamma^{T-2}}{1-\gamma}u_{i2} + \ldots + u_{i,T-1}.\end{aligned} \tag{3.2.11}$$

Under the assumption that u_{it} is uncorrelated with α_i and are independently identically distributed, by a law of large numbers (Rao 1973), and using (3.2.11), we can show that when N tends to infinity,

$$\begin{aligned}&\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_i) \\ &\quad = -\operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i,-1}\bar{u}_i \\ &\quad = -\frac{\sigma_u^2}{T^2}\cdot\frac{T(1-\gamma) - (1-\gamma^T)}{(1-\gamma)^2}.\end{aligned} \tag{3.2.12}$$

By similar manipulations we can show that the denominator of (3.2.10) converges to

$$\frac{\sigma_u^2}{1-\gamma^2}\left\{1 - \frac{1}{T} - \frac{2\gamma}{(1-\gamma)^2}\cdot\frac{(T-1) - T\gamma + \gamma^T}{T^2}\right\}. \tag{3.2.13}$$

If T is fixed, (3.2.12) is a nonzero constant, and (3.2.9) and (3.2.10) are inconsistent estimators no matter how large N is. The asymptotic bias of the CV of γ is

$$\underset{N\to\infty}{\text{plim}}\,(\hat{\gamma}_{\text{cv}} - \gamma) = -\frac{1+\gamma}{T-1}\left(1 - \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\right) \left\{1 - \frac{2\gamma}{(1-\gamma)(T-1)}\left[1 - \frac{1-\gamma^T}{T(1-\gamma)}\right]\right\}^{-1}. \tag{3.2.14}$$

The bias of $\hat{\gamma}_{cv}$ is caused by having to eliminate the unknown individual effects α_i from each observation, which creates the correlation of the order $(1/T)$ between the regressors and the residuals in the transformed model $(y_{it} - \bar{y}_i) = \gamma(y_{i,t-1} - \bar{y}_{i,-1}) + (u_{it} - \bar{u}_i)$. For a small T, this bias is always negative if $\gamma > 0$. Nor does the bias go to zero as γ goes to zero. A typical panel usually contains a small number of time series observations, this bias can hardly be ignored. For instance, when $T = 2$, the asymptotic bias is equal to $-(1+\gamma)/2$, and when $T = 3$, it is equal to $-(2+\gamma)(1+\gamma)/2$. Even with $T = 10$ and $\gamma = 0.5$, the asymptotic bias is -0.167. The CV for the dynamic fixed-effects model remains biased with the introduction of exogenous variables if T is small; for details of the derivation, see Anderson and Hsiao (1982) and Nickell (1981); for Monte Carlo studies, see Nerlove (1971a).

The process of eliminating the individual-specific effects α_i introduces an estimation error of order T^{-1}. When T is large, $(y_{i,t-1} - \bar{y}_{i,-1})$ and $(u_{it} - \bar{u}_i)$ become asymptotically uncorrelated, (3.2.12) converges to zero, and (3.2.13) converges to a nonzero constant $\sigma_u^2/(1-\gamma^2)$. Hence, when $T \longrightarrow \infty$, the CV estimator becomes consistent. It can be shown that when N is fixed and T is large, $\sqrt{T}(\hat{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean 0 and variance $1-\gamma^2$. When both N and T are large, although the correlation between $(y_{i,t-1} - \bar{y}_i)(u_{it} - \bar{u}_i)$ is of order $\frac{1}{T}$, its magnitude becomes aggravated by N. Hahn and Kuersteiner (2002) show that $\sqrt{NT}(\hat{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean $-\sqrt{c}(1+\gamma)$ and variance $1-\gamma^2$, if $c = \lim \frac{N}{T} \neq 0$ as $T \to \infty$. Since the standard error of CV is $O\left(\frac{1}{\sqrt{NT}}\right)$, the conventionally constructed t-statistic

$$\frac{(\hat{\gamma}_{CV} - \gamma)}{\text{standard error of } \hat{\gamma}_{CV}} \tag{3.2.15}$$

is no longer centered at zero. The asymptotic bias of the CV estimator could lead to severe size distortion when N also increases as T increases such that $\frac{N}{T} \to c \neq 0$ as $T \to \infty$ (e.g., Hsiao and Zhang 2015).

3.3 METHOD OF MOMENTS ESTIMATOR

3.3.1 Simple Instrumental Variable Estimator

Taking the first difference of (3.2.1) to eliminate α_i, we have

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \boldsymbol{\beta}'(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}) + u_{it} - u_{i,t-1}, t = 2, \ldots, T. \tag{3.3.1}$$

However, $\Delta y_{i,t-1} = (y_{i,t-1} - y_{i,t-2})$ is correlated with $\Delta u_{it} = u_{it} - u_{i,t-1}$. Regressing $\Delta y_{it} = (y_{it} - y_{i,t-1})$ on $\Delta y_{i,t-1}$ and $\Delta \boldsymbol{x}_{it} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})$ is inconsistent.

Because $y_{i,t-2}$ and $(y_{i,t-2} - y_{i,t-3})$ are correlated with $(y_{i,t-1} - y_{i,t-2})$ but are uncorrelated with $(u_{it} - u_{i,t-1})$, they can be used as an instrument for $(y_{i,t-1} - y_{i,t-2})$. Estimating γ and $\boldsymbol{\beta}$ by the instrumental-variable method, both

$$\begin{pmatrix} \hat{\gamma}_{iv} \\ \hat{\boldsymbol{\beta}}_{iv} \end{pmatrix} = \left[\sum_{i=1}^{N} \sum_{t=3}^{T} \begin{pmatrix} (y_{i,t-1} - y_{i,t-2})(y_{i,t-2} - y_{i,t-3}) & (y_{i,t-2} - y_{it-3})(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})' \\ (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(y_{i,t-1} - y_{i,t-2}) & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})' \end{pmatrix} \right]^{-1}$$
$$\left[\sum_{i=1}^{N} \sum_{t=3}^{T} \begin{pmatrix} y_{i,t-2} - y_{i,t-3} \\ \boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1} \end{pmatrix} (y_{it} - y_{i,t-1}) \right], \tag{3.3.2}$$

and

$$\begin{pmatrix} \tilde{\gamma}_{iv} \\ \tilde{\boldsymbol{\beta}}_{iv} \end{pmatrix} = \left[\sum_{i=1}^{N} \sum_{t=2}^{T} \begin{pmatrix} y_{i,t-2}(y_{i,t-1} - y_{i,t-2}) & y_{i,t-2}(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})' \\ (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(y_{i,t-1} - y_{i,t-2}) & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})' \end{pmatrix} \right]^{-1}$$
$$\left[\sum_{i=1}^{N} \sum_{t=2}^{T} \begin{pmatrix} y_{i,t-2} \\ \boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1} \end{pmatrix} (y_{it} - y_{i,t-1}) \right], \tag{3.3.3}$$

are consistent.

Both (3.3.2) and (3.3.3) are derived using the sample moments $\frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=2}^{T} \boldsymbol{q}_{it}$ $(u_{it} - u_{i,t-1}) = 0$ to approximate the population moments $E[\boldsymbol{q}_{it}(u_{it} - u_{i,t-1})] = \mathbf{0}$, where $\boldsymbol{q}_{it} = [(y_{i,t-2} - y_{i,t-3}), (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})']'$ for (3.3.2) and $\boldsymbol{q}_{it} = [y_{i,t-2}, (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})']$ for (3.3.3). Therefore, (3.3.2) or (3.3.3) are consistent, and $\sqrt{NT}[(\hat{\gamma}_{iv} - \gamma), (\hat{\boldsymbol{\beta}}_{iv} - \boldsymbol{\beta})']'$ is asymptotically normally distributed with mean zero either N or T or both tend to infinity (in other words, there is no asymptotic bias).

Estimator (3.3.3) has an advantage over (3.3.2) in the sense that the minimum number of time periods required is two, whereas (3.3.2) requires $T \geq 3$. In practice, if $T \geq 3$, the choice between (3.3.2) and (3.3.3) depends on the correlations between $(y_{i,t-1} - y_{i,t-2})$ and $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$. For a comparison of asymptotic efficiencies of the instruments $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$, see Han and Phillips (2013).

When T is short, Chudik and Pesaran (2020) note that under the assumption that u_{it} is independently distributed over t,

$$E(\Delta u_{it}^2) + E(\Delta u_{it} \Delta y_{i,t+1}) + E(\Delta u_{it} \Delta y_{i,t-1}) = 0. \tag{3.3.4}$$

Substituting $\Delta u_{it} = \Delta y_{it} - \Delta y_{i,t-1}\gamma - \Delta \boldsymbol{x}_{it}' \boldsymbol{\beta}$ and equating sample moments to the population moments, one obtains the quadratic function

$$\frac{1}{N(T-1)} \sum_{i=1}^{N} \left\{ \sum_{t=2}^{T} \left[(\Delta y_{it} - \Delta y_{i,t-1}\,\gamma - \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta})^2 + (\Delta y_{it} - \Delta y_{i,t-1}\,\gamma - \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta}) \Delta y_{i,t-1} \right] \right.$$
$$\left. + \frac{T-1}{T-2} \sum_{t=2}^{T-1} \left(\Delta y_{it} - \Delta y_{i,t-1}\gamma - \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} \right) \Delta y_{i,t+1} \right\} \tag{3.3.5}$$

Their simulation studies show combining (3.3.5) together with the moment conditions $E(\Delta \boldsymbol{x}_{it} \Delta u_{i,t+1}) = \mathbf{0}$ and $E(\Delta y_{i,t-2} \Delta u_{it}) = 0$ or $E(y_{i,t-2} \Delta u_{it}) = 0$ optimally (through a GMM method discussed in Section 3.3.2) significantly improves the efficiency of the simple instrumental variable estimator (3.3.2) or (3.3.3) (in terms of the relative variance of their respective estimators).

The instrumental variable method is simple to implement. When the individual varying but time-invariant variables z_i appear as explanatory variables of y_{it},

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}' \boldsymbol{x}_{it} + \boldsymbol{\rho}' z_i + \alpha_i + u_{it}, \tag{3.3.6}$$

if α_i are fixed constants and correlated with z_i, there is no way to separate the impacts of z_i and α_i. If α_i are random and uncorrelated with z_i, then the coefficients of z_i can be obtained by the following two-step procedure.

Step 1: Taking the first difference of (3.3.4) yields (3.3.1). Apply the simple IV (3.3.2) or (3.3.3) or the Chudik and Pesaran augmented estimator to obtain $\boldsymbol{\beta}$ and γ.

Step 2: Substitute the estimated $\boldsymbol{\beta}$ and γ into the equation

$$\bar{y}_i - \gamma \bar{y}_{i,-1} - \boldsymbol{\beta}' \bar{\boldsymbol{x}}_i = \boldsymbol{\rho}' z_i + \alpha_i + \bar{u}_i \qquad i = 1, \ldots, N, \tag{3.3.7}$$

where $\bar{y}_i = \sum_{t=1}^{T} y_{it}/T$, $\bar{y}_{i,-1} = \sum_{t=1}^{T} y_{i,t-1}/T$, $\bar{\boldsymbol{x}}_i = \sum_{t=1}^{T} \boldsymbol{x}_{it}/T$, and $\bar{u}_i = \sum_{t=1}^{T} u_{it}/T$. Estimate $\boldsymbol{\rho}$ by the OLS method.

Under the assumption that $E(\alpha_i) = 0$, the σ_u^2 and σ_α^2 can be obtained by

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} \left[(y_{it} - y_{i,t-1}) - \hat{\gamma}(y_{i,t-1} - y_{i,t-2}) - \hat{\boldsymbol{\beta}}'(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}) \right]^2}{N(T-1)}, \tag{3.3.8}$$

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{N} \left(\bar{y}_i - \hat{\gamma} \bar{y}_{i,-1} - \hat{\boldsymbol{\rho}}' z_i - \hat{\boldsymbol{\beta}}' \bar{\boldsymbol{x}}_i \right)^2}{N} - \frac{1}{T} \hat{\sigma}_u^2. \tag{3.3.9}$$

The consistency of these estimators is independent of initial conditions. The instrumental-variable estimators of γ, $\boldsymbol{\beta}$, and σ_u^2 are consistent when N or T or both tend to infinity. The estimators of $\boldsymbol{\rho}$ and σ_α^2 are consistent only when N goes to infinity. They are inconsistent if N is fixed and T tends to infinity.

3.3.2 Generalized Method of Moments Estimator (GMM)

We note that $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$ is not the only instrument for $(y_{i,t-1} - y_{i,t-2})$. In fact, as noted by Amemiya and MaCurdy (1986), Arellano and Bond (1991), and Breusch, Mizon, and Schmidt (1989), all $y_{i,t-2-j}, j = 0, 1, \ldots$ satisfy the conditions that $E[y_{i,t-2-j}(y_{i,t-1} - y_{i,t-2})] \neq 0$ and $E[y_{i,t-2-j}(u_{it} - u_{i,t-1})] = 0$. Therefore, they all are legitimate instruments for $(y_{i,t-1} - y_{i,t-2})$. Let $\boldsymbol{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2}, \boldsymbol{x}_i')'$, where $\boldsymbol{x}_i' = (\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{iT}')$, we have

$$E\boldsymbol{q}_{it} \Delta u_{it} = 0, \quad t = 2, \ldots, T. \tag{3.3.10}$$

Stacking the $(T-1)$ first differenced equation of (3.3.1) in matrix form, we have

$$\Delta \boldsymbol{y}_i = \Delta \boldsymbol{y}_{i,-1} \gamma + \Delta X_i \boldsymbol{\beta} + \Delta \boldsymbol{u}_i, \quad i = 1, \ldots, N, \tag{3.3.11}$$

where $\Delta \boldsymbol{y}_i$, $\Delta \boldsymbol{y}_{i,-1}$, and $\Delta \boldsymbol{u}_i$ are $(T-1) \times 1$ vectors of the form $(y_{i2} - y_{i1}, \ldots, y_{iT} - y_{i,T-1})'$, $(y_{i1} - y_{i0}, \ldots, y_{i,T-1} - y_{i,T-2})'$, $(u_{i2} - u_{i1}, \ldots, u_{iT} - u_{i,T-1})'$, respectively, and ΔX_i is the $(T-1) \times K_1$ matrix of $(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT} - \boldsymbol{x}_{i,T-1})'$. The $T(T-1)[K_1 + \frac{1}{2}]$ orthogonality (or moment) conditions of (3.3.10) can be represented as

$$EW_i \Delta \boldsymbol{u}_i = \boldsymbol{0}, \tag{3.3.12}$$

where

$$W_i = \begin{pmatrix} \boldsymbol{q}_{i2} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \boldsymbol{q}_{i3} & \cdots & \mathbf{0} \\ \vdots & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{q}_{i,T} \end{pmatrix}, \tag{3.3.13}$$

is of dimension $[T(T-1)(K_1+\frac{1}{2})] \times (T-1)$. The dimension of (3.3.13) in general is much larger than $K_1 + 1$. Thus, Arellano and Bond (1991) suggest a generalized method of moments estimator (GMM).

The standard method of moments estimator consists of solving the unknown parameter vector $\boldsymbol{\theta}$ by equating the theoretical moments with their empirical counterparts or estimates. For instance, suppose that $\boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})$ denote some population moments of $\boldsymbol{y}$ and/or $\boldsymbol{x}$, say the first and second moments of $\boldsymbol{y}$ and/or $\boldsymbol{x}$, which are functions of the unknown parameter vector $\boldsymbol{\theta}$ and are supposed to equal to some known constants, say zero. Let $\hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{x}_i;\boldsymbol{\theta})$ be their sample estimates based on N independent samples of $(\boldsymbol{y}_i,\boldsymbol{x}_i)$. Then the method of moments estimator $\boldsymbol{\theta}$ is the $\hat{\boldsymbol{\theta}}_{mm}$, such that

$$\boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) = \hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\hat{\boldsymbol{\theta}}_{mm}) = \mathbf{0}. \tag{3.3.14}$$

For instance, the orthogonality conditions between QX_i and $Q\boldsymbol{u}_i$ for the fixed-effects linear static model (2.2.2), $E(X_i'Q\boldsymbol{u}_i) = E[X_i'Q(\boldsymbol{y}_i - \boldsymbol{e}\alpha_i^* - X_i\boldsymbol{\beta})] = \mathbf{0}$, lead to the LSDV estimator (2.2.10). In this sense, the IV method is a method of moments estimator.

If the number of equations in (3.3.12) is equal to the dimension of $\boldsymbol{\theta}$, it is in general possible to solve for $\hat{\boldsymbol{\theta}}_{mm}$ uniquely. If the number of equations is greater than the dimension of $\boldsymbol{\theta}$, (3.3.14) in general has no solution. It is then necessary to minimize some norm (or distance measure) of $\hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\hat{\boldsymbol{\theta}}) - \boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})$, say,

$$[\hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\hat{\boldsymbol{\theta}}) - \boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})]'A[\hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\hat{\boldsymbol{\theta}}) - \boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})], \tag{3.3.15}$$

where A is some positive definite matrix.

The statistical property of the moment estimator thus obtained depends on A. The optimal choice of A turns out to be

$$A^* = \left\{E[\hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) - \boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})][\hat{\boldsymbol{m}}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta}) - \boldsymbol{m}(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})]'\right\}^{-1} \tag{3.3.16}$$

(Hansen 1982). The generalized method of moments estimator of $\boldsymbol{\theta}$ is to choose $\hat{\boldsymbol{\theta}}_{GMM}$ such that it minimizes (3.3.15) when $A = A^*$.

The Arellano and Bond (1991) GMM estimator of $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}')'$ is obtained by minimizing

$$\left(\frac{1}{N}\sum_{i=1}^{N} \Delta\boldsymbol{u}_i' W_i'\right) \Psi^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \Delta\boldsymbol{u}_i\right), \tag{3.3.17}$$

with respect to $\boldsymbol{\theta}$, where $\Psi = E\left[\frac{1}{N^2}\sum_{i=1}^{N} W_i \Delta\boldsymbol{u}_i \Delta\boldsymbol{u}_i' W_i'\right]$. Under the assumption that u_{it} is i.i.d. with mean zero and variance σ_u^2, Ψ can be approximated by $\frac{\sigma_u^2}{N^2}\sum_{i=1}^{N} W_i \tilde{A} W_i'$, where

$$\underset{(T-1)\times(T-1)}{\tilde{A}} = \begin{bmatrix} 2 & -1 & 0 & . & 0 \\ -1 & 2 & -1 & . & 0 \\ 0 & \ddots & \ddots & & \\ 0 & \ddots & \ddots & . & -1 \\ 0 & & . & -1 & 2 \end{bmatrix}. \tag{3.3.18}$$

Thus, the Arellano and Bond GMM estimator takes the form

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{GMM,AB} = & \left\{ \left[\sum_{i=1}^{N} \begin{pmatrix} \Delta \boldsymbol{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} W'_i \right] \left[\sum_{i=1}^{N} W_i \tilde{A} W'_i \right]^{-1} \left[\sum_{i=1}^{N} W_i (\Delta \boldsymbol{y}_{i,-1}, \Delta X_i) \right] \right\}^{-1} \\ & \left\{ \left[\sum_{i=1}^{N} \begin{pmatrix} \Delta \boldsymbol{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} W'_i \right] \left[\sum_{i=1}^{N} W_i \tilde{A} W'_i \right]^{-1} \left[\sum_{i=1}^{N} W_i \Delta \boldsymbol{y}_i \right] \right\}, \end{aligned} \tag{3.3.19}$$

with asymptotic covariance matrix

$$\operatorname{Cov}\left(\hat{\boldsymbol{\theta}}_{GMM,AB}\right) = \left\{ \left[\sum_{i=1}^{N} \begin{pmatrix} \Delta \boldsymbol{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} W'_i \right] \left[\sum_{i=1}^{N} W_i \tilde{A} W'_i \right]^{-1} \left[\sum_{i=1}^{N} W_i (\Delta \boldsymbol{y}_{i,-1}, \Delta X_i) \right] \right\}^{-1}. \tag{3.3.20}$$

In addition to the moment conditions (3.3.10), Arellano and Bover (1995) also note that $E\bar{v}_i = 0$, where $\bar{v}_i = \bar{y}_i - \bar{y}_{i,-1}\gamma - \bar{\boldsymbol{x}}'_i\boldsymbol{\beta} - \boldsymbol{\rho}' \boldsymbol{z}_i$. Therefore, if instruments $\tilde{\boldsymbol{q}}_i$ exist (for instance, the constant 1 is a valid instrument) such that

$$E\tilde{\boldsymbol{q}}_i \bar{v}_i = \boldsymbol{0}, \tag{3.3.21}$$

then a more efficient GMM estimator can be derived by incorporating this additional moment condition.

Apart from the linear moment conditions (3.3.12) and (3.3.21), Ahn and Schmidt (1995) note that the homoscedasticity condition of $E(v_{it}^2)$ implies the following $T-2$ linear conditions

$$E(y_{it}\Delta u_{i,t+1} - y_{i,t+1}\Delta u_{i,t+2}) = 0, \quad t = 1, \ldots, T-2. \tag{3.3.22}$$

Combining (3.3.12), (3.3.21), and (3.3.22), a more efficient GMM estimator can be derived by minimizing

$$\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^{+\prime} W_i^{+\prime} \right) \Psi^{+-1} \left(\frac{1}{N} \sum_{i=1}^{N} W_i^{+} \boldsymbol{u}_i^{+} \right) \tag{3.3.23}$$

with respect to $\boldsymbol{\theta}$, where $\boldsymbol{u}_i^+ = (\Delta \boldsymbol{u}'_i, \bar{v}_i)'$, $\Psi^+ = E\left(\frac{1}{N^2} \sum_{i=1}^{N} W_i^+ \boldsymbol{u}_i^+ \boldsymbol{u}_i^{+\prime} W_i^{+\prime} \right)$, and

$$W_i^{+'} = \begin{pmatrix} W'_i & W_i^{*\prime} & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{0}' & \tilde{\boldsymbol{q}}'_i \end{pmatrix}$$

where

$$\underset{(T-2)\times(T-1)}{W_i^*} = \begin{pmatrix} y_{i1} & -y_{i2} & 0 & 0 & \dots & 0 \\ 0 & y_{i2} & -y_{i3} & 0 & \dots & . \\ & & & & & \\ & & & & & 0 \\ & & & 0 & y_{i,T-2} & -y_{i,T-1} \end{pmatrix}.$$

However, because the covariance matrix (3.3.22) depends on the unknown $\boldsymbol{\theta}$, it is impractical to implement the GMM. A less efficient but computationally feasible GMM estimator is to ignore the information that Ψ^+ also depends on $\boldsymbol{\theta}$ and simply substitute Ψ by its consistent estimator

$$\hat{\Psi}^+ = \left(\frac{1}{N^2} \sum_{i=1}^{N} W_i^+ \hat{\boldsymbol{u}}_i^+ \hat{\boldsymbol{u}}_i^{+\prime} W_i^{+\prime} \right) \qquad (3.3.24)$$

into the objective function (3.3.23) to derive a linear estimator of form (3.3.19) where $\hat{\boldsymbol{u}}_i^+$ is derived by using some simple consistent estimator of γ and $\boldsymbol{\beta}$, say, the IV discussed in Section 3.3.1, into (3.3.23) and the $\bar{v}_i$ equation.

In principle, one can improve the asymptotic efficiency of the GMM type estimator by adding more moment conditions. For instance, Ahn and Schmidt (1995) note that in addition to the linear moment conditions of (3.3.12), (3.3.21), and (3.3.22), there exist $(T-1)$ nonlinear moment conditions of the form $E((\bar{y}_i - \boldsymbol{\beta}'\bar{\boldsymbol{x}}_i)\Delta u_{it}) = 0, t = 2, \dots, T$, implied by the homoscedasticity conditions of Ev_{it}^2. Under the additional assumption that $E(\alpha_i y_{it})$ is the same for all t, this condition and condition (3.3.22) can be transformed into the $(2T-2)$ linear moment conditions

$$E[u_{iT}\Delta y_{it}] = 0,\ t = 1, \dots, T-1, \qquad (3.3.25)$$

and

$$E[u_{it}y_{it} - u_{i,t-1}y_{i,t-1}] = 0,\ t = 2, \dots, T. \qquad (3.3.26)$$

While theoretically, it is possible to add additional moment conditions to improve the asymptotic efficiency of GMM, it is doubtful how much efficiency gain one can achieve by using a huge number of moment conditions in a finite sample. Moreover, if higher moment conditions are used, the estimator can be very sensitive to outlying observations. Through a simulation study, Ziliak (1997) has found that the downward bias in GMM is quite severe as the number of moment conditions expands, outweighing the gains in efficiency. The strategy of exploiting all the moment conditions for estimation is actually not recommended for panel-data applications in a finite sample, mainly due to bias. The bias is proportional to the number of each instruments for each equation. In addition, when γ is close to one, the lagged instruments $y_{i,t-2-j}, j \geq 0$ are weak instruments. There is also a bias-variance trade-off in the number of moment conditions used for estimation. Koenker and Machado (1999) show that the usual asymptotic theory holds only if the number of moments used is less than the cubic root of sample size. Okui (2009) proposes a moment selection method based on minimizing (Nagar 1959) approximate mean square error. In general, when T is small, it is optimal to use all moment conditions. When T is not very small $\left(\frac{T^2}{N \log T} \longrightarrow \infty\right)$, the optimal number of moment conditions chosen is of $O((NT)^{1/3})$ assuming there exists a natural rank ordering of instruments for each

$(y_{i,t-1} - y_{i,t-2})$, say $(y_{i0}, y_{i1}, \ldots, y_{i,t-2})$ in increasing order.[4] When σ_α^2 is large relative to σ_u^2, it is also advisable to use many moment conditions. For further discussions, see Judson and Owen (1999), Kiviet (1995), and Wansbeek and Bekker (1996).

To improve the efficiency of GMM when γ is close to 1, Hahn, Hausman, and Kuersteiner (2007) suggest not to use the first difference equation as in (3.3.1), but to use the long difference $y_{iT} - y_{i1}$. In the case of the first-order autoregressive process (3.2.3),

$$y_{iT} - y_{i1} = \gamma(y_{i,T-1} - y_{io}) + (u_{iT} - u_{i1}). \tag{3.3.27}$$

Then $y_{io}, y_{i,T-1} - \gamma y_{i,T-2}, \ldots, y_{i2} - \gamma y_{i1}$ are valid instruments. Their long difference (LD) estimator is equivalent to applying the GMM based on the "reduced set" of moment conditions

$$E\begin{pmatrix} y_{io} \\ y_{iT-1} - \gamma y_{i,T-2} \\ \vdots \\ y_{i2} - \gamma y_{i1} \end{pmatrix} [(y_{iT} - y_{i1}) - \gamma(y_{i,T-1} - y_{io})] = \mathbf{0}. \tag{3.3.28}$$

The instruments $y_{it} - \gamma y_{i,t-1}$ for $t = 2, \ldots, T-1$ require the knowledge of γ. A feasible LD estimator could be to use the Arellano–Bond GMM estimator (3.3.19) to obtain a preliminary consistent estimator $\hat{\gamma}_{GMM,AB}$, then to use $(y_{io}, y_{i,T-1} - y_{i,T-2}\hat{\gamma}_{GMM,AB}, \ldots, y_{i2} - y_{i1}\hat{\gamma}_{GMM,AB})$ as instruments.

The reason that the LD estimator can improve the efficiency of the GMM based on the first difference equation of (3.3.1) is because GMM can be viewed as the two-stage least squares method (Theil 1958). As shown by Donald and Newey (2001), the bias of 2SLS (GMM) depends on four factors: "explained" variance of the first stage reduced form equation; "covariance" between the stochastic disturbance of the structural equation and the reduced form equation; the number of instruments; and the sample size,

$$E\left[\hat{\gamma}_{2SLS} - \gamma\right] \simeq \frac{1}{n}a, \tag{3.3.29}$$

where n denotes the sample size and

$$a = \frac{(\text{number of instruments}) \times (\text{“covariance”})}{\text{“Explained” variance of the first stage reduced form equation}}. \tag{3.3.30}$$

Based on this formula, Hahn, Hausman, and Kuersteiner (2007) show that the $a = -\frac{1+\gamma}{1-\gamma}$ for the Arellano–Bond (1991) GMM estimator when $T = 3$. When $\gamma = .9$, $a = -19$. For $N = 100$, this implies a percentage bias of -105.56. On the other hand, using the LD estimator, $a = -.37$, which is much smaller than -19 in absolute magnitude.

Remark 3.3.1 We derive the simple IV (3.3.3) or the GMM estimator (3.3.19) assuming that u_{it} is independently distributed across i and over t. If u_{it} is serially correlated, $E(y_{i,t-j}\Delta u_{it}) \neq 0$ for $j \geq 2$. Then neither (3.3.3) nor (3.3.19) is a consistent estimator. On the other hand, the estimator $\hat{\boldsymbol{\theta}}^*$ that replaces W_i in (3.3.19) by the block diagonal instrument matrix W_i^* whose tth block is given by $\boldsymbol{x}_i$ if $\boldsymbol{x}_{it}$ is strictly exogenous (i.e., $E\boldsymbol{x}_{it}u_{is} = 0$ for all s) or $(\boldsymbol{x}'_{it}, \boldsymbol{x}'_{i,t-1}, \ldots, \boldsymbol{x}'_{i1})'$ if $\boldsymbol{x}_{it}$ is weakly exogenous (i.e.,

[4] Actually, Okui (2009) derives his selection method using the forward orthogonal deviation operation of Arellano and Bover (1995): $\Delta u_{it}^* = \sqrt{\frac{T-t}{T-T+1}}[u_{it} - \frac{1}{T-t}(u_{i,t+1} + \cdots + u_{iT})]$.

$E(\boldsymbol{x}_{i,t+j+1}u_{it}) \neq 0$ and $E(u_{it}\boldsymbol{x}_{i,t-j}) = 0$ for $j \geq 0$) remains consistent. Therefore, a Hausman (1978) type of test statistic can be constructed to test if u_{it} is serially uncorrelated by comparing the difference of $(\hat{\boldsymbol{\theta}}_{GMM,AB} - \hat{\boldsymbol{\theta}}^*)$.

Arellano and Bond (1991) note that if u_{it} is not serially correlated, then $E(\Delta u_{it}\Delta u_{i,t-2}) = 0$. They show that the statistic

$$\frac{\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}}{\hat{s}} \tag{3.3.31}$$

is asymptotically normally distributed with mean 0 and variance 1 when $T \geq 5$ and $N \longrightarrow \infty$, where

$$\begin{aligned}
\hat{s}^2 = {} & \sum_{i=1}^{N}\left(\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}\right)^2 - 2\left(\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{i,t-2}\Delta\boldsymbol{x}'_{it}\right) \\
& \times\left\{\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\boldsymbol{y}'_{i,-1}\\ \Delta X'_i\end{pmatrix}W'_i\right]\left(\frac{1}{N}\sum_{i=1}^{N}W_i\tilde{A}W'_i\right)^{-1}\left[\sum_{i=1}^{N}W_i(\Delta\boldsymbol{y}_{i,-1},\Delta X_i)\right]\right\}^{-1} \\
& \times\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\boldsymbol{y}'_{i,-1}\\ \Delta X'_i\end{pmatrix}W'_i\right]\left(\frac{1}{N}\sum_{i=1}^{N}W_i\tilde{A}W'_i\right)^{-1}\left[\sum_{i=1}^{N}W_i\Delta\hat{\boldsymbol{u}}_i\left(\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}\right)\right] \\
& + \left(\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{i,t-2}\Delta\boldsymbol{x}'_{it}\right)\left(\text{Cov}\,(\hat{\boldsymbol{\theta}}_{GMM,AB})\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta\boldsymbol{x}_{it}\Delta\hat{u}_{i-t-2}\right),
\end{aligned} \tag{3.3.32}$$

where $\Delta\hat{u}_{it} = \Delta y_{it} - (\Delta y_{i,t-1}, \Delta\boldsymbol{x}'_{it})\hat{\boldsymbol{\theta}}_{GMM,AB}, \Delta\hat{\boldsymbol{u}}_i = (\Delta\hat{u}_{i2}, \ldots, \Delta\hat{u}_{iT})$.

This statistic or the Hausman-type test statistic can be used to test serial correlation. The statistic (3.3.31) is only defined if $T \geq 5$. When $T < 5$, Arellano and Bond (1991) suggest using the Sargan (1958) test statistic of overidentification,

$$\left(\sum_{i=1}^{N}\Delta\hat{\boldsymbol{u}}'_i W^{*'}_i\right)\left(\sum_{i=1}^{N}W^*_i\Delta\hat{\boldsymbol{u}}_i\Delta\hat{\boldsymbol{u}}'_iW^{*'}_i\right)^{-1}\left(\sum_{i=1}^{N}W^*_i\Delta\hat{\boldsymbol{u}}_i\right), \tag{3.3.33}$$

where W^*_i could be W_i or any number of instruments that satisfy the orthogonality condition $E(W^*_i\Delta\boldsymbol{u}_i) = 0$. Under the null of no serial correlation, (3.3.33) is asymptotically chi-square distributed with $p - (K+1)$ degrees of freedom for any $p > (K+1)$ where p denotes the number of rows in W^*_i.

Remark 3.3.2 Because the individual-specific effects α_i are time-invariant, taking the deviation of the individual y_{it} equation from any transformation of y_{it} equation that maintains the time-invariance property of α_i can eliminate α_i. For instance, Alvarez and Arellano (2003) consider the forward deviation transformation of y_{it} equation into an equation of the form,

$$c_t\left[y_{it} - \frac{1}{(T-t)}(y_{i,t+1} + \cdots + y_{iT})\right],\ t = 1, \ldots, T-1, \tag{3.3.34}$$

where $c_t^2 = \frac{(T-t)}{(T-t+1)}$. The advantage of considering the equation specified by the transformation (3.3.34) is that the residuals $u^*_{it} = c_t[u_t - \frac{1}{(T-t)}(u_{i,t+1} + \cdots + u_{iT})]$, $t = 1, \ldots, T-1$ is orthogonal, that is, $Eu^*_{it}u^*_{is} = 0$ if $t \neq s$ and $Eu^{*2}_{it} = \sigma_u^2$. However, if transformation (3.3.33) is used to remove α_i, then u^*_{it} are uncorrelated with $y_{i,t-1}$. Thus the instruments $\boldsymbol{q}_{it}$ now take the form $(y_{i0}, \ldots, y_{i,t-1}, \boldsymbol{x}'_i)$ in the application of GMM (for further discussion on the pros and cons of the first difference and forward deviation, see Hsiao and Zhou 2018).

3.4 THE QUASI-LIKELIHOOD APPROACH FOR RANDOM-EFFECTS MODELS

The method of moments approach does not need to consider the issue of initial value distribution, y_{i0}, because the orthogonality conditions exploited are independent of how y_{i0} is generated. On the other hand, the quasi-likelihood approach needs to take account of how y_{i0} is generated because for each i, we have $T+1$ observations of $(y_{i0}, y_{i1}, \ldots, y_{iT})$, while the assumed data generating process (DGP) of y_{it}, (3.2.1), is well defined only for $t = 1, \ldots, T$ not for y_{i0}. It turns out that the consistency and asymptotic properties of the quasi-maximum likelihood estimator depends on how y_{i0} is treated and how N and T go to infinity. Since formulating the DGP depends on the assumption of the DGP of α_i, we shall follow the convention for treating α_i as random if α_i are uncorrelated with $\boldsymbol{x}_{it}$ and α_i as fixed if α_i are correlated with $\boldsymbol{x}_{it}$.

3.4.1 The Likelihood Approach

When the specific effects are treated as random, they can be considered to be either correlated or not correlated with the explanatory variables. In the case in which the effects are correlated with the explanatory variables, ignoring this correlation no longer yields the desirable properties. Thus, a more appealing approach here would be to take explicit account of the linear dependence between the effects and the exogenous variables by letting $\alpha_i = \boldsymbol{a}'\bar{\boldsymbol{x}}_i + \omega_i$ (Mundlak 1978a) (see Section 2.4) and use a random-effects framework of the model

$$\boldsymbol{y}_i = \boldsymbol{y}_{i,-1}\gamma + X_i\boldsymbol{\beta} + \boldsymbol{\rho}'\boldsymbol{z}_i + \boldsymbol{e}\bar{\boldsymbol{x}}'_i\boldsymbol{a} + \boldsymbol{e}\omega_i + \boldsymbol{u}_i, \tag{3.4.1}$$

where now $E(\boldsymbol{x}_{it}\omega_i) = \boldsymbol{0}$ and $E(\boldsymbol{x}_{it}u_{it}) = \boldsymbol{0}$. However, because $\bar{\boldsymbol{x}}_i$ is time-invariant and the (residual) individual effect ω_i possesses the same property as α_i under the assumption $E\alpha_i\boldsymbol{x}'_{it} = \boldsymbol{0}'$, the estimation of (3.4.1) is formally equivalent to the estimation of the model

$$\boldsymbol{y}_i = \boldsymbol{y}_{i,-1}\gamma + X_i\boldsymbol{\beta} + \boldsymbol{e}\boldsymbol{z}'_i\boldsymbol{\rho} + \boldsymbol{e}\alpha_i + \boldsymbol{u}_i,, \quad i = 1, \ldots, N. \tag{3.4.2}$$

with X_i now denoting the $T \times K_1$ time-varying explanatory variables, $\boldsymbol{z}'_i$ being the $1 \times K_2$ time-invariant explanatory variables including the intercept term and $\bar{\boldsymbol{x}}_i$, and $E\alpha_i = 0$, $E\alpha_i\boldsymbol{z}'_i = \boldsymbol{0}'$, and $E\alpha_i\boldsymbol{x}'_{it} = \boldsymbol{0}'$. So, for ease of exposition, we assume in this section that the effects are uncorrelated with $\boldsymbol{x}_{it}, \boldsymbol{z}_i$. However, unlike the static case, the effects and the regressors remain correlated because of the presence of lagged dependent variables.

We assume y_{i0} are observable, $|\ \gamma\ | < 1$, $v_{it} = \alpha_i + u_{it}$,

$$\begin{aligned}
E\alpha_i &= Eu_{it} = 0, \\
E\alpha_i\boldsymbol{z}'_i &= \boldsymbol{0}', \qquad E\alpha_i\boldsymbol{x}'_{it} = \boldsymbol{0}', \\
E\alpha_i u_{jt} &= 0, \\
E\alpha_i\alpha_j &= \begin{cases} \sigma^2_\alpha & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \\
Eu_{it}u_{js} &= \begin{cases} \sigma^2_u & \text{if } i = j, \quad t = s, \\ 0 & \text{otherwise,} \end{cases}
\end{aligned} \tag{3.4.3}$$

3.4.1.1 Initial Observations Are Exogeneous

The likelihood function $f(y_{i0}, y_{i1}, \ldots, y_{iT}) = f(y_{i1}, \ldots, y_{iT} | y_{i0}) f(y_{i0})$ depends on the relationship between y_{i0} and y_{it}. If the DGP for y_{i0} is independent of y_{it}, say y_{i0} is a fixed constant, then $\Pi_{i=1}^{N} f(y_{i1}, \ldots, y_{iT} | y_{i0})$ takes the form

$$L = (2\pi)^{-\frac{NT}{2}} |V|^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{v}_i' V^{-1} \boldsymbol{v}_i\right), \tag{3.4.4}$$

under the normality assumption for α_i and u_{it}, where $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iT})' = (\alpha_i + u_{i1}, \ldots, \alpha_i + u_{iT})$,

$$V = \sigma_u^2 I_T + \sigma_\alpha^2 \boldsymbol{e}\boldsymbol{e}'. \tag{3.4.5}$$

When α_i and u_{it} are not normally distributed, (3.4.4) is their quasi-likelihood function. Conditional on V, the maximum likelihood estimator (MLE) or the quasi-maximum likelihood estimator (QMLE) is the generalized least squares estimator (GLS),

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \tilde{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \boldsymbol{y}_i'\right) \tag{3.4.6}$$

where $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}', \rho')'$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\tilde{X}_i = (\boldsymbol{y}_{i,-1}, X_i, \boldsymbol{e}\boldsymbol{z}_i')'$, $\boldsymbol{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$ The GLS is consistent and asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\left(\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \tilde{X}_i\right)^{-1}.$$

3.4.1.2 Initial Values Are Endogenous

However, the starting date of collecting panel data is arbitrary and if the DGP for y_{i0} is not different from any other y_{it}, then

$$\begin{aligned} y_{i0} &= \gamma y_{i,-1} + \boldsymbol{x}_{i0}'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\rho} + \alpha_i + u_{i0} \\ &= \boldsymbol{\beta}' \sum_{j=0} \boldsymbol{x}_{i,-j}\gamma^j + \frac{1}{1-\gamma}\boldsymbol{z}_i'\boldsymbol{\rho} + \frac{1}{1-\gamma}\alpha_i + \sum_{j=0} u_{i,-j}\gamma^j \\ &= \theta_{i0} + \epsilon_{i0} \end{aligned} \tag{3.4.7}$$

where

$$\theta_{i0} = \boldsymbol{\beta}' \sum_{j=0} \boldsymbol{x}_{i,-j}\gamma^j + \frac{1}{1-\gamma}\boldsymbol{z}_i'\boldsymbol{\rho} \quad \text{and} \quad \epsilon_{i0} = \frac{1}{1-\gamma}\alpha_i + \sum_{j=0} u_{i,-j}\gamma^j$$

Then the covariance between y_{i0} and y_{it} is not zero, but

$$\operatorname{Cov}(y_{i0}, y_{it}) = \frac{1}{1-\gamma}\sigma_\alpha^2 \tag{3.4.8}$$

In other words, $\Pi_{i=1}^{N} f(\boldsymbol{y}_i | y_{i0})$ is not (3.4.4). Under the assumption that $(\boldsymbol{x}_{it}, \boldsymbol{z}_i)$ are strictly exogenous with respect to α_i and u_{it},

$$\operatorname{plim} \frac{1}{NT} \sum_{i=1}^{N} X_i{}' V^{-1} \boldsymbol{v}_i = \boldsymbol{0} \quad \text{and} \quad \operatorname{plim} \frac{1}{NT} \sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{e}' V^{-1} \boldsymbol{v}_i = \boldsymbol{0}$$

but

$$\begin{aligned}
&\operatorname{plim} \frac{1}{NT} \sum_{i=1}^{N} \boldsymbol{y}_{i,-1}' \left[\mathrm{I}_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + \mathrm{T}\sigma_\alpha^2} \boldsymbol{e}\boldsymbol{e}' \right] (\boldsymbol{e}\alpha_i + \boldsymbol{u}_i) \\
&= \frac{\sigma_\alpha^2}{\sigma_u^2 + \mathrm{T}\sigma_a^2} \left(\operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i \alpha_i \right) - \frac{\mathrm{T}\sigma_\alpha^2}{\sigma_u^2 + \mathrm{T}\sigma_\alpha^2} \left(\operatorname{plim} \frac{1}{N} \sum_{i=1}^{N} \bar{y}_{i,-1} \bar{u}_i \right) \\
&= -\frac{\sigma_\alpha^2 \sigma_u^2 (1-\gamma^T)}{(1-\gamma)^2 (\sigma_u^2 + \mathrm{T}\sigma_\alpha^2)\mathrm{T}^2} = O\left(\frac{1}{T^2}\right)
\end{aligned} \tag{3.4.9}$$

The second equality of (3.4.9) follows from (3.2.7) and (3.2.12). The estimator (3.4.6) is inconsistent if T is small no matter how large N is. When T is large, (3.4.6) is consistent, but if N is also large and $N/T^3 \to a \neq 0$ as $T \to \infty$, $\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is not centered at zero, but biased of order $\sqrt{a}$. When an estimator is asymptotically biased, the conventional t- or F-test statistics could have significant size distortion (e.g., Hsiao and Zhang 2015, Hsiao and Zhou 2019).

Combining (3.4.2) and (3.4.7), the joint likelihood function of $\Pi_{i=1}^{N} f(y_{i0}, y_{i1}, \ldots, y_{iT})$ takes the form

$$\begin{aligned}
\tilde{L} = (2\pi)^{-\frac{N(T+1)}{2}} \mid \Omega \mid^{-\frac{N}{2}} \exp \Bigg\{ -\frac{1}{2} \sum_{i=1}^{N} (&y_{i0} - \theta_{i0}, y_{i1} - \gamma y_{i0} - \boldsymbol{\rho}' \boldsymbol{z}_i - \boldsymbol{\beta}' \boldsymbol{x}_{i1}, \ldots, \\
&y_{iT} - \gamma y_{i,T-1} - \boldsymbol{\rho}' \boldsymbol{z}_i - \boldsymbol{\beta}' \boldsymbol{x}_{iT}) \Omega^{-1} (y_{i0} - \theta_{i0}, \ldots, y_{iT} - \gamma y_{i,T-1} - \boldsymbol{\rho}' \boldsymbol{z}_i - \boldsymbol{\beta}' \boldsymbol{x}_{iT})' \Bigg\},
\end{aligned} \tag{3.4.10}$$

where

$$\begin{aligned}
\underset{(T+1)\times(T+1)}{\Omega} &= \sigma_u^2 \begin{bmatrix} \frac{1}{1-\gamma^2} & \boldsymbol{0}', \\ \boldsymbol{0} & I_T \end{bmatrix} + \sigma_\alpha^2 \begin{bmatrix} \frac{1}{1-\gamma} \\ \boldsymbol{e} \end{bmatrix} \left(\frac{1}{1-\gamma}, \boldsymbol{e}' \right), \\
\mid \Omega \mid &= \frac{\sigma_u^{2T}}{1-\gamma^2} \left(\sigma_u^2 + T\sigma_\alpha^2 + \frac{1+\gamma}{1-\gamma} \sigma_\alpha^2 \right), \\
\Omega^{-1} &= \frac{1}{\sigma_u^2} \Bigg[\begin{bmatrix} 1-\gamma^2 & \boldsymbol{0}' \\ \boldsymbol{0} & I_T \end{bmatrix} \\
&\quad - \left(\frac{\sigma_u^2}{\sigma_\alpha^2} + T + \frac{1+\gamma}{1-\gamma} \right)^{-1} \begin{bmatrix} 1+\gamma \\ \boldsymbol{e} \end{bmatrix} (1+\gamma, \boldsymbol{e}') \Bigg].
\end{aligned} \tag{3.4.11}$$

The initial value, y_{i0}, have means θ_{i0}, which are different for different i due to differences in $\boldsymbol{x}_{i,-j}$, thus introducing incidental parameter problems. The MLE in the presence of incidental parameters is inconsistent. Bhargava and Sargan (1983) suggest predicing θ_{i0} by all the observed $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_i$ as a means to get around the incidental-parameters problem, $\theta_{i0} = E(\theta_{i0} | \boldsymbol{x}_i', \boldsymbol{z}_i) + \epsilon_{i0}$. If $\boldsymbol{x}_{it}$ is generated by a homogeneous stochastic process

$$\boldsymbol{x}_{it} = \boldsymbol{c} + \sum_{j=0}^{\infty} \boldsymbol{b}_j \boldsymbol{\xi}_{i,t-j}, \tag{3.4.12}$$

where $\boldsymbol{\xi}_{it}$ is independently, identically distributed, then the minimum mean square error predictor of $\boldsymbol{x}_{i,-j}$ by $\boldsymbol{x}_{it}$ is the same for all i. Substituting these predictive formulae into (3.4.7) yields

$$y_{i0} = \sum_{t=1}^{T} \boldsymbol{\pi}'_{0t} \boldsymbol{x}_{it} + \boldsymbol{\rho}^{*\prime} \boldsymbol{z}_i + v_{i0}, \tag{3.4.13}$$

and

$$v_{i0} = \epsilon_{i0} + u^*_{i0} + \eta_i, \qquad i = 1, \ldots, N. \tag{3.4.14}$$

The coefficients $\boldsymbol{\pi}_{0t}$ are identical across i (Hsiao, Pesaran, and Tahmiscioglu 1999; Hsiao and Zhang 2015). The error term v_{i0} is the sum of three components: the prediction error of θ_{i0}, ϵ_{i0}, the cumulative shocks before time zero, $u^*_{i0} = u_{i0} + \gamma u_{i,-1} + \gamma^2 u_{i,-2} + \cdots$, and the individual effects, $\eta_i = \frac{1}{1-\gamma}\alpha_i$. The prediction error ϵ_{i0} is independent of u_{it} and η_i, with mean zero and variance $\sigma^2_{\epsilon 0}$. Depending on whether or not the error process of y_{i0} conditional on the exogenous variables has achieved stationarity (i.e., whether or not the variance of y_{i0} is the same as any other y_{it}), we have the case of

$$\begin{aligned} &\mathrm{Var}(v_{i0}) = \sigma^2_{\epsilon 0} + \frac{\sigma^2_u}{1-\gamma^2} + \frac{\sigma^2_\alpha}{(1-\gamma)^2} \\ \text{and} \quad &\mathrm{Cov}(v_{i0}, v_{it}) = \frac{\sigma^2_\alpha}{(1-\gamma)}, \qquad t = 1, \ldots, T, \end{aligned} \tag{3.4.15}$$

or

$$\mathrm{Var}(v_{i0}) = \sigma^2_{w0} \quad \text{and} \quad \mathrm{Cov}(v_{i0}, v_{it}) = \sigma^2_\tau, \qquad t = 1, \ldots, T. \tag{3.4.16}$$

Then the likelihood function of (3.4.2) and (3.4.13) is

$$\tilde{L} = (2\pi)^{-\frac{N(T+1)}{2}} \mid \tilde{V} \mid^{-\frac{N}{2}} \cdot \exp\left\{ -\frac{1}{2} \tilde{\boldsymbol{v}}'_i \tilde{V}^{-1} \tilde{\boldsymbol{v}}_i \right\}, \tag{3.4.17}$$

where $(v_{i0}, \boldsymbol{v}_i) = (y_{io} - \sum_{t=1}^{T} \boldsymbol{\pi}'_{ot} \boldsymbol{x}_{it} - \boldsymbol{\rho}^{*\prime} z_i, y_{i1} - \gamma y_{i0} - \boldsymbol{\beta}' \boldsymbol{x}_{i1} - \boldsymbol{\rho}' z_i, \ldots, y_{iT} - \gamma y_{i,T-1} - \boldsymbol{\beta}' \boldsymbol{x}_{iT} - \boldsymbol{\rho}' z_i)'$ and for notational ease we use z_i to indicate both genuine time-invariant explanatory variables over t and $\bar{\boldsymbol{x}}_i$. The covariance matrix $\tilde{V}$ take the form,

$$\tilde{V} = E[(v_{i0}, \boldsymbol{v}'_i)'(v_{i0}, \boldsymbol{v}'_i)] = \begin{bmatrix} \sigma^2_0 & \sigma^2_\tau \boldsymbol{e}' \\ \sigma^2_\tau \boldsymbol{e} & V \end{bmatrix}, \tag{3.4.18}$$

and V is defined as (3.4.5).

Equation (3.4.13) transforms (3.4.10), in which the number of parameters increases with the number of observations, into a situation in which N independently distributed $(T+1)$-component vectors depend only on a fixed number of parameters. Therefore, the MLE is consistent when $N \to \infty$ or both N and $T \to \infty$. Moreover, the MLE multiplied by the scale factor $\sqrt{NT}$ is centered at the true values independent of the way N or T go to infinity (for detail, see Hsiao and Zhang 2015). When N is fixed and $T \to \infty$, $(\gamma, \boldsymbol{\beta}')$ remain consistent and asymptotically normally distributed but $\boldsymbol{\pi}'_0, \boldsymbol{\rho}', \boldsymbol{\rho}^{*\prime}, \sigma^2_\alpha$, cannot be consistently estimated.

The MLE is obtained by solving the first-order conditions of the likelihood function with respect to unknown parameters. If there is a unique solution to these partial-derivative equations with $\sigma_\alpha^2 > 0$, the solution is the MLE. However, just as in the static case discussed in Section 3.3, a boundary solution with $\sigma_\alpha^2 = 0$ may occur for dynamic error-components models as well. Anderson and Hsiao (1981) have derived the conditions under which the boundary solution will occur for various cases. Trognon (1978) provided analytic explanations based on asymptotic approximations where the number of time periods tends to infinity. Nerlove (1967, 1971a) conducted Monte Carlo experiments to explore the properties of the MLE. These results show that the autocorrelation structure of the exogenous variables is a determinant of the existence of boundary solutions. In general, the more autocorrelated the exogenous variables or the more important the weight of the exogenous variables, the less likely it is that a boundary solution will occur.

The solution for the MLE is complicated. We can apply the Newton–Raphson type iterative procedure or the sequential iterative procedure suggested by Anderson and Hsiao (1982) to obtain a solution. Alternatively, because we have a cross section of size N repeated successively in T time periods, we can regard the problems of estimation (and testing) of (3.4.2) as akin to those for a simultaneous-equations system with T or $T+1$ structural equations with N observations available on each of the equations. That is, the dynamic relationship (3.4.17) in a given time period is written as an equation in a system of simultaneous equations,

$$\Gamma Y' + BX' + PZ' = U', \tag{3.4.19}$$

where we now let[5]

$$\underset{N\times(T+1)}{Y} = \begin{bmatrix} y_{10} & y_{11} & \cdots y_{1T} \\ y_{20} & y_{21} & \cdots y_{2T} \\ \vdots & & \\ y_{N0} & y_{N1} & \cdots y_{NT} \end{bmatrix},$$

$$\underset{N\times TK_1}{X} = \begin{bmatrix} \boldsymbol{x}'_{11} & \boldsymbol{x}'_{12} & \cdots & \boldsymbol{x}'_{1T} \\ \boldsymbol{x}'_{21} & \boldsymbol{x}'_{22} & \cdots & \boldsymbol{x}'_{2T} \\ \vdots & & & \\ \boldsymbol{x}'_{N1} & \boldsymbol{x}'_{N2} & \cdots & \boldsymbol{x}'_{NT} \end{bmatrix},$$

$$\underset{N\times K_2}{Z} = \begin{bmatrix} \boldsymbol{z}'_1 \\ \boldsymbol{z}'_2 \\ \vdots \\ \boldsymbol{z}'_N \end{bmatrix},$$

and U is the $N \times T$ matrix of the errors if the initial values, y_{i0}, are treated as constants, and the $N \times (T+1)$ matrix of errors if the initial values are treated as stochastic. The structural-form coefficient matrix $A = [\Gamma \;\; B \;\; P]$ is $(T+1) \times [(T+1) + TK_1 + K_2]$. The earlier serial covariance matrix [e.g., (3.4.11) or (3.4.18)] now becomes the variance–covariance matrix of the errors on $(T+1)$ structural equations. We can then use the algorithm for solving the full-information maximum-likelihood estimator to obtain the MLE.

5 Previously, we combined the intercept term and the time-varying exogenous variables into the $\boldsymbol{x}_{it}$ vector because the property of the MLE for the constant is the same as that of the MLE for the coefficients of time-varying exogenous variables. Now we separate $\boldsymbol{x}'_{it}$ as $(1, \tilde{\boldsymbol{x}}'_{it})$, because we wish to avoid having the constant term appearing more than once in (3.4.10).

There are cross-equation linear restrictions on the structural-form coefficient matrix and restrictions on the variance–covariance matrix. The structural-form coefficient matrix A is a $(T+1) \times [(T+1) + TK_1 + K_2]$ matrix of the form

$$A = \begin{bmatrix} 1 & 0 & . & . & . & 0 & \boldsymbol{\pi}'_{01} & \boldsymbol{\pi}'_{02} & . & . & . & \boldsymbol{\pi}'_{0T} & \boldsymbol{\rho}^{*\prime} \\ -\gamma & 1 & . & . & . & . & \boldsymbol{\beta}' & \mathbf{0}' & . & . & . & \mathbf{0}' & \boldsymbol{\rho}' \\ 0 & -\gamma & . & . & . & . & \mathbf{0}' & \boldsymbol{\beta}' & . & . & . & \mathbf{0}' & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & . & . & . & -\gamma & 1 & \mathbf{0}' & \mathbf{0}' & . & . & . & \boldsymbol{\beta}' & \boldsymbol{\rho}' \end{bmatrix}, \qquad (3.4.20)$$

and the variance–covariance matrix of U is block-diagonal, with the diagonal block a $(T+1) \times (T+1)$ matrix of the form $\tilde{V}$(3.4.18). Bhargava and Sargan (1983) suggest maximizing the likelihood function of (3.4.19) by directly substituting the restrictions into the structural-form coefficient matrix A and the variance–covariance matrix of U'.

To summarize, if initial value y_{i0} is treated as a fixed constant, then the likelihood function takes the form (3.4.4). The MLE of (3.4.4) is inconsistent if $E(y_{i0}\alpha_i) \neq 0$ and T is fixed and $N \longrightarrow \infty$. When T is large, it is consistent. However, if $\frac{N}{T^3} \longrightarrow a \neq 0$, it is asymptotically biased of order $\sqrt{a}$. On the other hand, the MLE of (3.4.17) is consistent and asymptotically unbiased when $N \longrightarrow \infty$ whether T is fixed or large (Hsiao and Zhou 2019). When N is fixed and T is large, the MLE of γ and $\boldsymbol{\beta}$ remains consistent and asymptotically normally distributed. However, ρ cannot be consistently estimated.

Alternatively, we can ignore the restrictions on the variance–covariance matrix of U' and use three-stage least squares (3SLS) methods. Because the restrictions on A are linear, it is easy to obtain the constrained 3SLS estimator of γ, $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, and $\boldsymbol{\rho}^*$ from the unconstrained 3SLS estimator.[6] Or we can use the Chamberlain (1982, 1984) minimum-distance estimator by first obtaining the unconstrained reduced-form coefficients matrix Π, then solving for the structural-form parameters (see Section 2.9). The Chamberlain minimum-distance estimator has the same limiting distribution as the constrained generalized 3SLS estimator (see Chapter 4). However, because the maintained hypothesis in the model implies that the covariance matrix of U' is constrained and in some cases dependent on the parameter γ occurring in the structural form, the constrained 3SLS or the constrained generalized 3SLS is inefficient in comparison with the (full-information) MLE.[7] But if the restrictions on the variance-covariance matrix are not true, the (full-information) MLE imposing the wrong restrictions will generally be inconsistent. But the (constrained) 3SLS or the Chamberlain minimum-distance estimator, because it does not impose any restriction on the covariance matrix of U$'$, remains consistent and is efficient within the class of estimators that do not impose restrictions on the variance–covariance matrix.

3.4.1.3 Testing Some Maintained Hypotheses on Error-Components Assumption or Initial Conditions

As discussed in Sections 3.4.1.1 and 3.4.1.2 the consistency and asymptotic properties for the MLE and GLS of a random-effects model depend on the initial conditions. Unfortunately, in practice we have very little information on the characteristics of the initial observations. Because some of these hypotheses are nested, Bhargava and Sargan (1983) suggest relying on the likelihood principle to test them. For instance, when y_{i0} are

[6] For the formula of the constrained estimator, see Theil (1971, p. 285, equation 8.5).
[7] See Chapter 4.

exogenous, we can test the validity of the error-components formulation by maximizing (3.4.4) with or without the restrictions on the covariance matrix V. Let L^* denote the maximum of log L subject to the restriction of model (3.4.5), and let L^{**} denote the maximum of log L with V being an arbitrary positive definite matrix. Under the null hypothesis, the resulting test statistic $2(L^{**} - L^*)$ is asymptotically chi-square distributed, with $[T(T+1)/2 - 2]$ degrees of freedom.

Similarly, we can test the validity of the error-components formulation under the assumption that y_{i0} are endogenous. Let $\tilde{L}^*$ and $\tilde{L}^{**}$ denote the maximum of the log-likelihood function (3.4.17) with $\tilde{V}$ under the restriction (3.4.18), and no restriction respectively. Then, under the null, $2(\tilde{L}^{**} - \tilde{L}^*)$ is asymptotically chi-square, with $[(T+1)(T+2)/2 - 2]$ degrees of freedom.

With regard to the test that y_{i0} are exogenous, unfortunately, it is not possible to directly compare L with the likelihood function $\tilde{L}$, because in the former case we are considering the density of $(y_{i1}, \ldots, y_{iT})$ assuming y_{i0} to be exogenous, whereas the latter case is the joint density of $(y_{i0}, \ldots, y_{iT})$. However, we can write the joint likelihood function of (3.4.4) and (3.4.13) under the restriction that v_{i0} are independent of η_i (or α_i) and have variance $\sigma^2_{\epsilon o}$. Namely, we impose the restriction that $\text{Cov}(v_{i0}, v_{it}) = 0, t = 1, \ldots, T$, in the $(T+1) \times (T+1)$ variance–covariance matrix of $(y_{i0}, \ldots, y_{iT})$. We denote this log likelihood function by L_e and its maximum by L^*_e. Let L^{**}_e denote the maximum of log L_e with an unrestricted variance-covariance matrix for $(v_{i0}, \ldots, v_{iT})$. Then we can test the exogeneity of y_{i0} using $2(L^{**}_e - L^*_e)$, which is asymptotically chi-square, with T degrees of freedom under the null. It is also possible to test the exogeneity of y_{i0} by constraining the error terms to have a variance-components structure. Suppose the variance–covariance matrix of $(v_{i1}, \ldots, v_{iT})$ is of the form V (Equation 3.4.5). Let $\tilde{L}^*_e$ denote the maximum of the log likelihood function L_e under this restriction and $E(v_{i0}v_{it}) = 0$. Let $\tilde{L}^{**}_e$ denote the maximum of the log likelihood function of $(y_{i0}, \ldots, y_{iT})$ under the restriction that $E\boldsymbol{v}_i\boldsymbol{v}_i' = \tilde{V}^*$, but allowing the variance of v_{i0} and the covariance between v_{i0} and $v_{it}, t = 1, \ldots, T$, to be arbitrary constants σ^2_{w0} *and* σ^2_τ. The statistic $2(\tilde{L}^{**}_e - \tilde{L}^*_e)$ is asymptotically chi-square with one degree of freedom if y_{i0} are exogenous.

3.4.1.4 Some Simulation Evidence

In order to investigate the performance of maximum-likelihood estimators under various assumptions about the initial conditions, Bhargava and Sargan (1983) conducted Monte Carlo studies. Their true model was generated by

$$y_{it} = 1 + 0.5y_{i,t-1} - 0.16z_i + 0.35x_{it} + \alpha_i + u_{it}, \qquad i = 1, \ldots, 100, \quad t = 1, \ldots, 20,$$

where α_i and u_{it} were independently normally distributed, with means zero and variances 0.09 and 0.4225, respectively. The time-varying exogenous variables x_{it} were generated by

$$x_{it} = 0.1t + \phi_i x_{i,t-1} + \omega_{it}, \qquad i = 1, \ldots, 100, \quad t = 1, \ldots, 20,$$

with ϕ_i and ω_{it} independently normally distributed, with means zero and variances 0.01 and 1, respectively. The time-invariant exogenous variables z_i were generated by

$$z_i = -0.2x_{i4} + \omega^*_i, \qquad i = 1, \ldots, 100,$$

and ω^*_i were independently normally distributed, with mean zero and variance 1. The z and the x were held fixed over the replications, and the first 10 observations were discarded.

Table 3.1. *Simulation results for the biases of the MLEs for dynamic random-effects models*

Coefficient of	y_{i0} exogenous, unrestricted covariance matrix	y_{i0} exogenous, error-components formulation	y_{i0} endogenous, unrestricted covariance matrix	y_{i0} endogenous, error-components formulation
Intercept	−0.1993 (0.142)[a]	−0.1156 (0.1155)	−0.0221 (0.1582)	0.0045 (0.105)
z_i	0.0203 (0.0365)	0.0108 (0.0354)	0.0007 (0.0398)	−0.0036 (0.0392)
x_{it}	0.0028 (0.0214)	0.0044 (0.0214)	0.0046 (0.0210)	0.0044 (0.0214)
$y_{i,t-1}$	0.0674 (0.0463)	0.0377 (0.0355)	0.0072 (0.0507)	−0.0028 (0.0312)
$\sigma_\alpha^2/\sigma_u^2$		−0.0499 (0.0591)		0.0011 (0.0588)

[a] Means of the estimated standard errors in parentheses.
Source: Bhargava and Sargan (1983).

Thus, the y_{i0} are in fact stochastic and are correlated with the individual effects α_i. Table 3.1 reproduces their results on the biases in the estimates for various models obtained in 50 replications.

In cases where the y_{i0} are treated as endogenous, the MLE performs extremely well, and the biases in the parameters are almost negligible. But this is not so for the MLE where y_{i0} are treated as exogenous. The magnitude of the bias is about one standard error. The boundary solution of $\sigma_\alpha^2 = 0$ occurs in a number of replications for the error-components formulation as well. The likelihood-ratio statistics also rejected the exogeneity of y_{i0} 46 and 50 times, respectively, using the tests $2[L_e^{**} - L_e^*]$ and $2[\tilde{L}_e^{**} - \tilde{L}_e^*]$. Under the endogeneity assumption, the likelihood-ratio statistic $2[L_e^{**} - \tilde{L}_e^{**}]$ rejected the error-components formulation 4 times (out of 50), whereas under the exogeneity assumption, the statistic $2(L^{**} - L^*)$ rejected the error-components formulation 7 times.[8]

3.4.2 Generalized Least Squares Estimator

The likelihood function depends only on a fixed number of parameters. Furthermore, conditional on Ω or $\sigma_u^2, \sigma_\alpha^2, \sigma_0^2$ and σ_τ^2, the MLE is equivalent to the generalized least squares estimator of $\boldsymbol{\delta}' = (\boldsymbol{\pi}', \boldsymbol{\rho}^{*\prime}, \gamma, \boldsymbol{\beta}', \boldsymbol{\rho}')$,

$$\hat{\boldsymbol{\delta}}_{GLS} = \left(\sum_{i=1}^N \tilde{X}_i' \tilde{V}^{-1} \tilde{X}_i\right)^{-1} \left(\sum_{i=1}^N \tilde{X}_i' \tilde{V}^{-1} \tilde{\boldsymbol{y}}_i\right), \tag{3.4.21}$$

where $\tilde{\boldsymbol{y}}_i' = (y_{io}, \ldots, y_{iT})$,

$$\tilde{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1}' & \boldsymbol{x}_{i2}' & \ldots & \boldsymbol{x}_{iT}' & \boldsymbol{z}_i' & 0 & \mathbf{0}' & \mathbf{0} \\ \mathbf{0}' & \ldots & \ldots & \ldots & \mathbf{0}' & y_{i0} & \boldsymbol{x}_{i1}' & \boldsymbol{z}_i' \\ \vdots & & & & \vdots & y_{i1} & \boldsymbol{x}_{i2}' & \boldsymbol{z}_i' \\ \vdots & & & & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & & & & \mathbf{0}' & y_{i,T-1} & \boldsymbol{x}_{iT}' & \boldsymbol{z}_i' \end{pmatrix}.$$

[8] Bhargava and Sargan (1983) did not report the significance level of their tests. Presumably, they used the conventional 5% significance level.

The estimator is consistent and asymptotically normally distributed centered at the true value when $N \to \infty$, whether T is fixed or $\to \infty$.

Blundell and Smith (1991) suggest a conditional GLS procedure. Note conditioning on $v_{i0} = y_{i0} - E(y_{i0} \mid \boldsymbol{x}_i', z_i)$,[9]

$$\boldsymbol{y}_i = \boldsymbol{y}_{i,-1}\gamma + Z_i\boldsymbol{\rho} + X_i\boldsymbol{\beta} + \boldsymbol{\tau} v_{i0} + \boldsymbol{v}_i^*, \tag{3.4.22}$$

where $\boldsymbol{v}_i^* = (v_{i1}^*, \ldots, v_{iT}^*)'$, and $\boldsymbol{\tau}$ is a $T \times 1$ vector of constants with the values depending on the correlation pattern between y_{i0} and α_i. When the covariances between y_{i0} and $(y_{i1}, \ldots, y_{iT})$ are arbitrary, $\boldsymbol{\tau}$ is a $T \times 1$ vector of unrestricted constants. Application of the GLS to (3.4.22) is consistent and asymptotically unbiased as $N \to \infty$ whether T is fixed or $\to \infty$ (Hsiao and Zhou 2019).

When the covariance matrix of $\boldsymbol{v}_i$ or $\boldsymbol{v}_i^*$ is unknown, a feasible GLS estimator can be applied. In the first step, we obtain some consistent estimates of the covariance matrix from the estimated $\boldsymbol{v}_i$ or $\boldsymbol{v}_i^*$. For instance, we can use the IV estimator discussed in Section 3.3 to obtain consistent estimators of γ and $\boldsymbol{\beta}$, then substitute them into $y_{it} - \gamma y_{i,t-1} - \boldsymbol{\beta}'\boldsymbol{x}_{it}$ and regress the resulting value on $\boldsymbol{z}_i$ across individuals to obtain a consistent estimate of $\boldsymbol{\rho}$ provided α_i and z_i are uncorrelated. Substituting estimated $\gamma, \boldsymbol{\beta}$, and $\boldsymbol{\rho}$ into (3.4.2), we obtain estimates of v_{it} for $t = 1, \ldots, T$. The estimates of v_{i0} can be obtained as the residuals of the cross-section regression of (3.4.13). The covariance matrix of $\boldsymbol{v}_i$ can then be estimated using the procedures discussed in Chapter 2. The estimated $\boldsymbol{v}_i^*$ can also be obtained as the residuals of the cross-sectional regression of $\boldsymbol{y}_i - \boldsymbol{y}_{i,-1}\gamma - X_i\boldsymbol{\beta}$ on Z_i and $\boldsymbol{e}\hat{v}_{i0}$. In the second step, we treat the estimated covariance matrix of $\boldsymbol{v}_i$ or $\boldsymbol{v}_i^*$ as if they were known, and we apply the GLS to the system composed of (3.4.2) and (3.4.13), or to the conditional system (3.4.22).

It should be noted that if Cov $(y_{i0}, \alpha_i) \neq 0$, the GLS applied to the system (3.4.2) is inconsistent when T is fixed and $N \to \infty$. This is easily seen by noting that conditional on y_{i0}, the system is of the form (3.4.22). Applying GLS to (3.4.2) is therefore subject to omitted variable bias. However, the asymptotic bias of (3.4.6) is still smaller than that of the OLS or the within estimator (Sevestre and Trognon 1982). When T tends to infinity, the GLS of (3.4.2) is again consistent because the GLS converges to the within (or LSDV) estimator, which becomes consistent.

It should also be noted that contrary to the static case, the feasible GLS is asymptotically less efficient than the GLS knowing the true covariance matrix because when a lagged dependent variable appears as one of the regressors, the estimation of slope coefficients is no longer asymptotically independent of the estimation of the parameters of the covariance matrix (Amemiya and Fuller 1967; Hsiao, Pesaran, and Tahmiscioglu 1999; see also Appendix 3A).

3.4.3 An Example: Demand for Natural Gas

We have discussed the properties of various estimators for dynamic models with individual-specific effects. In this section we report results from the study of demand for natural gas conducted by Balestra and Nerlove (1966) to illustrate the specific issues involved in estimating dynamic models using observations drawn from a time series of cross sections.

Balestra and Nerlove (1966) assumed that the new demand for gas (inclusive of demand due to the replacement of gas appliances and the demand due to net increases in the stock of such appliances), G^*, was a linear function of the relative price of gas, P, and the total

[9] It should be noted that y_{it} conditional on $y_{i,t-1}$ and y_{i0} will not give a consistent estimator because $E(y_{i0}) = \theta_{i0}$. In other words, the residual will have a mean different from 0 and the mean varying with i will give rise to the incidental parameters problem.

new requirements for all types of fuel, F^*. Let the depreciation rate for gas appliances be r, and assume that the rate of utilization of the stock of appliances is constant; the new demand for gas and the gas consumption at year t, G_t, follow the relation

$$G_t^* = G_t - (1-r)G_{t-1}. \tag{3.4.23}$$

They also postulated a similar relation between the total new demand for all types of fuel and the total fuel consumption, F, with F approximated by a linear function of total population, N, and per-capita income, I. Substituting these relations into (3.4.23), they obtained

$$G_t = \beta_0 + \beta_1 P_t + \beta_2 \Delta N_t + \beta_3 N_{t-1} + \beta_4 \Delta I_t + \beta_5 I_{t-1} + \beta_6 G_{t-1} + v_t, \tag{3.4.24}$$

where $\Delta N_t = N_t - N_{t-1}, \Delta I_t = I_t - I_{t-1}$, and $\beta_6 = 1 - r$.

Balestra and Nerlove used annual U.S. data from 36 states over the period 1957–1967 to estimate the model for residential and commercial demand for natural gas (3.4.24). Because the average age of the stock of gas appliances during this period was relatively young, it was expected that the coefficient of the lagged gas-consumption variable, β_6, would be less than 1, but not too much below 1. The OLS estimates of (3.4.24) are reported in the second column of Table 3.2. The estimated coefficient of G_{t-1} is 1.01. It is clearly incompatible with a priori theoretical expectations, as it implies a negative depreciation rate for gas appliances.

One possible explanation for the foregoing result is that when cross-sectional and time series data are combined in the estimation of (3.4.24), certain effects specific to the individual state may be present in the data. To account for such effects, dummy variables corresponding to the 36 different states were introduced into the model. The resulting dummy-variable estimates are shown in the third column of Table 3.2. The estimated coefficient of the lagged endogenous variable is drastically reduced; in fact, it is reduced to such a low level that it implies a depreciation rate of gas appliances of over 30% – again, highly implausible.

Table 3.2. *Various estimates of the parameters of Balestra and Nerlove's demand-for-gas model (3.4.24) from the pooled sample, 1957–1962.*

Coefficient	OLS	LSDV	GLS
β_0	−3.650 (3.316)[a]	–	−4.091 (11.544)
β_1	−0.0451 (0.0270)	−0.2026 (0.0532)	−0.0879 (0.0468)
β_2	0.0174 (0.0093)	−0.0135 (0.0215)	−0.00122 (0.0190)
β_3	0.00111 (0.00041)	0.0327 (0.0046)	0.00360 (0.00129)
β_4	0.0183 (0.0080)	0.0131 (0.0084)	0.0170 (0.0080)
β_5	0.00326 (0.00197)	0.0044 (0.0101)	0.00354 (0.00622)
β_6	1.010 (0.014)	0.6799 (0.0633)	0.9546 (0.0372)

[a] Figures in parentheses are standard errors for the corresponding coefficients.
Source: Balestra and Nerlove (1966).

Instead of assuming the regional effect to be fixed, they again estimated (3.4.24) by explicitly incorporating individual state-effects variables into the error term, so that $v_{it} = \alpha_i + u_{it}$, where α_i and u_{it} are independent random variables. The two-step GLS estimates under the assumption that the initial observations are fixed are shown in the fourth column of Table 3.2. The estimated coefficient of lagged consumption is 0.9546. The implied depreciation rate is approximately 4.5%, which is in agreement with a priori expectation.[10]

The foregoing results illustrate that by properly taking account of the unobserved heterogeneity in the panel data, Balestra and Nerlove (1966) were able to obtain results that were reasonable on the basis of a priori theoretical considerations that they were not able to obtain through attempts to incorporate other variables into the equation by conventional procedures. Moreover, the least squares and the least squares dummy-variables estimates of the coefficient of the lagged gas-consumption variable were 1.01 and 0.6799, respectively. In Section 3.2 we showed that for dynamic models with individual-specific effects, the least squares estimate of the coefficient of the lagged dependent variable is biased upward and the least-squares dummy-variable estimate is biased downward if T is small. Their estimates are in agreement with these theoretical results.

3.5 THE LIKELIHOOD APPROACH-FIXED-EFFECTS MODELS

When the effects α_i are correlated with the $\boldsymbol{x}_{it}$ model (3.4.2)

$$\begin{aligned}
&\operatorname{plim}\frac{1}{NT}\sum_{i=1}^{N} X_i{}'\left(I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\boldsymbol{e}\boldsymbol{e}'\right)(\boldsymbol{e}\alpha_i + \boldsymbol{u}_i)\\
&\quad = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}\operatorname{plim}\frac{1}{N}\sum_{i=1}^{N}\bar{\boldsymbol{x}}_i\alpha_i\\
&\quad = O\left(\frac{1}{T}\right) \qquad (3.5.1)
\end{aligned}$$

neither the estimator (3.4.6) nor the estimator (3.2.10) is consistent when T is finite. When $T \to \infty$, (3.4.6) or (3.2.10) converges to the covariance estimator and is consistent. However, if N is large, it is asymptotically biased of order $\sqrt{c}$ where $c = \lim \frac{N}{T}$ as $T \to \infty$.

3.5.1 Conditional Mean Approach

As discussed at the beginning of Section 3.4, if the assumption of Mundlak (1978a) and Chamberlain (1982) holds, $E(\alpha_i|\boldsymbol{x}_i) = \boldsymbol{a}'\bar{\boldsymbol{x}}_i$, we can consider $\boldsymbol{x}_i$ as part of the $\boldsymbol{z}_i$ in model (3.4.2) and apply the QMLE or the feasible GLS (3.4.21). The resulting estimator is consistent and asymptotically normally distributed centered at the true value when $N \to \infty$ whether T is finite or goes to infinity.

[10] We do not know the value of the GLS estimates when the initial observations are treated as endogenous. My conjecture is that it is likely to be close to the two-step GLS estimates with fixed initial observations. As discussed in Section 3.4, Sevestre and Trognon (1982) have shown that even the initial values are correlated with the effects; the asymptotic bias of the two-step GLS estimator under the assumption of fixed initial observations is still smaller than the OLS or the within estimator. Moreover, if Bhargava and Sargan's simulation result is any indication, the order of bias due to the wrong assumption about initial observations when T is greater than 10 is about 1 standard error or less. Here, the standard error of the lagged dependent variable for the two-step GLS estimates with fixed initial values is only 0.037.

3.5.2 Linear Difference: The Transformed Likelihood Approach

The first difference equation (3.3.1) no longer contains the individual effects α_i and is well defined for $t = 2, 3, \ldots, T$, under the assumption that the initial observations y_{i0} and $\boldsymbol{x}_{i1}$ are available. But $\Delta y_{i1} = (y_{i1} - y_{i0})$ is not defined because Δy_{i0} and $\Delta \boldsymbol{x}_{i0}$ are missing. However, by continuous substitution, we can write Δy_{i1} as

$$\Delta y_{i1} = a_{i1} + \sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j}, \tag{3.5.2}$$

where $a_{i1} = \boldsymbol{\beta}' \sum_{j=0}^{\infty} \Delta \boldsymbol{x}_{i,1-j} \gamma^j$. Since $\Delta \boldsymbol{x}_{i,1-j}, j = 1, 2, \ldots,$ are unavailable, a_{i1} is unknown. Treating a_{i1} as a free parameter to be estimated will again introduce the incidental parameters problem. To get around this problem, the expected value of a_{i1}, conditional on the observables, has to be a function of a finite number of parameters of the form,

$$E(a_{i1} \mid \Delta \boldsymbol{x}_i) = c^* + \boldsymbol{\pi}' \Delta \boldsymbol{x}_i, \quad i = 1, \ldots, N, \tag{3.5.3}$$

where $\boldsymbol{\pi}$ is a $TK_1 \times 1$ vector of constants, $\Delta \boldsymbol{x}_i$ is a $TK_1 \times 1$ vector of $(\Delta \boldsymbol{x}'_{i1}, \ldots, \Delta \boldsymbol{x}'_{iT})'$. Hsiao, Pesaran, and Tahmiscioglu (2002) have shown that if $\boldsymbol{x}_{it}$ are generated by

$$\boldsymbol{x}_{it} = \boldsymbol{\mu}_i + \boldsymbol{g}t + \sum_{j=0}^{\infty} \boldsymbol{b}'_j \boldsymbol{\xi}_{i,t-j}, \quad \sum_{j=0}^{\infty} \mid \boldsymbol{b}_j \mid < \infty, \tag{3.5.4}$$

where $\boldsymbol{\xi}_{it}$ are assumed to be i.i.d. with mean zero and constant covariance matrix, then (3.5.3) holds. The data generating process of the exogenous variables $\boldsymbol{x}_{it}$ (3.5.4) can allow fixed and different intercepts $\boldsymbol{\mu}_i$ across i, or to have $\boldsymbol{\mu}_i$ randomly distributed with a common mean. However, if there exists a trend term in the data generating process of $\boldsymbol{x}_{it}$, then they must be identical across i for (3.5.3) to hold.

Given (3.5.3), Δy_{i1} can be written as

$$\Delta y_{i1} = c^* + \boldsymbol{\pi}' \Delta \boldsymbol{x}_i + v^*_{i1}, \tag{3.5.5}$$

where $v^*_{i1} = \sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j} + [a_{i1} - E(a_{i1} \mid \Delta \boldsymbol{x}_i)]$. By construction, $E(v^*_{i1} \mid \Delta \boldsymbol{x}_i) = 0$, $E(v^{*2}_{i1}) = \sigma^2_{v^*}$, $E(v^*_{i1} \Delta u_{i2}) = -\sigma^2_u$, and $E(v^*_{i1} \Delta u_{it}) = 0$, for $t = 3, 4, \ldots, T$. It follows that the $T \times T$ covariance matrix of $\Delta \boldsymbol{u}^*_i = (v^*_{i1}, \Delta \boldsymbol{u}'_i)'$ has the form

$$\Omega^* = \sigma^2_u \begin{bmatrix} h & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & & \\ 0 & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & & \\ 0 & & & -1 & 2 \end{bmatrix} = \sigma^2_u \tilde{\Omega}^*, \tag{3.5.6}$$

where $h = \frac{\sigma^2_{v^*}}{\sigma^2_u}$.

Combining (3.3.1) and (3.5.5), we can write the likelihood function of $\Delta \boldsymbol{y}^*_i = (\Delta y_{i1}, \ldots, \Delta y_{iT})', i = 1, \ldots, N$, in the form of

$$(2\pi)^{-\frac{NT}{2}} \mid \Omega^* \mid^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} \Delta \boldsymbol{u}^{*\prime}_i \Omega^{*-1} \Delta \boldsymbol{u}^*_i \right\}, \tag{3.5.7}$$

if $\Delta \boldsymbol{u}_i^*$ is normally distributed, where

$$\Delta \boldsymbol{u}_i^* = [\Delta y_{i1} - c^* - \boldsymbol{\pi}' \Delta \boldsymbol{x}_i, \Delta y_{i2} - \gamma \Delta y_{i1} - \boldsymbol{\beta}' \Delta \boldsymbol{x}_{i2}, \ldots, \Delta y_{iT} - \gamma \Delta y_{i,T-1} - \boldsymbol{\beta}' \Delta \boldsymbol{x}_{iT}]'. \tag{3.5.8}$$

The likelihood function again depends only on a fixed number of parameters and satisfies the standard regularity conditions, so that the MLE is consistent and asymptotically normally distributed as $N \to \infty$.

Since $\mid \tilde{\Omega}^* \mid = 1 + T(h-1)$ and

$$\tilde{\Omega}^{*-1} = [1 + T(h-1)]^{-1} \begin{bmatrix} T & T-1 & \ldots & 2 & 1 \\ T-1 & (T-1)h & & 2h & h \\ \vdots & \vdots & & \vdots & \vdots \\ 2 & 2h & & 2[(T-2)h-(T-3)] & (T-2)h-(T-3) \\ 1 & h & & (T-2)h-(T-3) & (T-1)h-(T-2) \end{bmatrix}, \tag{3.5.9}$$

the logarithm of the likelihood function (3.5.7) is

$$\begin{aligned} lnL = & -\frac{NT}{2} \log 2\pi - \frac{NT}{2} \log \sigma_u^2 - \frac{N}{2} \log [1 + T(h-1)] \\ & - \frac{1}{2} \sum_{i=1}^{N} \left[(\Delta \boldsymbol{y}_i^* - H_i \boldsymbol{\psi})' \Omega^{*-1} (\Delta \boldsymbol{y}_i^* - H_i \boldsymbol{\psi}) \right], \end{aligned} \tag{3.5.10}$$

where $\boldsymbol{\psi} = (c^*, \boldsymbol{\pi}', \gamma, \boldsymbol{\beta}')'$, and

$$H_i = \begin{bmatrix} 1 & \Delta \boldsymbol{x}_i' & 0 & \boldsymbol{0}' \\ 0 & \boldsymbol{0}' & \Delta y_{i1} & \Delta \boldsymbol{x}_{i2}' \\ \vdots & & \vdots & \vdots \\ 0 & \boldsymbol{0}' & \Delta y_{i,T-1} & \Delta \boldsymbol{x}_{iT}' \end{bmatrix}.$$

The MLE is obtained by solving the following equations simultaneously:

$$\hat{\boldsymbol{\psi}} = \left(\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} H_i \right)^{-1} \left(\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} \Delta \boldsymbol{y}_i^* \right), \tag{3.5.11}$$

$$\hat{\sigma}_u^2 = \frac{1}{NT} \sum_{i=1}^{N} \left[(\Delta \boldsymbol{y}_i^* - H_i \hat{\boldsymbol{\psi}})' (\hat{\tilde{\Omega}}^*)^{-1} (\Delta \boldsymbol{y}_i^* - H_i \hat{\boldsymbol{\psi}}) \right], \tag{3.5.12}$$

$$\hat{h} = \frac{T-1}{T} + \frac{1}{\hat{\sigma}_u^2 NT^2} \sum_{i=1}^{N} \left[(\Delta \boldsymbol{y}_i^* - H_i \hat{\boldsymbol{\psi}})' (\boldsymbol{J}\boldsymbol{J}') (\Delta \boldsymbol{y}_i^* - H_i \hat{\boldsymbol{\psi}}) \right], \tag{3.5.13}$$

where $\boldsymbol{J}' = (T, T-1, \ldots, 2, 1)$. One way to obtain the MLE is to iterate among (3.5.11)–(3.5.13) conditionally on the early round estimates of the other parameters until the solution converges or to use the Newton–Raphson type iterative scheme (Hsiao, Pesaran, and Tahmiscioglu 2002).

For finite N, occasionally, the transformed MLE breaks down, giving estimated γ greater than unity or negative variance estimates. However, the problem quickly disappears as N becomes large. For further discussions on the properties of transformed MLE when $\gamma = 1$ or approaches to 1 or explosive, see Kruiniger (2009) and Han and Phillips (2013).

3.5.3 Minimum Distance Estimator

Conditional on Ω^*, the MLE (3.5.11) is the minimum distance estimator (MDE) of the form

$$\text{Min} \sum_{i=1}^{N} \Delta \boldsymbol{u}_i^{*\prime} \Omega^{*-1} \Delta \boldsymbol{u}_i^*. \tag{3.5.14}$$

In the case that Ω^* is unknown, a two-step feasible MDE can be implemented. In the first step we obtain consistent estimators of σ_u^2 and σ_{v*}^2. For instance, we can regress (3.5.5) across i to obtain the least squares residuals $\hat{v}_{i1}^*$, then estimate

$$\hat{\sigma}_{v*}^2 = \frac{1}{N - TK_1 - 1} \sum_{i=1}^{N} \hat{v}_{i1}^{*2}. \tag{3.5.15}$$

Similarly, we can apply the IV to (3.3.1) and obtain the estimated residuals $\Delta \hat{u}_{it}$, then estimate σ_u^2 by

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \Delta \hat{\boldsymbol{u}}_i' \tilde{A}^{-1} \Delta \hat{\boldsymbol{u}}_i \tag{3.5.16}$$

where $\tilde{A}$ is defined in (3.3.18).

In the second step, we substitute estimated σ_u^2 and σ_{v*}^2 into (3.5.6) and treat them as if they were known, then use (3.5.11) to obtain the MDE of $\boldsymbol{\psi}$, $\hat{\boldsymbol{\psi}}_{MDE}$.

The asymptotic covariance matrix of MDE, Var $(\hat{\boldsymbol{\psi}}_{MDE})$, using the true Ω^* as the weighting matrix is equal to $(\sum_{i=1}^{N} H_i' \Omega^{*-1} H_i)^{-1}$. The asymptotic covariance of the feasible MDE using a consistently estimated Ω^*, Var $(\hat{\boldsymbol{\psi}}_{FMDE})$, contrary to the static case, is equal to (Hsiao, Pesaran, and Tahmiscioglu 2002)

$$\begin{aligned} &\left(\frac{1}{N}\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1} + \left(\frac{1}{N}\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1} \\ &\begin{bmatrix} 0 & \mathbf{0}' & 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0}' & d & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \left(\frac{1}{N}\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1}, \end{aligned} \tag{3.5.17}$$

where

$$\begin{aligned} d = {} & \frac{[\gamma^{T-2} + 2\gamma^{T-3} + \cdots + (T-1)]^2}{[1 + T(h-1)]^2 \sigma_u^4} \\ & \times \left(\sigma_u^4 \operatorname{Var}(\hat{\sigma}_{v*}^2) + \sigma_{v*}^4 \operatorname{Var}(\hat{\sigma}_u^2) - 2\sigma_u^2\sigma_{v*}^2 \operatorname{Cov}(\hat{\sigma}_{v*}^2, \hat{\sigma}_u^2)\right). \end{aligned}$$

The second term of (3.5.17) arises because the estimation of $\boldsymbol{\psi}$ and Ω^* are not asymptotically independent when the lagged dependent variables also appear as regressors.

3.5.4 Issues of Random- versus Fixed-Effects Specification

When T is fixed and $N \to \infty$, the GMM or the MLE of the transformed likelihood function (3.5.7) or the MDE (3.5.11) is consistent and asymptotically normally distributed whether α_i are fixed or random when $N \to \infty$ whether T is fixed or $\to \infty$. When α_i are random and uncorrelated with $\boldsymbol{x}_{it}$, the likelihood function of (3.4.17) uses the level variables while (3.5.7) uses the first difference variables. In general, the variation across individuals are greater than the variation within individuals. Moreover, first differencing reduces the number of time series observations by one per cross-sectional unit; hence, maximizing (3.5.7) yields estimators that will not be as efficient as the MLE of (3.4.17) when α_i are indeed random and uncorrelated with $\boldsymbol{x}_i$. However, if α_i are fixed or correlated with $\boldsymbol{x}_{it}$, the MLE of (3.4.17) yields an inconsistent estimator.

The transformed MLE or MDE is consistent under a more general data generating process of $\boldsymbol{x}_{it}$ than the MLE of (3.4.17) or the GLS (3.4.21). In order for the Bhargava and Sargan (1983) MLE of the random-effects model to be consistent, we will have to assume that the $\boldsymbol{x}_{it}$ are generated from the same stationary process with common means (Equation 3.4.12). Otherwise, $E(y_{i0} \mid \boldsymbol{x}_i) = \boldsymbol{c}_i + \boldsymbol{\pi}_i' \boldsymbol{x}_i$, where $\boldsymbol{c}_i$ and $\boldsymbol{\pi}_i$ vary across i, and we will have the incidental parameters problem again. On the other hand, the transformed likelihood approach allows $\boldsymbol{x}_{it}$ to have different means (or intercepts) (Equation 3.5.4). Limited Monte Carlo analysis conducted by Hsiao (2020b) indicates that there is not much difference between the conditional mean method (3.4.17) and the linear difference method (3.5.7) in terms of bias, root mean square error, and the size of the test when the homogeneity assumption for the data generating process of $\boldsymbol{x}_{it}$ holds (Equation 3.4.12). However, if $\boldsymbol{x}_{it}$ have different means for different i, the linear difference method (3.5.7) or (3.5.14) appears to dominate (Hsiao 2020b).

Therefore, if one is not sure about the assumption of the effects, α_i, or the homogeneity assumption about the data generating process of $\boldsymbol{x}_{it}$, one probably should work with the transformed likelihood function (3.5.7) or the MDE (3.5.14) despite that one may lose efficiency under the ideal condition.

The use of the transformed likelihood approach also offers the possibility of using a Hausman (1978)-type test for fixed versus random effects specification or test for the homogeneity and stationarity assumption about the $\boldsymbol{x}_{it}$ process under the assumption that α_i are random. Under the null of random effects and homogeneity of the $\boldsymbol{x}_{it}$ process, the MLE of (3.4.17) is asymptotically efficient. The transformed MLE of (3.5.7) is consistent, but not efficient. On the other hand, if α_i are fixed or $\boldsymbol{x}_{it}$ is not generated by a homogeneous process but satisfies (3.5.4), the transformed MLE of (3.5.7) is consistent, but the MLE of (3.4.17) is inconsistent. Therefore, a Hausman-type test statics (2.5.2) can be constructed by comparing the difference between the two estimators.

3.6 RELATIONS BETWEEN THE LIKELIHOOD BASED ESTIMATOR AND THE GMM

The likelihood approach considers either the likelihood function of $T+1$ random variables $(y_{i0}, y_{i1}, \ldots, y_{iT})$ or T random variables $\Delta \boldsymbol{y}_i = (\Delta y_{i1}, \ldots, \Delta y_{iT})$ for each i, while the GMM considers the $(T-1)$ random variables $(\Delta y_{i2}, \ldots, \Delta y_{iT})$. Given that the GMM is based on the first difference equation, we focus our discussion on the first difference form.

We note that GMM is inconsistent if N is fixed no matter how large T is, while the transformed MLE remains consistent and asymptotically normally distributed. Although normality is assumed to derive the transformed MLE and MDE, it is not required. Both estimators remain consistent and asymptotically normally distributed, and even the errors are not normally distributed. Under normality, the transformed MLE achieves the Cramér-Rao lower bound; hence, it is fully efficient. Even without normality, the transformed MLE (or MDE if Ω^* is known) is more efficient than the GMM that only uses second-moment restrictions.

GMM and the transformed MLE of (3.5.7) or (3.5.11) are consistent and asymptotically normally distributed when T is fixed and $N \rightarrow \infty$. Using the formula of partitioned inverse (e.g., Amemiya 1985), the covariance matrix of the minimum distance estimator of $(\gamma, \boldsymbol{\beta})$ is of the form

$$\text{Cov}\begin{pmatrix}\hat{\gamma}_{MDE}\\ \hat{\boldsymbol{\beta}}_{MDE}\end{pmatrix} = \sigma_u^2\left[\sum_{i=1}^{N}\left(\Delta \boldsymbol{y}_{i,-1}, \Delta X_i\right)'\left(\tilde{A} - \frac{1}{h}\boldsymbol{g}\boldsymbol{g}'\right)^{-1}\left(\Delta \boldsymbol{y}_{i,-1},\ \Delta X_i\right)\right]^{-1} \tag{3.6.1}$$

where $\boldsymbol{g}' = (-1, 0, \ldots, 0)$.

We note that (3.6.1) is smaller than

$$\sigma_u^2\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta \boldsymbol{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix}\tilde{A}^{-1}(\Delta \boldsymbol{y}_{i,-1}, \Delta X_i)\right]^{-1}, \tag{3.6.2}$$

in the sense that the difference between the two matrices is a nonpositive semi-definite matrix, because $\tilde{A} - (\tilde{A} - \frac{1}{h}\boldsymbol{g}\boldsymbol{g}')$ is a positive semi-definite matrix. Furthermore,

$$\begin{aligned}
&\sum_{i=1}^{N}\begin{pmatrix}\Delta \boldsymbol{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix}\tilde{A}^{-1}(\Delta \boldsymbol{y}_{i,-1}, \Delta X_i) - \left[\textstyle\sum_{i=1}^{N}\begin{pmatrix}\Delta \boldsymbol{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix} W_i'\right]\left(\sum_{i=1}^{N} W_i\tilde{A}W_i'\right)^{-1}\\
&\quad\cdot\left[\sum_{i=1}^{N} W_i(\Delta y_{i,-1}, \Delta X_i)\right]\\
&\quad = D'[I - Q(Q'Q)^{-1}Q]D,
\end{aligned} \tag{3.6.3}$$

is a positive semi-definite matrix, where $D = (D_1', \ldots, D_N')', Q = (Q_1', Q_2', \ldots, Q_N')'$, $D_i = \Lambda'(\Delta \boldsymbol{y}_{i,-1}, \Delta X_i), Q_i = \Lambda^{-1}W_i$, and $\Lambda\Lambda' = \tilde{A}^{-1}$. Therefore, the asymptotic covariance matrix of the GMM estimator (3.3.17), (3.3.18), is greater than (3.6.2), which is in the sense that the difference of the two covariance matrix is a positive semi-definite matrix.

When $\tilde{\Omega}^*$ is unknown, the asymptotic covariance matrix of the GMM (3.3.17) remains as (3.3.18). But the asymptotic covariance matrix of the feasible MDE is (3.5.17). Although the first term of (3.5.17) is smaller than (3.3.18), it is not clear that with the addition of the second term, it will remain smaller than (3.3.18). However, it is very likely so due to several factors. First, additional information due to the Δy_{i1} equation is utilized which can be substantial (e.g., see Hahn 1999). Second, the GMM method uses the $(t-1)$ instruments $(y_{i0}, \ldots, y_{i,t-2})$ for the Δy_{it} equation for $t = 2, 3, \ldots, T$; the likelihood approach uses the t instruments $(y_{i0}, y_{i1}, \ldots, y_{i,t-1})$. Third, the likelihood approach uses the condition that $E(H_i'\Omega^{*-1}\Delta\boldsymbol{u}_i^*) = \boldsymbol{0}$ which is approximated by $\frac{1}{NT}\sum_{i=1}^{\mathbf{N}}\sum_{t=1}^{T} H_i'\Omega^{*-1}\Delta\boldsymbol{u}_i^* = \boldsymbol{0}$, while the GMM method uses the moment condition $E(W_i\Delta\boldsymbol{u}_i) = 0$ which is approximated by $\frac{1}{N}\sum_{i=1}^{\mathbf{N}} W_i\Delta\hat{\boldsymbol{u}}_i = \boldsymbol{0}$. The grouping of observations in general will lead to a loss of information.[11]

[11] For additional discussions on the contribution of initial observations, see Blundell and Bond (1998) and Hahn (1999).

Although both the GMM and the likelihood-based estimator are consistent, when N is large, the process of removing the individual-specific effects in a dynamic model creates the order 1, $O(1)$, correlation between $(y_{i,t-1} - y_{i,t-1})$ and $(u_{it} - u_{i,t-1})$. The likelihood approach uses all NT observations to approximate the population moment $E(H_i' \Omega^{*-1} \Delta \boldsymbol{u}_i^*) = \mathbf{0}$; hence it is asymptotically unbiased independent of the way N or $T \longrightarrow \infty$ (Hsiao and Zhang 2015). The GMM (or instrumental variable) approach transforms the correlation between $(y_{it} - y_{i,t-1})$ and $(u_{it} - u_{i,t-1})$ into the correlation between $\frac{1}{N}\sum_{i=1}^{\mathbf{N}} \boldsymbol{q}_{it}(y_{it} - y_{i,t-1})$ and $\frac{1}{N}\sum_{i=1}^{\mathbf{N}} \boldsymbol{q}_{it}(u_{it} - u_{i,t-1})$, which is of order $\frac{1}{N}, O\left(\frac{1}{N}\right)$. Therefore, when T is fixed and N is large, the GMM estimator is consistent and $\sqrt{N}(\hat{\gamma}_{GMM} - \gamma)$ is centered at zero. However, the number of moment conditions for the GMM (say, Equation 3.3.10) is (or increases) at the order of T^2. This could create finite sample bias (e.g., see Ziliak 1997). When both N and T are large, and $\frac{T}{N} \longrightarrow a, 0 < a < \infty$ as $N \rightarrow \infty$, the effects of the correlations due to $\frac{1}{N}\sum_{i=1}^{\mathbf{N}} \boldsymbol{q}_{it}(y_{it} - y_{i,t-1})$ and $\frac{1}{N}\sum_{i=1}^{\mathbf{N}} \boldsymbol{q}_{it}(u_{it} - u_{i,t-1})$ become magnified. Alvarez and Arellano (2003) show that $\sqrt{NT}\hat{\gamma}_{GMM}$ has asymptotic bias equal to $-\sqrt{a}(1+\gamma)$ for models with $\boldsymbol{\beta} = 0$. On the other hand, the likelihood-based estimator is asymptotically unbiased (Hsiao and Zhang 2015). In other words, the difference between γ and the GMM estimator multiplied by the scale factor $\sqrt{NT}$ is not centered at 0, but the likelihood-based estimator is.[12] Whether an estimator is asymptotically biased or not has important implications in statistical inference because in hypothesis testing, we typically normalize the estimated γ by the inverse of its standard error, which is equivalent to multiplying the estimator by the scale factor $\sqrt{NT}$.

Hsiao, Pesaran, and Tahmiscioglu (2002) conducted Monte Carlo studies to compare the performance of the IV of (3.3.2), the GMM of (3.3.17), the MLE and the MDE when T is small and N is finite. They generate y_{it} by

$$y_{it} = \alpha_i + \gamma y_{i,t-1} + \beta x_{it} + u_{it}, \tag{3.6.4}$$

where the error term u_{it} is generated from two schemes. One is from $N(0, \sigma_u^2)$. The other is from mean-adjusted chi-square with two degrees of freedom. The regressor x_{it} is generated according to

$$x_{it} = \mu_i + gt + \xi_{it}, \tag{3.6.5}$$

where ξ_{it} follows an autoregressive moving average process,

$$\xi_{it} - \phi \xi_{i,t-1} = \epsilon_{it} + \theta \epsilon_{i,t-1} \tag{3.6.6}$$

and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$. The fixed effects μ_i and α_i are generated from a variety of schemes, such as being correlated with x_{it} or uncorrelated with x_{it}, but from a mixture of different distributions. Table 3.3 gives a summary of the different designs of the Monte Carlo study.

In generating y_{it} and x_{it}, both are assumed to start from zero. But the first 50 observations are discarded. The bias and root mean squares error (RMSE) of various estimators of γ and β when $T = 5$ and $N = 50$ based on 2,500 replications are reported in Tables 3.4 and 3.5, respectively. The results show that the bias of the MLE of γ as a percentage of the true value is smaller than 1% in most cases. The bias of the IV of γ can be significant for certain data generating processes. In particular, if γ is close to 1, the GMM

12 As a matter of fact, Alvarez and Arellano (2003) show that the least variance ratio estimator (which they call "the limited information maximum likelihood estimator") has asymptotic bias of order $\frac{1}{2N-T}$ when $0 < c < 2$. However, it appears that their forward deviation approach works only under fixed initial conditions. When the initial condition is treated as random, there is no asymptotic bias (see Hsiao and Zhang 2015).

Table 3.3. *Monte Carlo design*

Design number	γ	β	ϕ	θ	g	$R^2_{\Delta y}$	σ_e
1	0.4	0.6	0.5	0.5	0.01	0.2	0.800
2	0.4	0.6	0.9	0.5	0.01	0.2	0.731
3	0.4	0.6	1	0.5	0.01	0.2	0.711
4	0.4	0.6	0.5	0.5	0.01	0.4	1.307
5	0.4	0.6	0.9	0.5	0.01	0.4	1.194
6	0.4	0.6	1	0.5	0.01	0.4	1.161
7	0.8	0.2	0.5	0.5	0.01	0.2	1.875
8	0.8	0.2	0.9	0.5	0.01	0.2	1.302
9	0.8	0.2	1	0.5	0.01	0.2	1.104
10	0.8	0.2	0.5	0.5	0.01	0.4	3.062
11	0.8	0.2	0.9	0.5	0.01	0.4	2.127
12	0.8	0.2	1	0.5	0.01	0.4	1.803

Source: Balestra and Nerlove (1966).

Table 3.4. *Bias of estimators ($T = 5$ and $N = 50$)*

		Bias			
Design	Coeff.	IVE	MDE	MLE	GMM
1	$\gamma = 0.4$	0.0076201	−0.050757	−0.000617	−0.069804
	$\beta = 0.6$	−0.001426	0.0120812	0.0023605	0.0161645
2	$\gamma = 0.4$	0.0220038	−0.052165	−0.004063	−0.072216
	$\beta = 0.6$	−0.007492	0.0232612	0.0027946	0.0321212
3	$\gamma = 0.4$	1.3986691	−0.054404	−0.003206	−0.075655
	$\beta = 0.6$	−0.386998	0.0257393	0.0002997	0.0365942
4	$\gamma = 0.4$	0.0040637	−0.026051	−0.001936	−0.03616
	$\beta = 0.6$	0.0004229	0.0066165	0.0019218	0.0087369
5	$\gamma = 0.4$	0.1253257	−0.023365	−0.000211	−0.033046
	$\beta = 0.6$	−0.031759	0.0113724	0.0016388	0.0155831
6	$\gamma = 0.4$	−0.310397	−0.028377	−0.00351	−0.040491
	$\beta = 0.6$	0.0640605	0.0146638	0.0022274	0.0209054
7	$\gamma = 0.8$	−0.629171	−0.108539	0.009826	−0.130115
	$\beta = 0.2$	−0.018477	0.0007923	0.0026593	0.0007962
8	$\gamma = 0.8$	−1.724137	−0.101727	0.0027668	−0.128013
	$\beta = 0.2$	0.0612431	0.0109865	−0.000011	0.013986
9	$\gamma = 0.8$	−0.755159	−0.102658	0.00624	−0.133843
	$\beta = 0.2$	−0.160613	0.0220208	0.0002624	0.0284606
10	$\gamma = 0.8$	0.1550445	−0.045889	0.001683	−0.05537
	$\beta = 0.2$	0.0096871	0.0000148	0.0007889	−0.000041
11	$\gamma = 0.8$	−0.141257	−0.038216	−0.000313	−0.050427
	$\beta = 0.2$	0.0207338	0.0048828	0.0007621	0.0063229
12	$\gamma = 0.8$	0.5458734	−0.039023	0.0005702	−0.053747
	$\beta = 0.2$	−0.069023	0.0079627	0.0003263	0.010902

Source: Hsiao, Pesaran, and Tahmiscioglu (2002, Table 2).

method could run into a weak IV problem (for analytical results, see Kruiniger 2009). The MDE and GMM of γ also have substantial downward biases in all designs. The bias of the GMM estimator of γ can be as large as 15%–20% in many cases and is larger than the bias of the MDE. The MLE also has the smallest RMSE followed by the MDE, then the GMM. The IV has the largest RMSE.

Table 3.5. *Root mean square error ($T = 5$ and $N = 50$)*

		Root mean square error			
Design	Coeff.	IVE	MDE	MLE	GMM
1	$\gamma = 0.4$	0.1861035	0.086524	0.0768626	0.1124465
	$\beta = 0.6$	0.1032755	0.0784007	0.0778179	0.0800119
2	$\gamma = 0.4$	0.5386099	0.0877669	0.0767981	0.11512
	$\beta = 0.6$	0.1514231	0.0855346	0.0838699	0.091124
3	$\gamma = 0.4$	51.487282	0.0889483	0.0787108	0.1177141
	$\beta = 0.6$	15.089928	0.0867431	0.0848715	0.0946891
4	$\gamma = 0.4$	0.1611908	0.0607957	0.0572515	0.0726422
	$\beta = 0.6$	0.0633505	0.0490314	0.0489283	0.0497323
5	$\gamma = 0.4$	2.3226456	0.0597076	0.0574316	0.0711803
	$\beta = 0.6$	0.6097378	0.0529131	0.0523433	0.0556706
6	$\gamma = 0.4$	14.473198	0.0620045	0.0571656	0.0767767
	$\beta = 0.6$	2.9170627	0.0562023	0.0550687	0.0607588
7	$\gamma = 0.8$	27.299614	0.1327602	0.116387	0.1654403
	$\beta = 0.2$	1.2424372	0.0331008	0.0340688	0.0332449
8	$\gamma = 0.8$	65.526156	0.1254994	0.1041461	0.1631983
	$\beta = 0.2$	3.2974597	0.043206	0.0435698	0.0450143
9	$\gamma = 0.8$	89.83669	0.1271169	0.104646	0.1706031
	$\beta = 0.2$	5.2252014	0.0535363	0.0523473	0.0582538
10	$\gamma = 0.8$	12.201019	0.074464	0.0715665	0.0884389
	$\beta = 0.2$	0.6729934	0.0203195	0.020523	0.0203621
11	$\gamma = 0.8$	17.408874	0.0661821	0.0642971	0.0822454
	$\beta = 0.2$	1.2541247	0.0268981	0.026975	0.02756742
12	$\gamma = 0.8$	26.439613	0.0674678	0.0645253	0.0852814
	$\beta = 0.2$	2.8278901	0.0323355	0.0323402	0.0338716

Source: Hsiao, Pesaran, and Tahmiscioglu (2002, Table 5).

Table 3.6. *GMM with many IVs*

N	T	Gamma	Mean	MSE	Actual size	Bias	Bias percentage
50	50	0.2	0.1750	0.0011	0.2290	−0.0250	−0.1251
50	50	0.5	0.4671	0.0015	0.3860	−0.0329	−0.0657
50	50	0.8	0.7536	0.0024	0.7630	−0.0464	−0.0580
50	100	0.2	0.1872	0.0004	0.1380	−0.0128	−0.0639
50	100	0.5	0.4831	0.0005	0.2350	−0.0169	−0.0339
50	100	0.8	0.7790	0.0005	0.5120	−0.0210	−0.0263
100	100	0.2	0.1879	0.0003	0.2090	−0.0121	−0.0603
100	100	0.5	0.4837	0.0004	0.4170	−0.0163	−0.0326
100	100	0.8	0.7794	0.0005	0.8130	−0.0206	−0.0258
25	100	0.2	0.1869	0.0006	0.1030	−0.0131	−0.0656
25	100	0.5	0.4830	0.0006	0.1450	−0.0170	−0.0341
25	100	0.8	0.7790	0.0006	0.3490	−0.0210	−0.0262

Source: Hsiao and Zhang (2015).

Hsiao and Zhang (2015) further conducted Monte Carlo studies on the model

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + u_{it}, \quad \begin{aligned} i &= 1, \ldots, N \\ t &= 0, 1, \ldots, T \end{aligned} \tag{3.6.7}$$

with $\gamma = 0.2, 0.5$, and 0.8 and u_{it} of either $N(0, 1)$ or mean-adjusted chi-square distributed with 2 degrees of freedom for different combinations of N and T. Their findings are

Table 3.7. *Simple IV*

N	T	Gamma	Mean	MSE	Actual size	Bias	Relative MSE	Relative bias
50	50	0.2	0.2011	0.0043	0.0470	0.0011	4.0396	−0.0440
50	50	0.5	0.5067	0.0125	0.0430	0.0067	8.5183	−0.2044
50	50	0.8	0.8444	0.1008	0.0550	0.0444	41.1601	−0.9572
50	100	0.2	0.1994	0.0020	0.0490	−0.0006	5.3755	0.0439
50	100	0.5	0.5029	0.0056	0.0560	0.0029	11.9963	−0.1707
50	100	0.8	0.8144	0.0466	0.0450	0.0144	85.3687	−0.6860
100	100	0.2	0.2004	0.0011	0.0480	0.0004	4.2987	−0.0368
100	100	0.5	0.5019	0.0029	0.0500	0.0019	8.3923	−0.1138
100	100	0.8	0.8148	0.0206	0.0430	0.0148	42.9504	−0.7158
25	100	0.2	0.2019	0.0041	0.0560	0.0019	7.1513	−0.1455
25	100	0.5	0.5032	0.0121	0.0510	0.0032	18.9890	−0.1883
25	100	0.8	0.8441	0.0974	0.0470	0.0441	157.6636	−2.1020
100	1000	0.2	0.2001	0.0001	0.0480	0.0001		
100	1000	0.5	0.5004	0.0003	0.0520	0.0004		
100	1000	0.8	0.8015	0.0020	0.0560	0.0015		
1000	100	0.2	0.2000	0.0001	0.0460	0.0000		
1000	100	0.5	0.4996	0.0003	0.0490	−0.0004		
1000	100	0.8	0.8004	0.0021	0.0440	0.0004		
1000	1000	0.2	0.2001	0.0000	0.0470	0.0001		
1000	1000	0.5	0.5000	0.0000	0.0460	0.0000		
1000	1000	0.8	0.8001	0.0002	0.0550	0.0001		

Source: Hsiao and Zhang (2015).

Table 3.8. *Fully fledged MLE*

N	T	Gamma	Mean	MSE	Actual Size	Bias	Relative MSE	Relative Bias
50	50	0.2	0.1999	0.0004	0.0470	−0.0001	0.3725	0.0036
50	50	0.5	0.4996	0.0003	0.0550	−0.0004	0.2275	0.0121
50	50	0.8	0.8011	0.0003	0.0530	0.0011	0.1049	−0.0244
50	100	0.2	0.1998	0.0002	0.0490	−0.0002	0.5553	0.0173
50	100	0.5	0.5006	0.0002	0.0570	0.0006	0.3372	−0.0344
50	100	0.8	0.8001	0.0001	0.0490	0.0001	0.1849	−0.0050
100	100	0.2	0.2001	0.0001	0.0480	0.0001	0.3919	−0.0108
100	100	0.5	0.4997	0.0001	0.0540	−0.0003	0.2194	0.0196
100	100	0.8	0.7998	0.0000	0.0460	−0.0002	0.0982	0.0118
25	100	0.2	0.1989	0.0004	0.0490	−0.0011	0.7396	0.0809
25	100	0.5	0.4992	0.0003	0.0440	−0.0008	0.5286	0.0462
25	100	0.8	0.7998	0.0002	0.0520	−0.0002	0.3629	0.0085

Source: Hsiao and Zhang (2015).

reproduced in Tables 3.6–3.8. The relative mean square error (MSE) of IV to GMM in Tables 3.6 and 3.7 show that the GMM is much more efficient than the IV. However, the actual size for a nominal 5% significance level test shows that the IV actual size is close to the nominal size, but the GMM has significant size distortion when N and T are of similar magnitude. The actual size could be 40% when $\gamma = 0.5$, and 80% when $\gamma = 0.8$. On the other hand, the transformed MLE (3.5.7) dominates both the IV and the GMM in terms of efficiency and closeness between the nominal size and actual size.

We summarize the consistent and asymptotic bias properties of the GMM and QMLE in Table 3.9.

Table 3.9. *Asymptotic properties of the likelihood based estimator and methods of moment estimator for dynamic models*

	QMLE		Simple IV	GMM
	y_{i0} fixed	y_{i0} random		
N fixed T large	consistent unbiased	consistent unbiased	consistent unbiased	inconsistent $\rightarrow \infty$
T fixed N large	inconsistent $\rightarrow \infty$	consistent unbiased	consistent unbiased	consistent unbiased
(N, T) $\rightarrow \infty$	consistent biased $O\left(\sqrt{\frac{N}{T^3}}\right)$	consistent unbiased	consistent unbiased	consistent biased $O\left(\sqrt{\frac{T}{N}}\right)$

3.7 ESTIMATION OF DYNAMIC MODELS WITH ARBITRARY SERIAL CORRELATIONS IN THE RESIDUALS

In previous sections we discussed estimation of the dynamic model

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}'\boldsymbol{x}_{it} + \alpha_i^* + u_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{3.7.1}$$

under the assumption that u_{it} are serially uncorrelated, where we now again let $\boldsymbol{x}_{it}$ stand for a $K \times 1$ vector of time-varying exogenous variables. When T is fixed and N tends to infinity, we can relax the restrictions on the serial correlation structure of u_{it} and still obtain efficient estimates of γ and $\boldsymbol{\beta}$.

Taking the first difference of (3.7.1) to eliminate the individual effect α_i^*, and stacking all equations for a single individual, we have a system of $(T-1)$ equations,

$$\begin{aligned} y_{i2} - y_{i1} &= \gamma(y_{i1} - y_{i0}) + \boldsymbol{\beta}'(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}) + (u_{i2} - u_{i1}), \\ y_{i3} - y_{i2} &= \gamma(y_{i2} - y_{i1}) + \boldsymbol{\beta}'(\boldsymbol{x}_{i3} - \boldsymbol{x}_{i2}) + (u_{i3} - u_{i2}), \\ &\vdots \\ y_{iT} - y_{i,T-1} &= \gamma(y_{i,T-1} - y_{i,T-2}) + \boldsymbol{\beta}'(\boldsymbol{x}_{iT} - \boldsymbol{x}_{i,T-1}) \\ &\quad + (u_{iT} - u_{i,T-1}), \quad i = 1, \ldots, N, \end{aligned} \tag{3.7.2}$$

We complete the system (3.7.2) with the identities

$$\begin{aligned} y_{i0} &= E^*(y_{i0} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}) + [y_{i0} - E^*(y_{i0} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})] \\ &= a_0 + \sum_{t=1}^{T} \boldsymbol{\pi}_{0t}' \boldsymbol{x}_{it} + \epsilon_{i0} \end{aligned} \tag{3.7.3}$$

and

$$\begin{aligned} y_{i1} &= E^*(y_{i1} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}) + [y_{i1} - E^*(y_{i1} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})] \\ &= a_1 + \sum_{t=1}^{T} \boldsymbol{\pi}_{1t}' \boldsymbol{x}_{it} + \epsilon_{i1}, \qquad i = 1, \ldots, N. \end{aligned} \tag{3.7.4}$$

where E^* denotes the projection operator. Because (3.7.3) and (3.7.4) are exactly identified equations, we can ignore them and apply the three-stage least squares (3SLS) or generalized 3SLS (see Chapter 4) to the system (3.7.2) only. With regard to the cross-equation constraints in (3.7.2), one can either directly substitute them out or first obtain unknown nonzero coefficients of each equation ignoring the cross-equation linear constraints, then impose the constraints and use the constrained estimation formula (Theil 1971, p. 281; equation 8.5).

Because the system (3.7.2) does not involve the individual effects, α_i^*, nor does the estimation method rely on specific restrictions on the serial-correlation structure of u_{it}, the method is applicable whether α_i^* are treated as fixed or random or as being correlated with $\boldsymbol{x}_{it}$. However, in order to implement simultaneous-equations estimation methods to (3.7.2) without imposing restrictions on the serial-correlation structure of u_{it}, there must exist strictly exogenous variables $\boldsymbol{x}_{it}$ such that

$$E(u_{it} \mid \boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT}) = 0. \tag{3.7.5}$$

Otherwise, the coefficient γ and the serial correlations of u_{it} cannot be disentangled (e.g., Binder, Hsiao, and Pesaran 2005).

3.8 MODELS WITH BOTH INDIVIDUAL- AND TIME-SPECIFIC ADDITIVE EFFECTS

For notational ease and without loss of generality, we illustrate the fundamental issues of dynamic model with both individual- and time-specific additive effects model by restricting $\boldsymbol{\beta} = \boldsymbol{0}$ in (3.1.2); thus, the model becomes

$$y_{it} = \gamma y_{i,t-1} + v_{it}, \tag{3.8.1}$$

$$\begin{aligned} v_{it} &= \alpha_i + \lambda_t + u_{it}, \quad & i &= 1, \dots, N, \\ & & t &= 1, \dots, T, \\ & & & y_{i0} \text{ observable}. \end{aligned} \tag{3.8.2}$$

The panel data estimators discussed in Sections 3.3–3.6 assume no presence of λ_t (i.e., $\lambda_t = 0$ for all t). When λ_t are indeed present, those estimators are not consistent if T is finite when $N \to \infty$. For instance, the consistency of GMM (3.3.17) is based on the assumption that $\frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta v_{it}$ converges to the population moments (3.3.8) of zero. However, if λ_t are also present as in (3.8.2), this condition is likely to be violated. To see this, taking the first difference of (3.8.1) yields

$$\begin{aligned} \Delta y_{it} &= \gamma \Delta y_{i,t-1} + \Delta v_{it} \\ &= \gamma \Delta y_{i,t-1} + \Delta \lambda_t + \Delta u_{it}, \\ & \quad i = 1, \dots, N, \\ & \quad t = 2, \dots, T. \end{aligned} \tag{3.8.3}$$

Although under the assumption λ_t are independently distributed over t with mean zero,

$$E(y_{i,t-j}\Delta v_{it}) = 0 \text{ for } j = 2, \dots, t, \tag{3.8.4}$$

the sample moment, as $N \longrightarrow \infty$,

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta v_{it} &= \frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta\lambda_t \\ &\quad + \frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta u_{it} \end{aligned} \tag{3.8.5}$$

converges to $\bar{y}_{t-j}\Delta\lambda_t$, which in general is not equal to zero, in particular if y_{it} has mean different from zero,[13] where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^N y_{it}$.

To obtain consistent estimators of γ, we need to take explicit account of the presence of λ_t in addition to α_i. When α_i and λ_t are fixed constants, under the assumption that u_{it} are independent normal and fixed y_{i0}, the MLE of the FE model (3.8.1) is equal to:

$$\tilde{\gamma}_{cv} = \frac{\sum_{i=1}^N \sum_{t=1}^T y_{i,t-1}^* y_{it}^*}{\sum_{i=1}^N \sum_{t=1}^T y_{i,t-1}^{*2}}, \tag{3.8.6}$$

where $y_{it}^* = (y_{it} - \bar{y}_i - \bar{y}_t + \bar{y})$, $\bar{y}_i = \frac{1}{T}\sum_{t=1}^T y_{it}$, $\bar{y}_t = \frac{1}{N}\sum_{i=1}^N y_{it}$, $\bar{y} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T y_{it}$, and similarly for $\bar{y}_{t-1}$, $\bar{y}_{i,-1}$, $\bar{\boldsymbol{x}}_i$, $\bar{\boldsymbol{x}}_t$, $\boldsymbol{x}_{it}^*$, v_{it}^*, $\bar{v}_i$, $\bar{v}_t$, and $\bar{v}$. The FE MLE of γ is also called the covariance estimator because it is equivalent to first applying covariance transformation to sweep out α_i and λ_t,

$$y_{it}^* = \gamma y_{i,t-1}^* + v_{it}^*, \tag{3.8.7}$$

then apply the least squares estimator of (3.8.7).

The probability limit of $\tilde{\gamma}_{cv}$ is identical to the case where $\lambda_t = 0$ for all t (Equation 3.2.14; Hahn and Moon 2006; Hsiao and Tahmiscioglu 2008). The bias is to the order of $(1/T)$, and it is identical independent of whether α_i and λ_t are fixed or random and are identical whether λ_t are present or not (e.g. Hahn and Moon 2006; Hsiao and Tahmiscioglu 2008). When $T \longrightarrow \infty$, the MLE of the FE model is consistent. However, if N also goes to infinity and $\lim\left(\frac{N}{T}\right) = c > 0$, Hahn and Moon (2006) have shown that $\sqrt{NT}(\tilde{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean $-\sqrt{c}(1+\gamma)$ and variance $1-\gamma^2$. In other words, the usual t-statistic based on γ_{cv} could be subject to severe size distortion unless T increases faster than N.

If α_i and λ_t are random and satisfy (3.4.3), because $Ey_{i0}v_{it} \neq 0$, we either have to write (3.8.1) conditional on y_{i0} or to complete the system (3.8.1) by deriving the marginal distribution of y_{i0}. By continuous substitutions, we have

$$\begin{aligned} y_{i0} &= \frac{1-\gamma^m}{1-\gamma}\alpha_i + \sum_{j=0}^{m-1}\lambda_{i,-j}\gamma^j + \sum_{j=0}^{m-1}\epsilon_{i,-j}\gamma^j \\ &= v_{i0}, \end{aligned} \tag{3.8.8}$$

assuming the process started at period $-m$.

Under (3.4.3), $Ey_{i0} = Ev_{i0} = 0$, var $(y_{i0}) = \sigma_0^2$, $E(v_{i0}v_{it}) = \frac{1-\gamma^m}{1-\gamma}\sigma_\alpha^2 = c^*$, $Ev_{it}v_{jt} = d^*$. Stacking the $T+1$ time series observations for the ith individual into a vector, $\boldsymbol{y}_i = (y_{i0}, \ldots, y_{i,T})'$ and $\boldsymbol{y}_{i,-1} = (0, y_{i1}, \ldots, y_{i,T-1})'$, $\boldsymbol{v}_i = (v_{i0}, \ldots, v_{iT})'$. Let $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')'$, $\boldsymbol{y}_{-1} = (\boldsymbol{y}_{1,-1}', \ldots, \boldsymbol{y}_{N,-1}')$, $\boldsymbol{v} = (\boldsymbol{v}_1', \ldots, \boldsymbol{v}_N')'$, then

$$\boldsymbol{y} = \boldsymbol{y}_{-1}\gamma + \boldsymbol{v}, \tag{3.8.9}$$

$$\begin{aligned} E\boldsymbol{v} &= \boldsymbol{0}, \\ E\boldsymbol{v}\boldsymbol{v}' &= \sigma_u^2 I_N \otimes \begin{pmatrix} \omega & \boldsymbol{0}' \\ \boldsymbol{0} & I_T \end{pmatrix} + \sigma_\alpha^2 I_N \otimes \begin{pmatrix} 0 & c^*\boldsymbol{e}_T' \\ c^*\boldsymbol{e}_T & \boldsymbol{e}_T\boldsymbol{e}_T' \end{pmatrix} \\ &\quad + \sigma_\lambda^2 \boldsymbol{e}_N\boldsymbol{e}_N' \otimes \begin{pmatrix} d^* & \boldsymbol{0}' \\ \boldsymbol{0} & I_T \end{pmatrix}, \end{aligned} \tag{3.8.10}$$

where $\otimes$ denotes the Kronecker product, and ω denotes the variance of v_{i0} divided by σ_u^2. The system (3.8.9) has a fixed number of unknowns $(\gamma, \sigma_u^2, \sigma_\alpha^2, \sigma_\lambda^2, \sigma_0^2, c^*, d^*)$ as N

[13] For instance, if y_{it} is also a function of exogenous variables as in (3.1.2), where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^N y_{it}$.

and T increase. Therefore, the MLE (or quasi-MLE or GLS of (3.8.9)) is consistent and asymptotically normally distributed.

When α_i and λ_t are fixed constants, we note that first differencing only eliminates α_i from the specification. The time-specific effects, $\Delta\lambda_t$, remain at (3.8.3). To further eliminate $\Delta\lambda_t$, we note that the cross-sectional mean $\Delta y_t = \frac{1}{N}\sum_{i=1}^{N}\Delta y_{it}$ is equal to

$$\Delta y_t = \gamma \Delta y_{t-1} + \Delta\lambda_t + \Delta u_t, \tag{3.8.11}$$

where $\Delta u_t = \frac{1}{N}\sum_{i=1}^{N}\Delta u_{it}$. Taking deviation of (3.8.3) from (3.8.11) yields

$$\Delta y^*_{it} = \gamma \Delta y^*_{i,t-1} + \Delta u^*_{it}, \qquad \begin{array}{l} i = 1, \dots, N, \\ t = 2, \dots, T, \end{array} \tag{3.8.12}$$

where $\Delta y^*_{it} = (\Delta y_{it} - \Delta y_t)$ and $\Delta u^*_{it} = (\Delta u_{it} - \Delta u_t)$. The system (3.8.12) no longer involves α_i and λ_t.

Since

$$E[y_{i,t-j}\Delta u^*_{it}] = 0 \text{ for } \begin{array}{l} j = 2, \dots, t, \\ t = 2, \dots, T. \end{array} \tag{3.8.13}$$

the $\frac{1}{2}T(T-1)$ orthogonality conditions can be represented as

$$E(W_i \Delta\tilde{\boldsymbol{u}}^*_i) = \mathbf{0}, \tag{3.8.14}$$

where $\Delta\tilde{\boldsymbol{u}}^*_i = (\Delta u^*_{i2}, \dots, \Delta u^*_{iT})'$,

$$W_i = \begin{pmatrix} \boldsymbol{q}_{i2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{q}_{i3} & & \\ \cdot & \cdot & \ddots & \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & & \boldsymbol{q}_{iT} \end{pmatrix}, \quad i = 1, \dots, N,$$

and $\boldsymbol{q}_{it} = (y_{i0}, y_{i1}, \dots, y_{i,t-2})', t = 2, 3, \dots, T$. Following Arellano and Bond (1991), we can propose a generalized method of moments (GMM) estimator,[14]

$$\begin{aligned} \tilde{\gamma}_{GMM} = & \left\{\left[\frac{1}{N}\sum_{i=1}^{N}\Delta\tilde{\boldsymbol{y}}^{*\prime}_{i,-1}W_i'\right]\hat{\Psi}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}W_i\Delta\tilde{\boldsymbol{y}}^*_{i,-1}\right]\right\}^{-1} \\ & \left\{\left[\frac{1}{N}\sum_{i=1}^{N}\Delta\tilde{\boldsymbol{y}}^{*\prime}_{i,-1}W_i'\right]\hat{\Psi}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}W_i\Delta\tilde{\boldsymbol{y}}^*_{i}\right]\right\}, \end{aligned} \tag{3.8.15}$$

where $\Delta\tilde{\boldsymbol{y}}^*_i = (\Delta y^*_{i2}, \dots, \Delta y^*_{iT})', \Delta\tilde{\boldsymbol{y}}^*_{i-1} = (\Delta y^*_{i1}, \dots, \Delta y^*_{i,T-1})$, and

$$\hat{\Psi} = \frac{1}{N^2}\left[\sum_{i=1}^{N}W_i\hat{\tilde{\boldsymbol{u}}}^*_i\right]\left[\sum_{i=1}^{N}W_i\hat{\tilde{\boldsymbol{u}}}^*_i\right]' \tag{3.8.16}$$

and $\Delta\hat{\tilde{\boldsymbol{u}}}^*_i = \Delta\tilde{\boldsymbol{y}}^*_i - \Delta\tilde{\boldsymbol{y}}^*_{i,-1}\tilde{\gamma}$, and $\tilde{\gamma}$ denotes some initial consistent estimator of γ, say, a simple instrumental variable estimator.

[14] For ease of exposition, we have considered only the GMM that makes use of orthogonality conditions. For additional moments conditions such as homoscedasticity or initial observations, see, for example, Ahn and Schmidt (1995) and Blundell and Bond (1998).

The asymptotic covariance matrix of $\tilde{\gamma}_{GMM}$ can be approximated by

$$\text{asy. cov}\ (\tilde{\gamma}_{GMM}) = \left\{ \left[\sum_{i=1}^{N} \Delta \tilde{\boldsymbol{y}}_{i,-1}^{*'} W_i' \right] \hat{\Psi}^{-1} \left[\sum_{i=1}^{N} W_i \Delta \tilde{\boldsymbol{y}}_{i,-1}^{*} \right] \right\}^{-1}. \tag{3.8.17}$$

To implement the likelihood approach, we need to complete the system (3.8.11) by deriving the marginal distribution of Δy_{i1}^* through continuous substitution,

$$\begin{aligned} \Delta y_{i1}^* &= \sum_{j=0}^{m-1} \Delta u_{i,1-j}^* \gamma^j \\ &= \Delta u_{i1}^*, \qquad i = 1, \ldots, N. \end{aligned} \tag{3.8.18}$$

Let $\Delta \boldsymbol{y}_i^* = (\Delta y_{i1}^*, \ldots, \Delta y_{iT}^*)', \Delta \boldsymbol{y}_{i,-1}^* = (0, \Delta y_{i1}^*, \ldots, \Delta y_{i,T-1}^*)'$, $\Delta \boldsymbol{u}_i^* = (\Delta u_{i1}^*, \ldots, \Delta u_{iT}^*)'$, the system

$$\Delta \boldsymbol{y}_i^* = \Delta \boldsymbol{y}_{i,-1}^* \gamma + \Delta \boldsymbol{u}_i^*, \tag{3.8.19}$$

does not involve α_i and λ_t. The MLE conditional on $\omega = \frac{\text{Var}\ (\Delta y_{i1}^*)}{\sigma_u^2}$ is identical to the GLS

$$\hat{\gamma}_{GLS} = \left[\sum_{i=1}^{N} \Delta \boldsymbol{y}_{i,-1}^{*'} \tilde{A}^{-1} \Delta \boldsymbol{y}_{i,-1}^* \right]^{-1} \left[\sum_{i=1}^{N} \Delta \boldsymbol{y}_{i,-1}^{*'} \tilde{A}^{-1} \Delta \boldsymbol{y}_i^* \right]. \tag{3.8.20}$$

where

$$\tilde{A} = \begin{bmatrix} \omega & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & . & . \\ 0 & -1 & 2 & -1 & \ldots & . & . \\ . & . & . & . & . & 2 & -1 \\ 0 & . & . & . & & -1 & 2 \end{bmatrix}, \tag{3.8.21}$$

The GLS is consistent and asymptotically normally distributed, with the covariance matrix equal to

$$\text{Var}\ (\hat{\gamma}_{GLS}) = \sigma_u^2 \left[\sum_{i=1}^{N} \Delta \boldsymbol{y}_{i,-1}^{*'} \tilde{A}^{-1} \Delta \boldsymbol{y}_{i,-1}^* \right]^{-1}. \tag{3.8.22}$$

Remark 3.8.1 The GMM or GLS with $\Delta \boldsymbol{\lambda}$ present is basically of the same form as the GMM or GLS without the time-specific effects (i.e. $\Delta \boldsymbol{\lambda} = \boldsymbol{0}$) (Hsiao, Pesaran, and Tahmiscioglu 2002), (Equation 3.5.14). However, there is an important difference between the two. The estimator (3.8.12) uses $\Delta y_{i,t-1}^*$ as the regressor for the equation Δy_{it}^* (Equation 3.8.19) and does not use $\Delta y_{i,t-1}$ as the regressor for the equation Δy_{it} (Equation 3.8.3). If there are indeed common shocks that affect all the cross-sectional units, then the estimator (3.5.12) is inconsistent while (3.8.20) is consistent (for detail, see Hsiao and Tahmiscioglu 2008). Note also that even though when there are no time-specific effects, (3.8.20) remains consistent, although it will not be as efficient as (3.5.11).

Remark 3.8.2 The estimator (3.8.20) and the estimator (3.8.15) remain consistent and asymptotically normally distributed when the effects are random because the transformation (3.8.12) effectively removes the individual- and time-specific effects from the specification. However, if the effects are indeed random, then the MLE or GLS of (3.8.7) is more efficient.

Remark 3.8.3 The GLS (3.8.20) assumes known ω. If ω is unknown, one may substitute it by a consistent estimator $\hat{\omega}$, then apply the feasible GLS. However, there is an important difference between the GLS and the feasible GLS in a dynamic setting. The feasible GLS is not asymptotically equivalent to the GLS when T is finite. However, if both N and $T \to \infty$ and $\lim\left(\frac{N}{T}\right) = c > 0$, then the FGLS will be asymptotically equivalent to the GLS. (Hsiao and Tahmiscioglu 2008).

Remark 3.8.4 The MLE or GLS of (3.8.20) can also be derived by treating $\Delta\lambda_t$ as fixed parameters in the system (3.8.3). Through continuous substitution, we have

$$\Delta y_{i1} = \lambda_1^* + \Delta\tilde{u}_{i1}, \tag{3.8.22}$$

where $\lambda_1^* = \sum_{j=0}^m \gamma^j \Delta\lambda_{1-j}$ and $\Delta\tilde{u}_{i1} = \sum_{j=0}^m \gamma^j \Delta u_{i,1-j}$. Let $\Delta\boldsymbol{y}_i' = (\Delta y_{i1}, \ldots, \Delta y_{iT}), \Delta\boldsymbol{y}_{i,-1}' = (0, \Delta y_{i1}, \ldots, \Delta y_{i,T-1})$, $\Delta\boldsymbol{u}_i' = (\Delta\tilde{u}_{i1}, \ldots, \Delta u_{iT})$, and $\Delta\boldsymbol{\lambda}' = (\lambda_1^*, \Delta\lambda_2, \ldots, \Delta\lambda_T)$, we may write

$$\begin{aligned}\underset{NT\times 1}{\Delta\boldsymbol{y}} &= \begin{pmatrix}\Delta\boldsymbol{y}_1\\ \vdots\\ \Delta\boldsymbol{y}_N\end{pmatrix} = \begin{pmatrix}\Delta\boldsymbol{y}_{1,-1}\\ \vdots\\ \Delta\boldsymbol{y}_{N,-1}\end{pmatrix}\gamma + (\boldsymbol{e}_N \otimes I_T)\Delta\boldsymbol{\lambda} + \begin{pmatrix}\Delta\boldsymbol{u}_1\\ \vdots\\ \Delta\boldsymbol{u}_N\end{pmatrix}\\ &= \Delta\boldsymbol{y}_{-1}\gamma + (\boldsymbol{e}_N \otimes I_T)\Delta\boldsymbol{\lambda} + \Delta\boldsymbol{u},\end{aligned} \tag{3.8.23}$$

If u_{it} is i.i.d. normal with mean 0 and variance σ_u^2, then $\Delta\boldsymbol{u}_i'$ is independently normally distributed across i with mean $\mathbf{0}$ and covariance matrix $\sigma_\epsilon^2\tilde{A}$, and $\omega = \frac{\text{Var}\,(\Delta\tilde{u}_{i1})}{\sigma_u^2}$.

The log-likelihood function of $\Delta\boldsymbol{y}$ takes the form

$$\begin{aligned}\log L = &-\frac{NT}{2}\log\sigma_u^2 - \frac{N}{2}\log\ |\tilde{A}|-\frac{1}{2\sigma_u^2}[\Delta\boldsymbol{y} - \Delta\boldsymbol{y}_{-1}\gamma - (e_N \otimes I_T)\Delta\boldsymbol{\lambda}]'\\ &(I_N \otimes \tilde{A}^{-1})[\Delta\boldsymbol{y} - \Delta\boldsymbol{y}_{-1}\gamma - (\boldsymbol{e}_N \otimes I_T)\Delta\boldsymbol{\lambda}].\end{aligned} \tag{3.8.24}$$

Taking partial derivative of (3.8.24) with respect to $\Delta\boldsymbol{\lambda}$ and solving for $\Delta\boldsymbol{\lambda}$ yields

$$\Delta\hat{\boldsymbol{\lambda}} = (N^{-1}\boldsymbol{e}_N' \otimes I_T)(\Delta\boldsymbol{y} - \Delta\boldsymbol{y}_{-1}\gamma). \tag{3.8.25}$$

Substituting (3.8.25) into (3.8.24) yields the concentrated log likelihood function.

$$\begin{aligned}\log L_c = &-\frac{NT}{2}\log\sigma_u^2 - \frac{N}{2}\log\ |\tilde{A}|\\ &-\frac{1}{2\sigma_u^2}(\Delta\boldsymbol{y}^* - \Delta\boldsymbol{y}_{-1}^*\gamma)'(I_N \otimes \tilde{A}^{-1})(\Delta\boldsymbol{y}^* - \Delta y_{-1}^*\gamma).\end{aligned} \tag{3.8.26}$$

Maximizing (3.8.26) conditional on ω yields (3.8.19).

Remark 3.8.5 When γ approaches to 1 and σ_α^2 is large relative to σ_u^2, the GMM estimator of the form (3.3.17) suffers from the weak instrumental variables issues and performs poorly (e.g. Binder, Hsiao, and Pesaran 2005). On the other hand, the performance of the likelihood or GLS estimator is not affected by these problems.

Remark 3.8.6 Hahn and Moon (2006) propose a bias corrected estimator as

$$\tilde{\gamma}_b = \tilde{\gamma}_{cv} + \frac{1}{T}(1 + \tilde{\gamma}_{cv}). \tag{3.8.27}$$

They show that when $N/T \to c$, as both N and T tend to infinity where $0 < c < \infty$,

$$\sqrt{NT}(\tilde{\gamma}_b - \gamma) \Longrightarrow N(0, 1 - \gamma^2). \tag{3.8.28}$$

The limited Monte Carlo studies conducted by Hsiao and Tahmiscioglu (2008) to investigate the finite sample properties of the feasible GLS (FGLS), GMM, bias corrected (BC) estimator of Hahn and Moon (2006) have shown that in terms of bias and root mean square errors, FGLS dominates. However, the BC rapidly improves as T increases. In terms of the closeness of actual size to the nominal size, again FGLS dominates and rapidly approaches the nominal size when N or T increases. The GMM with T fixed and N large also has actual sizes close to nominal sizes except for the cases when γ is close to unity (here $\gamma = 0.8$). The BC has significant size distortion, presumably because of the use of the asymptotic covariance matrix which is significantly downward biased in the finite sample.indexPesaran, M.H.

Remark 3.8.7 Hsiao and Tahmiscioglu (2008) also compare the FGLS and GMM with and without the correction of time-specific effects in the presence of both individual- and time-specific effects or in the presence of individual-specific effects only. It is interesting to note that when both individual- and time-specific effects are present, the biases and root mean squares errors are large for estimators assuming no time-specific effects; however, their biases decrease as T increases when the time-specific effects are independent of regressors. On the other hand, even in the case of no time-specific effects, there is hardly any efficiency loss for the FGLS or GMM that makes the correction of presumed presence of time-specific effects. Therefore, if an investigator is not sure if the assumption of cross-sectional independence is valid or not, it might be advisable to use estimators that take into account both individual- and time-specific effects when T is finite.

APPENDIX 3A DERIVATION OF THE ASYMPTOTIC COVARIANCE MATRIX OF THE FEASIBLE MDE

The estimation error of $\hat{\boldsymbol{\psi}}_{MDE}$ is equal to

$$\sqrt{N}(\hat{\boldsymbol{\psi}}_{MDE} - \boldsymbol{\psi}) = \left(\frac{1}{N}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} H_i\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} \Delta \boldsymbol{u}_i^*\right). \tag{3A.1}$$

When $N \longrightarrow \infty$

$$\frac{1}{N}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} H_i \longrightarrow \frac{1}{N}\sum_{i=1}^{N} H_i' \tilde{\Omega}^{*-1} H_i \tag{3A.2}$$

but

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} \Delta \boldsymbol{u}_i^* &\approx \frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i' \tilde{\Omega}^{*-1} \Delta \boldsymbol{u}_i^* \\ &\quad + \left[\frac{1}{N}\sum_{i=1}^{N} H_i' \left(\frac{\partial}{\partial h}\tilde{\Omega}^{*-1}\right) \Delta \boldsymbol{u}_i^*\right] \cdot \sqrt{N}(\hat{h} - h),\end{aligned} \tag{3A.3}$$

where the right-hand side follows from taking Taylor series expansion of $\hat{\tilde{\Omega}}^{*-1}$ around $\tilde{\Omega}^{*-1}$. By (3.5.9),

$$\frac{\partial}{\partial h}\tilde{\Omega}^{*-1} = \frac{-T}{[1+T(h-1)]^2}\tilde{\Omega}^{*-1}$$

$$+\frac{1}{1+T(h-1)]}\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & T-1 & \dots & 2 & 1 \\ \ddots & \ddots & & \ddots & \ddots \\ . & 2 & \dots & 2(T-2) & T-2 \\ 0 & 1 & \dots & T-2 & T-1 \end{bmatrix}. \quad (3A.4)$$

We have

$$\frac{1}{N}\sum_{i=1}^{N} H_i'\tilde{\Omega}^{*-1}\Delta \boldsymbol{u}_i^* \longrightarrow \mathbf{0},$$

$$\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} 1 & \Delta \boldsymbol{x}_i' & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \Delta X_i \end{bmatrix}' \cdot \frac{\partial}{\partial h}\tilde{\Omega}^{*-1}\Delta \boldsymbol{u}_i^* \longrightarrow \mathbf{0},$$

$$\frac{1}{N}\sum_{i=1}^{N}\Delta \boldsymbol{y}_{i,-1}'\begin{bmatrix} T-1 & \dots & 1 \\ & \ddots & \\ 2 & & T-2 \\ 1 & & T-1 \end{bmatrix}\Delta \boldsymbol{u}_i^* \longrightarrow [\gamma^{T-2}+2\gamma^{T-3}+\dots+(T-1)]\sigma_u^2.$$

Since $\operatorname{plim}\hat{\sigma}_u^2 = \sigma_u^2$, and

$$\sqrt{N}(\hat{h}-h) = \sqrt{N}\left[\frac{\hat{\sigma}_{v*}^2}{\hat{\sigma}_u^2} - \frac{\sigma_{v*}^2}{\sigma_u^2}\right] = \sqrt{N}\frac{\sigma_u^2(\hat{\sigma}_{v*}^2-\sigma_{v*}^2)-\sigma_{v*}^2(\hat{\sigma}_u^2-\sigma_u^2)}{\hat{\sigma}_u^2\sigma_u^2},$$

it follows that the limiting distribution of the feasible MDE converges to

$$\sqrt{N}(\hat{\boldsymbol{\psi}}_{MDE}-\boldsymbol{\psi}) \longrightarrow \left(\frac{1}{N}\sum_{i=1}^{N}H_i'\Omega^{*-1}H_i\right)^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}H_i'\Omega^{*-1}\Delta\boldsymbol{u}_i^*\right.$$

$$\left. -\begin{bmatrix} 0 \\ \mathbf{0} \\ 1 \\ \mathbf{0} \end{bmatrix}\frac{[\gamma^{T-2}+2\gamma^{T-3}+\dots+(T-1)]}{[1+T(h-1)]\sigma_u^2}[\sigma_u^2\cdot\sqrt{N}(\hat{\sigma}_{v*}^2-\sigma_{v*}^2)-\sigma_{v*}^2\cdot\sqrt{N}(\hat{\sigma}_u^2-\sigma_u^2)]\right\}, \quad (3A.5)$$

with the asymptotic covariance matrix equal to (3.5.17).

APPENDIX 3B LARGE N AND T ASYMPTOTICS

In cases when N is fixed and T is large, or T is fixed and N is large, standard one-dimension asymptotic techniques can be applied. However, in some panel data sets, the orders of magnitude of the cross-section and time series are similar, for instance the Penn-World tables. These large N, large T panels call for the use of large N, T asymptotics rather than just large N asymptotics. Moreover, when T is large, there is a need to consider serial correlations more generally, including both short memory and persistent components. In some panel data sets like the Penn-World Table, the time series components also have strongly evident nonstationarity. It turns out that panel data in this case can sometimes offer additional insights to the data generating process than a single time series or cross-section data.

In regressions with large N, large T panels, most of the interesting test statistics and estimators inevitably depend on the treatment of the two indexes, N and T, which tend to infinity together. Several approaches are possible:

(a) *Sequential Limits*. A sequential approach is to fix one index, say N, and allow the other, say T, to pass to infinity, giving an intermediate limit. Then, by letting N pass to infinity subsequently, a sequential limit theory is obtained.

(b) *Diagonal Path Limits*. This approach is to allow the two indexes, N and T, to pass to infinity along a specific diagonal path in the two-dimension array, say $T = T(N)$ such as $\frac{N}{T} \longrightarrow c \neq 0 < \infty$ as the index $N \to \infty$. This approach simplifies the asymptotic theory of a double-indexed process into a single-indexed process.

(c) *Joint Limits*. A joint limit theory allows both indexes, N and T, to pass to infinity simultaneously without placing specific diagonal path restrictions on the divergence, although it may still be necessary to exercise some control over the rate of expansion of the two indexes in order to get definitive results.

A double index process in this monograph typically takes the form,

$$X_{N,T} = \frac{1}{k_N} \sum_{i=1}^{N} Y_{i,T}, \tag{3B.1}$$

where k_N is an N-indexed standardizing factor, $Y_{i,T}$ are independent m-component random vectors across i for all T that is integrable and has the form

$$Y_{i,T} = \frac{1}{d_T} \sum_{t=1}^{T} f(Z_{i,t}), \tag{3B.2}$$

for h-component independently, identically distributed random vectors, $Z_{i,t}$ with finite fourth moments, $f(\cdot)$ is a continuous functional from R^h to R^m, and d_T is a T-indexed standardizing factor. Sequential limit theory is easy to derive and generally leads to quick results. However, it can also give asymptotic results that are misleading in cases where both indexes pass to infinity simultaneously. A joint limit will give a more robust result than either a sequential limit or diagonal path limit, but will also be substantially more difficult to derive and will usually apply only under stronger conditions, such as the existence of higher moments, that will allow for uniformity in the convergence arguments. Phillips and Moon (1999) give he conditions for sequential convergence to imply joint convergence as:

(i) $X_{N,T}$ converges to X_N, for all N, in probability as $T \longrightarrow \infty$ uniformly and X_N converges to X in probability as $N \longrightarrow \infty$. Then $X_{N,T}$ converges to X in probability jointly if and only if

$$\limsup_{T\to\infty} \sup_{N} P\{\| X_{N,T} - X_N \| > \epsilon\} = 0 \text{ for every } \epsilon > 0. \tag{3B.3}$$

(ii) $X_{N,T}$ converges to X_N in distribution for any fixed N as $T \longrightarrow \infty$ and X_N converges to X in distribution as $N \longrightarrow \infty$. Then, $X_{N,T}$ converges to distribution jointly if and only if

$$\limsup_{N,T} | E(f(X_{N,T}) - E(f(X)) | = 0, \tag{3B.4}$$

for all bounded, continuous, real function on R^m.

Suppose $Y_{i,T}$ converges to Y_i in distribution as $T \longrightarrow \infty$. Phillips and Moon (1999) have given the following set of sufficient conditions that ensures the sequential limits are equivalent to joint limits:

(i) $\limsup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_{i,T} \parallel < \infty$;

(ii) $\limsup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} \parallel EY_{i,T} - EY_i \parallel = 0$;

(iii) $\limsup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_{i,T} \parallel 1\left\{\parallel Y_{i,T} \parallel > N\epsilon\right\} = 0 \; \forall \epsilon > 0$;

(iv) $\limsup_{N} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_i \parallel 1\{\parallel Y_i \parallel > N\epsilon\} = 0 \; \forall \epsilon > 0$,

where $\parallel A \parallel$ is the Euclidean norm $(tr(A'A))^{\frac{1}{2}}$ and $1\{\cdot\}$ is an indicator function.

In general, if an estimator is of the form (3B.1) and $y_{i,T}$ is integrable for all T, and if this estimator is consistent in the fixed T, large N case, it will remain consistent if both N and T tend to infinity irrespective of how they tend to infinity. Moreover, even in the case that an estimator is inconsistent for fixed T and large N case, say, the covariance estimator for the fixed-effects dynamic model (3.2.1), it can become consistent if T also tends to infinity. The probability limit of an estimator, in general, is identical independent of the sequence of limits one takes. However, the properly scaled limiting distribution may be different depending on how the two indexes, N and T, tend to infinity. Consider the double sequence

$$X_{N,T} = \frac{1}{N} \sum_{i=1}^{N} Y_{i,T}. \tag{3B.5}$$

Suppose $Y_{i,T}$ is independently, identically distributed across i for each T with $E(Y_{i,T}) = \frac{1}{\sqrt{T}}b$ and $\text{var}\,(Y_{i,T}) \leq B < \infty$. For fixed N, $X_{N,T}$ converges to X_N in probability as $T \to \infty$ where $E(X_N) = 0$. Because $\text{var}\,(Y_{i,T})$ is bounded, by a law of large numbers, X_N converges to zero in probability as $N \to \infty$. Since (3B.5) satisfies (3B.3), the sequential limit is equal to the joint limit as $N, T \to \infty$. This can be clearly seen by writing

$$\begin{aligned} X_{N,T} &= \frac{1}{N} \sum_{i=1}^{N} [Y_{i,T} - E(Y_{i,T})] + \frac{1}{N} \sum_{i=1}^{N} E(Y_{i,T}) \\ &= \frac{1}{N} \sum_{i=1}^{N} [Y_{i,T} - E(Y_{i,T})] + \frac{b}{\sqrt{T}}. \end{aligned} \tag{3B.6}$$

Since the variance of $Y_{i,T}$ is uniformly bounded by B,

$$E(X_{N,T}^2) = \frac{1}{N}\,\text{var}\,(Y_{i,T}) + \frac{b^2}{T} \longrightarrow 0 \tag{3B.7}$$

as $N, T \longrightarrow \infty$. Equation (3B.7) implies that $X_{N,T}$ converges to zero jointly as $N, T \longrightarrow \infty$.

Alternatively, if we let

$$X_{N,T} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Y_{i,T}, \tag{3B.8}$$

the sequential limit would imply $X_{N,T}$ is asymptotically normally distributed with $E(X_{N,T}) = 0$. However, under (3B.8), the condition (3B.3) is violated. The joint limit would have

$$EX_{N,T} = \frac{\sqrt{c}}{N} \sum_{i=1}^{N} b \longrightarrow \sqrt{c}b, \tag{3B.9}$$

along some diagonal limit, $\frac{N}{T} \longrightarrow c \neq 0$ as $N \longrightarrow \infty$. In this case, T has to increase faster than N to make the $\sqrt{N}$-standardized sum of the biases small, say $\frac{N}{T} \longrightarrow 0$, to prevent the bias from having a dominating asymptotic effect on the standardized quantity (e.g. Alvarez and Arellano 2003; Hahn and Kuersteiner 2000).

If the time series component is an integrated process (nonstationary), panel regressions in which both T and N are large can behave very differently from time series regressions. For instance, consider the linear regression model

$$y = E(y \mid x) + v = \beta x + v. \tag{3B.10}$$

If v_t is stationary (or an $I(0)$ process), the least squares estimator of $\beta, \hat{\beta}$, gives the same interpretation irrespective of whether y and x are stationary or integrated of order 1, $I(1)$ (i.e., the first difference of a variable is stationary or $I(0)$). However, if both y_{it} and x_{it} are $I(1)$ but not cointegrated, then v_{it} is also $I(1)$. It is shown by Phillips (1986) that a time series regression coefficient $\hat{\beta}_i$ has a nondegenerating distribution as $T \longrightarrow \infty$. The estimate $\hat{\beta}_i$ is spurious in the sense that the time series regression of y_{it} on x_{it} does not identify any fixed long-run relation between y_{it} and x_{it}. On the other hand, with panel data, such regressions are not spurious in the sense that they do, in fact, identify a long-run average relation between y_{it} and x_{it}. To see this, consider the case that y and x are bivariate normally distributed as $N(\mathbf{0}, \sum)$ with

$$\sum = \begin{pmatrix} \sum_{yy} & \sum_{yx} \\ \sum_{xy} & \sum_{xx} \end{pmatrix}, \tag{3B.11}$$

then plim $\hat{\boldsymbol{\beta}} = \sum_{yx} \sum_{xx}^{-1}$. In a unit root framework of the form

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{yt} \\ u_{xt} \end{pmatrix}, \tag{3B.12}$$

where the errors $\boldsymbol{u}_t = (u_{yt}, u_{xt})'$ are stationary, then the panel regression under the assumption of cross-sectional independence yields

$$\text{plim } \hat{\beta} = \Omega_{yx}\Omega_{xx}^{-1}, \tag{3B.13}$$

which can be viewed as the long-run average relation between y and x, where Ω_{yx}, Ω_{xx} denote the long-run covariance between u_{yt} and u_{xt}, and the long-run variance of $\mathbf{u}_t$ is defined by

$$\begin{aligned} \Omega &= \lim_{T\to\infty} E\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{u}_t\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{u}_t'\right)\right] \\ &= \sum_{\ell=-\infty}^{\infty} E(\boldsymbol{u}_0\boldsymbol{u}_\ell') = \begin{pmatrix} \Omega_{yy} & \Omega_{yx} \\ \Omega_{xy} & \Omega_{xx} \end{pmatrix}. \end{aligned} \tag{3B.14}$$

When cross-sectional units have heterogeneous long-run covariance matrices Ω_i for $(y_{it}, x_{it}), i = 1, \ldots, N$ with $E\Omega_i = \Omega$, Phillips and Moon (1999, 2000) extend this concept of a long-run average relation among cross-sectional units further,

$$\beta = E(\Omega_{yx,i})(E\Omega_{xx,i})^{-1} = \Omega_{yx}\Omega_{xx}^{-1}, \tag{3B.15}$$

and show that the least squares estimator converges to (3B.15) as $N, T \to \infty$.

This generalized concept of average relation between cross-sectional units covers both the cointegrated case (Engle and Granger 1987) in which β is a cointegrating coefficient in the sense that the particular linear combination $y_t - \beta x_t$ is stationary, and the correlated but noncointegrated case, which is not available for a single time series. To see this point more clearly, suppose that the two nonstationary time series variables have the following relation:

$$\begin{aligned} y_t &= f_t + w_t, \\ x_t &= f_t, \end{aligned} \tag{3B.16}$$

with

$$\begin{pmatrix} w_t \\ f_t \end{pmatrix} = \begin{pmatrix} w_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} u_{wt} \\ u_{ft} \end{pmatrix}, \tag{3B.17}$$

where u_{ws} is independent of u_{ft} for all t and s and has nonzero long-run variance. Then f_t is a nonstationary common factor variable for y and x, and u_w is a nonstationary idiosyncratic factor variable. Since u_{wt} is nonstationary over time, it is apparent that there is no cointegrating relation between y_t and x_t. However, since the two nonstationary variables y_t and x_t share a common contributory nonstationary source in u_{ft}, we may still expect to find evidence of long-run correlation between y_t and x_t, and this is what is measured by the regression coefficient β in (3B.13).

Phillips and Moon (1999, 2000) show that for large N and T panels, the regression coefficient β converge so defined long-run average relation. However, if N is fixed, then as $T \to \infty$, the least squares estimator of β is a nondegenerate random variable that is a functional of Brownian motion that does not converge to β (Phillips 1986). In other words, with a single time series or a fixed number of time series, the regression coefficient β will not converge to the long-run average relation defined by (3.B.13) if only $T \to \infty$.

Therefore, if we define spurious regression as yielding nonzero β for the two independent variables, then contrary to the case of time series regression involving two linearly independent I(1) variables (Phillips 1986), the issue of spurious regression will not arise for the panel estimates of $N \to \infty$ (e.g., McCoskey and Kao 1998).

When data on cross-sectional dimension are correlated, the limit theorems become complicated. When there are strong correlations on cross-sectional dimensions, it is unlikely that the law of large numbers or central limit theory will hold if cross-sectional correlations are strong. They can hold only when cross-sectional dependence is weak in the sense of time series mixing condition in the cross-sectional dimension (e.g., Conley 1999; Pesaran and Tosetti 2011).