CHAPTER 4

Dynamic Models with Variable Intercepts

4.1 INTRODUCTION

In Chapter 3 we discussed the implications of treating the specific effects as fixed or random and the associated estimation methods for the linear static model

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i^* + \lambda_t + u_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{4.1.1}$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector of explanatory variables, 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 let u_{it} represent the effects of those unobserved variables that vary over i and t. Very often we also wish to use panel data to estimate behavioral relationships that are dynamic in character, namely, models containing lagged dependent variables such as[1]

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i^* + \lambda_t + u_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{4.1.2}$$

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. It turns out that in this circumstance the choice between a fixed-effects formulation and a random-effects formulation has implications for estimation that are of a different nature than those associated with the static model.

Roughly speaking, two issues have been raised in the literature regarding whether the effects, α_i and λ_t, should be treated as random or as fixed for a linear static model, namely, the efficiency of the estimates and the independence between the effects and the regressors (i.e., the validity of the strict exogeneity assumption of the regressors (3.4.1); e.g., Maddala 1971a; Mundlak 1978a (see Chapter 3)). When all the explanatory variables are fixed constants or strictly exogenous relative to u, the covariance estimator is the best linear

[1] We defer the discussion of estimating distributed-lag models to Chapter 11.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

unbiased estimator under the fixed-effects assumption and a consistent and unbiased estimator under the random-effects assumption, even though it is not efficient. However, when there exist omitted individual attributes that are correlated with the included exogenous variables, the covariance (CV) estimator does not suffer from bias due to omission of these relevant individual attributes because their impacts have been differenced out, but a generalized least-squares estimator for the random-effects model under the assumption of independence between the effects and explanatory variables is biased. Furthermore, in a linear static model if the effects are correlated with the mean of the explanatory variables, a correctly formulated random-effects model leads to the same CV estimator as the fixed-effects model (Mundlak (1978a); see also Section 3.4 in Chapter 3). Thus, the fixed-effects model has assumed paramount importance in empirical studies (e.g., Ashenfelter 1978; Hausman 1978; Kiefer 1979).

However, if lagged dependent variables also appear as explanatory variables, strict exogeneity of the regressors no longer holds. The initial values of a dynamic process raise another problem. It turns out that with a random-effects formulation, the interpretation of a model depends on the assumption of initial observation. In the case of fixed-effects formulation, the maximum-likelihood estimator (MLE) or the CV estimator is no longer consistent in the typical situation in which a panel involves a large number of individuals, but over only a short period of time. The consistency and asymptotic properties of various fixed-effects estimators to be discussed in this chapter depend on the way in which the number of time series observations T and the number of cross-sectional units N tend to infinity.

For ease of exposition, we shall first assume that the time-specific effects, λ_t, do not appear. In Section 4.2 we discuss the properties of the CV (or the least squares dummy variable) estimator. Section 4.3 discusses the random-effects model. We discuss the implications of various formulation and methods of estimation. We show that the ordinary least-squares estimator is inconsistent but the MLE, the instrumental variable (IV), and the generalized method of moments (GMM) estimator are consistent. Procedures to test initial conditions are also discussed. In Section 4.4 we use Balestra and Nerlove's (1966) model of demand for natural gas to illustrate the consequences of various assumptions for the estimated coefficients. Section 4.5 discusses the estimation of the fixed-effects dynamic model. We show that although the conventional MLE and CV estimator are inconsistent when T is fixed and N tends to infinity, there exists a transformed likelihood approach that does not involve the incidental parameter and is consistent and efficient under proper formulation of initial conditions. We also discuss the IV and GMM estimator that does not need the formulation of initial conditions. Procedures to test fixed versus random effects are also suggested. In Section 4.6 we relax the assumption on the specific serial-correlation structure of the error term and propose a system approach to estimating dynamic models. Models with both individual- and time-specific effects are discussed in Section 7.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

4.2 THE CV ESTIMATOR

The CV transformation removes the individual-specific effects from the specification; hence the issue of random- versus fixed-effects specification does not arise. The CV estimator is consistent for the static model when either N or T or both are large. In the case of dynamic model, the properties of CV (or LDSV) depend on the way in which N and T goes to infinity.

Consider[2]

$$y_{it} = \gamma y_{i,t-1} + \alpha_i^* + u_{it}, \quad |\, \gamma \,| < 1, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{4.2.1}$$

where for simplicity we let $\alpha_i^* = \alpha_i + \mu$ to avoid imposing the restriction that $\sum_{i=1}^N \alpha_i = 0$. We also assume that y_{i0} are observable, $Eu_{it} = 0$, and $Eu_{it}u_{js} = \sigma_u^2$ if $i = j$ and $t = s$, and $Eu_{it}u_{js} = 0$ otherwise.

Let $\bar{y}_i = \sum_{t=1}^T y_{it}/T, \bar{y}_{i,-1} = \sum_{t=1}^T y_{i,t-1}/T$, and $\bar{u}_i = \sum_{t=1}^T u_{it}/T$. The LSDV (CV) estimators for α_i^* and γ are

$$\hat{\alpha}_i^* = \bar{y}_i - \hat{\gamma}_{\text{cv}}\bar{y}_{i,-1}, \qquad i = 1, \ldots, N, \tag{4.2.2}$$

$$\begin{aligned} \hat{\gamma}_{\text{cv}} &= \frac{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \bar{y}_i)(y_{i,t-1} - \bar{y}_{i,-1})}{\sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})^2} \\ &= \gamma + \frac{\sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_i)/NT}{\sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \bar{y}_{i,-1})^2/NT}. \end{aligned} \tag{4.2.3}$$

The CV estimator exists if the denominator of the second term of (4.2.3) is nonzero. It is consistent if the numerator of the second term of (4.2.3) converges to 0 as sample size increases.

By continuous substitution, we have

$$y_{it} = u_{it} + \gamma u_{i,t-1} + \cdots + \gamma^{t-1}u_{i1} + \frac{1-\gamma^t}{1-\gamma}\alpha_i^* + \gamma^t y_{i0}. \tag{4.2.4}$$

[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 $T \to \infty$ (Anderson 1959). When T is fixed and $N \to \infty$, 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 Chapter 4, Section 4.3.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Summing $y_{i,t-1}$ over t, we have

$$\begin{aligned}\sum_{t=1}^{T} y_{i,t-1} = {} & \frac{1-\gamma^T}{1-\gamma} y_{i0} + \frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2}\alpha_i^* \\ & + \frac{1-\gamma^{T-1}}{1-\gamma} u_{i1} + \frac{1-\gamma^{T-2}}{1-\gamma} u_{i2} + \cdots + u_{i,T-1}.\end{aligned} \tag{4.2.5}$$

Under the assumption that u_{it} are uncorrelated with α_i^* and are independently identically distributed, by a law of large numbers (Rao 1973), and using (4.2.5), we can show that when N tends to infinity,

$$\begin{aligned} & \operatorname*{plim}_{N\to\infty} \frac{1}{NT} \sum_{i=1}^{N}\sum_{t=1}^{T} (y_{i,t-1} - \bar{y}_{i,-1})(u_{it} - \bar{u}_i) \\ & = - \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \bar{y}_{i,-1}\bar{u}_i \\ & = -\frac{\sigma_u^2}{T^2} \cdot \frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2}. \end{aligned} \tag{4.2.6}$$

By similar manipulations we can show that the denominator of (4.2.3) converges to

$$\frac{\sigma_u^2}{1-\gamma^2}\left\{1 - \frac{1}{T} - \frac{2\gamma}{(1-\gamma)^2} \cdot \frac{(T-1) - T\gamma + \gamma^T}{T^2}\right\}. \tag{4.2.7}$$

as $N \to \infty$. If T is fixed, (4.2.6) is a nonzero constant, and (4.2.2) and (4.2.3) are inconsistent estimators no matter how large N is. The asymptotic bias of the CV of γ is

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} (\hat{\gamma}_{\text{cv}} - \gamma) = {} & -\frac{1+\gamma}{T-1}\left(1 - \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\right) \\ & \cdot \left\{1 - \frac{2\gamma}{(1-\gamma)(T-1)}\left[1 - \frac{1-\gamma^T}{T(1-\gamma)}\right]\right\}^{-1}.\end{aligned} \tag{4.2.8}$$

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 small T, this bias is always negative if $\gamma > 0$. Nor does the bias go to 0 as γ goes to 0. Because 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 estimator for the dynamic fixed-effects model remains biased with the introduction of exogenous variables if T is

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

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, (4.2.6) converges to zero, and (4.2.7) 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, the CV estimator remains consistent. However, the standard error of the CV is now of order $(\frac{1}{\sqrt{NT}})$.

The t-statistic

$$\frac{(\hat{\gamma}_{CV} - \gamma)}{\text{standard error of } \hat{\gamma}_{CV}} \tag{4.2.9}$$

is no longer centered at 0 because the order $\left(\frac{1}{T}\right)$ correlation between $(y_{i,t-1} - \bar{y}_i)$ and $(u_{it} - \bar{u}_i)$ gets magnified by large N. The scale factor $\sqrt{NT} = \sqrt{c}T$ if $\frac{N}{T} = c \neq 0 < \infty$ as $T \longrightarrow \infty$, $(\hat{\gamma}_{cv} - \gamma)$ divided by its standard error is equivalent to multiplying $(\hat{\gamma}_{cv} - \gamma)$ by a scale factor T. Equation (4.2.6) multiplied by T will not go to 0 no matter how large T is. Hahn and Kuersteiner (2002) have shown that $\sqrt{NT}(\hat{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean $-\sqrt{c}(1+\gamma)$ and variance $1-\gamma^2$. In other words, the usual t-statistic based on $\hat{\gamma}_{cv}$ is not centered at 0, and hence could be subject to severe size distortion when N also increases as T increases such that $\frac{N}{T} \rightarrow c \neq 0$ as $T \rightarrow \infty$ (e.g., Hsiao and Zhang 2013).

4.3 RANDOM-EFFECTS MODELS

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 and simply using the CV estimator no longer yields the desirable properties as in the case of static regression models. 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 $\alpha_i = \mathbf{a}'\bar{\mathbf{x}}_i + \omega_i$ (Mundlak 1978a) (see Section 3.4) and use a random-effects framework of the model

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\gamma + X_i\boldsymbol{\beta} + \mathbf{e}\bar{\mathbf{x}}_i'\mathbf{a} + \mathbf{e}\omega_i + u_i, \tag{4.3.1}$$

where now $E(\mathbf{x}_{it}\omega_i) = \mathbf{0}$ and $E(\mathbf{x}_{it}u_{it}) = \mathbf{0}$. However, because $\bar{\mathbf{x}}_i$ is time-invariant and the (residual) individual effect ω_i possesses the same property as α_i when the assumption $E\alpha_i\mathbf{x}_{it}' = \mathbf{0}'$ holds, the estimation of (4.3.1) is formally equivalent to the estimation of the model

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\gamma + X_i\boldsymbol{\beta} + \mathbf{e}\mathbf{z}_i'\boldsymbol{\rho} + \mathbf{e}\alpha_i + u_i, \tag{4.3.2}$$

with X_i now denoting the $T \times K_1$ time-varying explanatory variables, $\mathbf{z}_i'$ being the $1 \times K_2$ time-invariant explanatory variables including the intercept term, and $E\alpha_i = 0$, $E\alpha_i \mathbf{z}_i' = \mathbf{0}'$, and $E\alpha_i \mathbf{x}_{it}' = \mathbf{0}'$. So, for ease of exposition, we assume in this section that the effects are uncorrelated with the exogenous variables.[3]

We first show that the ordinary least-squares (OLS) estimator for dynamic error-component models is biased. We then discuss how the assumption about the initial observations affects interpretation of a model. Finally we discuss estimation methods and their asymptotic properties under various assumptions about initial conditions and sampling schemes.

4.3.1 Bias in the OLS Estimator

In the static case in which all the explanatory variables are exogenous and are uncorrelated with the effects, we can ignore the error-component structure and apply the OLS method. The OLS estimator, although less efficient, is still unbiased and consistent. But this is no longer true for dynamic error-component models. The correlation between the lagged dependent variable and individual-specific effects would seriously bias the OLS estimator. We use the following simple model to illustrate the extent of bias. Let

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + u_{it}, \quad \mid \gamma \mid < 1, \quad \begin{aligned} i &= 1, \dots, N, \\ t &= 1, \dots T, \end{aligned} \tag{4.3.3}$$

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

$$\hat{\gamma}_{\mathrm{LS}} = \frac{\sum_{i=1}^N \sum_{t=1}^T y_{it} \cdot y_{i,t-1}}{\sum_{i=1}^N \sum_{t=1}^T y_{i,t-1}^2} = \gamma + \frac{\sum_{i=1}^N \sum_{t=1}^T (\alpha_i + u_{it}) y_{i,t-1}}{\sum_{i=1}^N \sum_{t=1}^T y_{i,t-1}^2}. \tag{4.3.4}$$

The asymptotic bias of the OLS estimator is given by the probability limit of the second term on the right-hand side of (4.3.4). Using a manipulation similar to that in Section 4.2, we can show that

$$\begin{aligned} &\operatorname*{plim}_{N\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (\alpha_i + u_{it}) y_{i,t-1} \\ &= \frac{1}{T} \frac{1-\gamma^T}{1-\gamma} \mathrm{Cov}(y_{i0}, \alpha_i) + \frac{1}{T} \frac{\sigma_\alpha^2}{(1-\gamma)^2} \left[(T-1) - T\gamma + \gamma^T \right], \end{aligned} \tag{4.3.5}$$

[3] This does not mean that we have resolved the issue of whether or not the effects are correlated with the exogenous variables. It only means that for estimation purposes we can let α_i stand for ω_i and treat (4.3.1) as a special case of (4.3.2).

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2 &= \frac{1-\gamma^{2T}}{T(1-\gamma^2)} \cdot \operatorname{plim} \frac{\sum_{i=1}^{N} y_{i0}^2}{N} \\
&+ \frac{\sigma_\alpha^2}{(1-\gamma)^2} \cdot \frac{1}{T}\left(T - 2\frac{1-\gamma^T}{1-\gamma} + \frac{1-\gamma^{2T}}{1-\gamma^2}\right) \\
&+ \frac{2}{T(1-\gamma)}\left(\frac{1-\gamma^T}{1-\gamma} - \frac{1-\gamma^{2T}}{1-\gamma^2}\right)\operatorname{Cov}(\alpha_i, y_{i0}) \\
&+ \frac{\sigma_u^2}{T(1-\gamma^2)^2}\left[(T-1) - T\gamma^2 + \gamma^{2T}\right].
\end{aligned} \tag{4.3.6}$$

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 $\operatorname{Cov}(y_{i0}, \alpha_i)$ is either 0 or positive.[4] Under the assumption that the initial values are bounded, namely, that $\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} y_{i0}^2/N$ is finite, the OLS method overestimates the true autocorrelation coefficient γ when N or T or both tend to infinity. The overestimation is more pronounced the greater the variance 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 lagged 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.

4.3.2 Model Formulation

Consider a model of the form[5]

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{4.3.7}$$

[4] For details, see Chapter 4, Section 4.3.2 or Sevestre and Trognon (1982).

[5] The presence of the term $\mathbf{x}_{it}'\boldsymbol{\beta}$ shows that the process $\{y_{it}\}$ is not generally stationary. But the statistical properties of the process $\{y_{it}\}$ vary fundamentally when $T \to \infty$ according to whether or not $\{y_{it}\}$ converges to a stationary process when the sequence of $\mathbf{x}_{it}$ is identically 0. As stated in footnote 2, we shall adopt the first position by letting $\mid \gamma \mid < 1$.

where $| \gamma |< 1$, $v_{it} = \alpha_i + u_{it}$,

$$\begin{aligned} E\alpha_i &= Eu_{it} = 0, \\ E\alpha_i \mathbf{z}_i' &= \mathbf{0}', \quad E\alpha_i \mathbf{x}_{it}' = \mathbf{0}', \\ E\alpha_i u_{it} &= 0, \\ E\alpha_i \alpha_j &= \begin{cases} \sigma_\alpha^2 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \\ Eu_{it}u_{js} &= \begin{cases} \sigma_u^2 & \text{if } i = j, \quad t = s, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{4.3.8}$$

and where $\mathbf{z}_i$ is a $K_2 \times 1$ vector of time-invariant exogenous variables such as the constant term or an individual's sex or race, $\mathbf{x}_{it}$ is a $K_1 \times 1$ vector of time-varying exogenous variables, γ is 1×1, and $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ are $K_2 \times 1$ and $K_1 \times 1$ vectors of parameters, respectively. Equation (4.3.7) can also be written in the form

$$w_{it} = \gamma w_{i,t-1} + \boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}, \tag{4.3.9}$$

$$y_{it} = w_{it} + \eta_i, \tag{4.3.10}$$

where

$$\alpha_i = (1-\gamma)\eta_i, \quad E\eta_i = 0, \quad \text{Var}(\eta_i) = \sigma_\eta^2 = \sigma_\alpha^2/(1-\gamma)^2. \tag{4.3.11}$$

Algebraically, (4.3.7) is identical to (4.3.9) and (4.3.10). However, the interpretation of how y_{it} is generated is not the same. Equation (4.3.7) implies that apart from a common response to its own lagged value and the exogenous variables, each individual process is also driven by the unobserved characteristics, α_i, which are different for different individuals. Equations (4.3.9) and (4.3.10) imply that the dynamic process $\{w_{it}\}$ is independent of the individual effect η_i. Conditional on the exogenous variables, individuals are driven by an identical stochastic process with independent (and different) shocks that are random draws from a common population [equation (4.3.9)]. It is the observed value of the latent variable w_{it}, y_{it}, that is shifted by the individual time-invariant random variable η_i [equation (4.3.10)]. This difference in means can be interpreted as a difference in individual endowments or a common measurement error for the ith process.

If we observe w_{it}, we can distinguish (4.3.7) from (4.3.9) and (4.3.10). Unfortunately, w_{it} are unobservable. However, knowledge of initial observations can provide information to distinguish these two processes. Standard assumptions about initial observations are either that they are fixed or that they are random. If (4.3.7) is viewed as the model, we have two fundamental cases: (I) y_{i0} fixed and (II) y_{i0} random. If (4.3.9) and (4.3.10) are viewed as the basic model, we have (III) w_{i0} fixed and (IV) w_{i0} random.

Case I: y_{i0} fixed. A cross-sectional unit may start at some arbitrary position y_{i0} and gradually move toward a level $(\alpha_i + \boldsymbol{\rho}'\mathbf{z}_i)/(1-\gamma) + \boldsymbol{\beta}' \sum_{j=0} \mathbf{x}_{i,t-j}\gamma^j$.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

This level is determined jointly by the unobservable effect (characteristic) α_i, observable time-invariant characteristics $\mathbf{z}_i$, and time-varying variables $\mathbf{x}_{it}$. The individual effect, α_i, is a random draw from a population with mean 0 and variance σ_α^2. This appears to be a reasonable model. But if the decision about when to start sampling is arbitrary and independent of the values of y_{i0}, treating y_{i0} as fixed might be questionable because the assumption $E\alpha_i y_{i0} = 0$ implies that the individual effects, α_i, are not brought into the model at time 0, but affect the process at time 1 and later. If the process has been going on for some time, there is no particular reason to believe that y_{i0} should be viewed differently than y_{it}.

Case II: y_{i0} random. We can assume that the initial observations are random, with a common mean μ_{y0} and variance σ_{y0}^2. Namely, let

$$y_{i0} = \mu_{y0} + \epsilon_i. \tag{4.3.12}$$

A rationalization of this assumption is that we can treat y_{it} as a state. We do not care how the initial state, y_{i0}, is reached as long as we know that it has a distribution with finite mean and variance. Or, alternatively, we can view ϵ_i as representing the effect of initial individual endowments (after correction for the mean). Depending on the assumption with regard to the correlation between y_{i0} and α_i, we can divide this case into two subcases:

Case IIa: y_{i0} independent of α_i; that is, $\mathrm{Cov}(\epsilon_i, \alpha_i) = 0$. In this case the impact of initial endowments gradually diminishes over time and eventually vanishes. The model is somewhat like case I, in which the starting value and the effect α_i are independent, except that now the starting observable value is not a fixed constant but a random draw from a population with mean μ_{y0} and variance σ_{y0}^2.

Case IIb: y_{i0} correlated with α_i. We denote the covariance between y_{i0} and α_i by $\phi\sigma_{y0}^2$. Then, as time goes on, the impact of initial endowments (ϵ_i) affects all future values of y_{it} through its correlation with α_i and eventually reaches a level $\phi\epsilon_i/(1-\gamma) = \lim_{t\to\infty} E[y_{it} - \boldsymbol{\rho}'\mathbf{z}_i/(1-\gamma) - \boldsymbol{\beta}' \sum_{j=0}^{t-1} \mathbf{x}_{i,t-j}\gamma^j \mid \epsilon_i]$. In the special case that $\phi\sigma_{y0}^2 = \sigma_\alpha^2$, namely, $\epsilon_i = \alpha_i$, the individual effect can be viewed as completely characterized by the differences in initial endowments. The eventual impact of this initial endowment equals $\alpha_i/(1-\gamma) = \eta_i$.

Case III: w_{i0} fixed. Here the unobserved individual process $\{w_{it}\}$ has an arbitrary starting value. In this sense, this case is similar to case I. However, the observed cross-sectional units, y_{it}, are correlated with the individual effects, η_i. That is, each of the observed cross-sectional units may start at some arbitrary position y_{i0} and gradually move toward a level $\eta_i + \boldsymbol{\rho}'\mathbf{z}_i/(1-\gamma)$ $+\boldsymbol{\beta}' \sum_{j=0}^{t-1} \mathbf{x}_{i,t-j}\gamma^j$. Nevertheless, we allow for the possibility that the starting period of the sample observations need not coincide with the beginning of a stochastic process by letting the individual effect η_i affect all sample observations, including y_{i0}.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Case IV: w_{i0} random. Depending on whether or not the w_{i0} are viewed as having common mean, we have four subcases:

Case IVa: w_{i0} random, with common mean μ_w and variance $\sigma_u^2/(1-\gamma^2)$

Case IVb: w_{i0} random, with common mean μ_w and arbitrary variance σ_{w0}^2

Case IVc: w_{i0} random, with mean θ_{i0} and variance $\sigma_u^2/(1-\gamma^2)$

Case IVd: w_{i0} random, with mean θ_{i0} and arbitrary variance σ_{w0}^2

In each of these four subcases we allow correlation between y_{i0} and η_i. In other words, η_i affects y_{it} in all periods, including y_{i0}. Cases IVa and IVb are similar to the state-space representation discussed in case IIa, in which the initial states are random draws from a distribution with finite mean. Case IVa assumes that the initial state has the same variance as the latter states. Case IVb allows the initial state to be nonstationary (with arbitrary variance). Cases IVc and IVd take a different view in that they assume that the individual states are random draws from different populations with different means. A rationalization for this can be seen through successive substitution of (4.3.9), yielding

$$w_{i0} = \frac{1}{1-\gamma}\boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\sum_{j=0}^{\infty}\mathbf{x}_{i,-j}\gamma^j + u_{i0} + \gamma u_{i,-1} + \gamma^2 u_{i,-2} + \ldots \qquad (4.3.13)$$

Because $\mathbf{x}_{i0}, \mathbf{x}_{i,-1}, \ldots$ are not observable, we can treat the combined cumulative effects of nonrandom variables for the ith individual as an unknown parameter and let

$$\theta_{i0} = \frac{1}{1-\gamma}\boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\sum_{j=0}^{\infty}\mathbf{x}_{i,-j}\gamma^j \qquad (4.3.14)$$

Case IVc assumes that the process $\{w_{it}\}$ was generated from the infinite past and has achieved stationarity of its second moments after conditioning on the exogenous variables (i.e., w_{i0} has the same variance as any other w_{it}). Case IVd relaxes this assumption by allowing the variance of w_{i0} to be arbitrary.

4.3.3 Estimation of Random-Effects Models

There are various ways to estimate the unknown parameters. Here we discuss four methods: the MLE, the GLS, the instrumental-variable (IV), and the GMM methods.

4.3.3.1 Maximum-Likelihood Estimator

Different assumptions about the initial conditions imply different forms of the likelihood functions. Under the assumption that α_i and u_{it} are normally

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

distributed, the likelihood function for case I is[6]

$$\begin{aligned} L_1 = (2\pi)^{-\frac{NT}{2}} \mid V \mid^{-\frac{N}{2}} \exp\Bigg\{ -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{y}_{i,-1}\gamma \\ - Z_i\boldsymbol{\rho} - X_i\boldsymbol{\beta})' \cdot V^{-1}(\mathbf{y}_i - \mathbf{y}_{i,-1}\gamma - Z_i\boldsymbol{\rho} - X_i\boldsymbol{\beta})\Bigg\}, \end{aligned} \tag{4.3.15}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $Z_i = \mathbf{e}\mathbf{z}_i'$, $\mathbf{e} = (1, \ldots, 1)'$, $X_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$, and $V = \sigma_u^2 I_T + \sigma_\alpha^2 \mathbf{e}\mathbf{e}'$. The likelihood function for case IIa is

$$L_{2\mathrm{a}} = L_1 \cdot (2\pi)^{-\frac{N}{2}}\left(\sigma_{y0}^2\right)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma_{y0}^2}\sum_{i=1}^{N}(y_{i0} - \mu_{y0})^2\right\}. \tag{4.3.16}$$

For case IIb, it is of the form

$$\begin{aligned} L_{2\mathrm{b}} = (2\pi)^{-\frac{NT}{2}}\left(\sigma_u^2\right)^{-\frac{N(T-1)}{2}}\left(\sigma_u^2 + Ta\right)^{-\frac{N}{2}} \exp\Bigg\{ -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}\sum_{t=1}^{T} \\ \cdot [y_{it} - \gamma y_{i,t-1} - \boldsymbol{\rho}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it} - \phi(y_{i0} - \mu_{y0})]^2 \\ + \frac{\mathrm{a}}{2\sigma_u^2\left(\sigma_u^2 + Ta\right)}\sum_{i=1}^{N}\Bigg\{\sum_{t=1}^{T}[y_{it} - \gamma y_{i,t-1} - \boldsymbol{\rho}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it} - \phi \\ \cdot (y_{i0} - \mu_{y0})]\Bigg\}^2\Bigg\} \cdot (2\pi)^{-\frac{N}{2}}\left(\sigma_{y0}^2\right)^{-\frac{N}{2}} \\ \cdot \exp\left\{ -\frac{1}{2\sigma_{y0}^2}\sum_{i=1}^{N}(y_{i0} - \mu_{y0})^2\right\}, \end{aligned} \tag{4.3.17}$$

where $a = \sigma_\alpha^2 - \phi^2\sigma_{y0}^2$. The likelihood function for case III is

$$\begin{aligned} L_3 = (2\pi)^{-\frac{NT}{2}}\left(\sigma_u^2\right)^{-\frac{NT}{2}} \exp\Bigg\{ -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it} - y_{i0} + w_{i0}) \\ - \gamma(y_{i,t-1} - y_{i0} + w_{i0}) - \boldsymbol{\rho}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it}]^2\Bigg\} \cdot (2\pi)^{-\frac{N}{2}}\left(\sigma_\eta^2\right)^{-\frac{N}{2}} \\ \cdot \exp\left\{ -\frac{1}{2\sigma_\eta^2}\sum_{i=1}^{N}(y_{i0} - w_{i0})^2\right\}, \end{aligned} \tag{4.3.18}$$

[6] V is the same as (3.3.4).

and for Case IVa it is

$$\begin{aligned} L_{4\mathrm{a}} = {} & (2\pi)^{-\frac{N(T+1)}{2}} \mid \Omega \mid^{-\frac{N}{2}} \\ & \cdot \exp\left\{ -\frac{1}{2}\sum_{i=1}^{N}(y_{i0}-\mu_w,\, y_{i1}-\gamma y_{i0}-\boldsymbol{\rho}'\mathbf{z}_i-\boldsymbol{\beta}'\mathbf{x}_{i1},\ldots, \right. \\ & y_{iT}-\gamma y_{i,T-1}-\boldsymbol{\rho}'\mathbf{z}_i-\boldsymbol{\beta}'\mathbf{x}_{iT}) \\ & \left. \Omega^{-1}(y_{i0}-\mu_w,\ldots,y_{iT}-\gamma y_{i,T-1}-\boldsymbol{\rho}'\mathbf{z}_i-\boldsymbol{\beta}'\mathbf{x}_{iT})' \right\}, \end{aligned} \tag{4.3.19}$$

where

$$\begin{aligned} \underset{(T+1)\times(T+1)}{\Omega} &= \sigma_u^2\begin{bmatrix} \frac{1}{1-\gamma^2} & \mathbf{0}' \\ \mathbf{0} & I_T \end{bmatrix} + \sigma_\alpha^2 \begin{bmatrix} \frac{1}{1-\gamma} \\ \mathbf{e} \end{bmatrix}\left(\frac{1}{1-\gamma}, \mathbf{e}'\right) \\ \mid \Omega \mid &= \frac{\sigma_u^{2T}}{1-\gamma^2}\left(\sigma_u^2 + T\sigma_\alpha^2 + \frac{1+\gamma}{1-\gamma}\sigma_\alpha^2\right), \\ \Omega^{-1} &= \frac{1}{\sigma_u^2}\left[\begin{bmatrix} 1-\gamma^2 & \mathbf{0}' \\ \mathbf{0} & I_T \end{bmatrix}\right. \\ & \qquad \left. - \left(\frac{\sigma_u^2}{\sigma_\alpha^2} + T + \frac{1+\gamma}{1-\gamma}\right)^{-1}\begin{bmatrix} 1+\gamma \\ \mathbf{e} \end{bmatrix}(1+\gamma, \mathbf{e}')\right]. \end{aligned} \tag{4.3.20}$$

The likelihood function for case IVb, $L_{4\mathrm{b}}$, is of the form (4.3.19), except that Ω is replaced by Λ, where Λ differs from Ω only in that the upper left element of the first term, $1/(1-\gamma^2)$, is replaced by σ_{w0}^2/σ_u^2. The likelihood function for case IVc, $L_{4\mathrm{c}}$, is similar to that for case IVa, except that the mean of y_{i0} in the exponential term is replaced by θ_{i0}. The likelihood function for case IVd, $L_{4\mathrm{d}}$, is of the form (4.3.17), with θ_{i0}, $(1-\gamma)\sigma_\eta^2/(\sigma_\eta^2+\sigma_{w0}^2)$, and $\sigma_\eta^2+\sigma_{w0}^2$ replacing μ_{y0}, ϕ, and σ_{y0}^2, respectively.

Maximizing the likelihood function with respect to unknown parameters yields the MLE. The consistency of the MLE depends on the initial conditions and on the way in which the number of time series observations T and the cross-sectional units N tends to infinity. For cases III and IVd, the MLEs do not exist. By letting y_{i0} equal to w_{i0} or θ_{i0}, the exponential term of the second function of their respective likelihood function becomes 1. If we let the variances σ_η^2 or $\sigma_\eta^2+\sigma_{w0}^2$ approach 0, the likelihood functions become unbounded. However, we can still take partial derivatives of these likelihood functions and solve for the first-order conditions. For simplicity of exposition, we shall refer to these interior solutions as the MLEs and examine their consistency properties in the same way as in other cases in which the MLEs exist.

When N is fixed, a necessary condition for $\boldsymbol{\rho}$ being identifiable is that $N \geq K_2$. Otherwise, the model is subject to strict multicollinearity. However, when T tends to infinity, even with N greater than K_2, the MLEs for $\boldsymbol{\rho}$ and σ_α^2 remain inconsistent because of insufficient variation across individuals. On

Table 4.1. *Consistency properties of the MLEs for dynamic random-effects models*[a]

Case		N fixed, $T \to \infty$	T fixed, $N \to \infty$
Case I: y_{i0} fixed	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Consistent
	$\boldsymbol{\rho}, \sigma_\alpha^2$	Inconsistent	Consistent
Case II: y_{i0} random			
IIa: y_{i0} independent of α_i	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Consistent
	$\mu_{y0}, \boldsymbol{\rho}, \sigma_\alpha^2, \sigma_{y0}^2$	Inconsistent	Consistent
IIb: y_{i0} correlated	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Consistent
with α_i	$\mu_{y0}, \boldsymbol{\rho}, \sigma_\alpha^2, \sigma_{y0}^2, \phi$	Inconsistent	Consistent
Case III: w_{i0} fixed	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Inconsistent
	$w_{i0}, \boldsymbol{\rho}, \sigma_\eta^2$	Inconsistent	Inconsistent
Case IV: w_{i0} random			
IVa: mean μ_w			
and variance	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Consistent
$\sigma_u^2/(1-\gamma^2)$	$\mu_w, \boldsymbol{\rho}, \sigma_\eta^2$	Inconsistent	Consistent
IVb: mean μ_w	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Consistent
and variance σ_{w0}^2	$\sigma_{w0}^2, \boldsymbol{\rho}, \sigma_\eta^2, \mu_w$	Inconsistent	Consistent
IVc: mean θ_{i0}	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Inconsistent
and variance $\sigma_u^2/(1-\gamma^2)$	$\theta_{i0}, \boldsymbol{\rho}, \sigma_\eta^2$	Inconsistent	Inconsistent
IVd: mean θ_{i0}	$\gamma, \boldsymbol{\beta}, \sigma_u^2$	Consistent	Inconsistent
and variance σ_{w0}^2	$\theta_{i0}, \boldsymbol{\sigma}_\eta^2, \sigma_{w0}^2$	Inconsistent	Inconsistent

[a] If an MLE does not exist, we replace it by the interior solution.

Source: Anderson and Hsiao (1982, Table 1).

the other hand, the MLEs of γ, $\boldsymbol{\beta}$, and σ_u^2 are consistent for all these different cases. When T becomes large, the weight of the initial observations becomes increasingly negligible, and the MLEs for different cases all converge to the same CV estimator.

For cases IVc and IVd, w_{i0} have means θ_{i0}, which introduces incidental parameter problems. The MLE in the presence of incidental parameters is inconsistent. Bhargava and Sargan (1983) suggest predicting θ_{i0} by all the observed $\mathbf{x}_{it}$ and $\mathbf{z}_i$ as a means to get around the incidental-parameters problem.[7] If $\mathbf{x}_{it}$ is generated by a homogeneous stochastic process

$$\mathbf{x}_{it} = \mathbf{c} + \sum_{j=0}^{\infty} \mathbf{b}_j \boldsymbol{\xi}_{i,t-j}, \tag{4.3.21}$$

[7] Bhargava and Sargan (1983) get around the issue of incidental parameter associated with initial value, y_{i0}, by projecting y_{i0} on $\mathbf{x}_i$ under the assumption that α_i and $\mathbf{x}_i$ are uncorrelated. Chamberlain (1984) and Mundlak (1978a) assume that the effects, α_i, are correlated with $\mathbf{x}_i$ and get around the issue of incidental parameters by projecting α_i on $\mathbf{x}_i$ under the assumption that $(\alpha_i, \mathbf{x}_i')$ are independently, identically distributed over i.

where $\boldsymbol{\xi}_{it}$ is independently, identically distributed, then the minimum mean square error predictor of $\mathbf{x}_{i,-j}$ by $\mathbf{x}_{it}$ is the same for all i. Substituting these predictive formulae into (4.3.14) yields

$$y_{i0} = \sum_{t=1}^{T} \boldsymbol{\pi}'_{0t}\mathbf{x}_{it} + \boldsymbol{\rho}^{*\prime}\mathbf{z}_i + v_{i0}, \tag{4.3.22}$$

and

$$v_{i0} = \epsilon_{i0} + u^*_{i0} + \eta_i. \qquad i = 1, \ldots, N. \tag{4.3.23}$$

The coefficients $\boldsymbol{\pi}_{0t}$ are identical across i (Hsiao, Pesaran, and Tahmiscioglu 2002). The error term v_{i0} is the sum of three components: the prediction error of θ_{i0}, ϵ_{i0}; the cumulative shocks before time 0, $u^*_{i0} = u_{i0} + \gamma u_{i,-1} + \gamma^2 u_{i,-2} + \ldots$; and the individual effects, η_i. The prediction error ϵ_{i0} is independent of u_{it} and η_i, with mean 0 and variance $\sigma^2_{\epsilon 0}$. Depending on whether or not the error process of w_{i0} conditional on the exogenous variables has achieved stationarity (i.e., whether or not the variance of w_{i0} is the same as any other w_{it}), we have[8] case IVc′,

$$\begin{aligned} \operatorname{Var}(v_{i0}) &= \sigma^2_{\epsilon 0} + \frac{\sigma^2_u}{1-\gamma^2} + \frac{\sigma^2_\alpha}{(1-\gamma)^2} \text{ and} \\ \operatorname{Cov}(v_{i0}, v_{it}) &= \frac{\sigma^2_\alpha}{(1-\gamma)}, \qquad t = 1, \ldots, T, \end{aligned} \tag{4.3.24}$$

or case IVd′,

$$\operatorname{Var}(v_{i0}) = \sigma^2_{w0} \quad \text{and} \quad \operatorname{Cov}(v_{i0}, v_{it}) = \sigma^2_\tau, \qquad t = 1, \ldots, T. \tag{4.3.25}$$

Cases IVc′ and IVd′ transform cases IVc and IVd, 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 $T \to \infty$ or both $N, T \to \infty$. Moreover, the MLE multiplied by the scale factor $\sqrt{NT}$ is centered at the true values independent of the way N or T goes to infinity (for details, see Hsiao and Zhang 2013).

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 $\sigma^2_\alpha > 0$, the solution is the MLE. However, just as in the static case discussed in Section 3.3, a boundary solution

[8] Strictly speaking, from (4.3.21), the nonstationary analogue of case IVd would imply that

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

However, given the existence of the prediction-error term ϵ_{i0}, it is not possible to distinguish this case from case IVc' based on the information of y_{i0} alone. So we shall follow Bhargava and Sargan (1983) in treating case IVd′ as the nonstationary analogue of case IVd.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

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) has provided analytic explanations based on asymptotic approximations where the number of time periods tends to infinity. Nerlove (1967, 1971a) has 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 (4.3.7) 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 (4.3.7) in a given time period is written as an equation in a system of simultaneous equations,

$$\Gamma Y' + BX' + PZ' = U', \tag{4.3.26}$$

where we now let[9]

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

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

$$\underset{N\times K_2}{Z} = \begin{bmatrix} \mathbf{z}'_1 \\ \mathbf{z}'_2 \\ \vdots \\ \mathbf{z}'_N \end{bmatrix}, \qquad i = 1\ldots, N,$$

and U is the $N \times T$ matrix of errors if the initial values, y_{i0}, are treated as constants, or 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

[9] Previously we combined the intercept term and the time-varying exogenous variables into the $\mathbf{x}_{it}$ 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 $\mathbf{x}'_{it}$ as $(1, \tilde{\mathbf{x}}'_{it})$, because we wish to avoid having the constant term appearing more than once in (4.3.22).

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

$T \times [(T+1) + TK_1 + K_2]$ or $(T+1) \times [(T+1) + TK_1 + K_2]$, depending on whether the initial values are treated as fixed or random. The earlier serial covariance matrix [e.g., (3.3.4), (4.3.20), (4.3.24), or (4.3.25)] now becomes the variance–covariance matrix of the errors on T or $(T+1)$ structural equations. We can then use the algorithm for solving the full-information maximum-likelihood estimator to obtain the MLE.

There are cross-equation linear restrictions on the structural form coefficient matrix and restrictions on the variance–covariance matrix. For instances, in case I, where y_{i0} are treated as fixed constants, we have

$$A = \begin{bmatrix} -\gamma & 1 & 0 & \cdot & \cdot & 0 & \boldsymbol{\beta}' & \mathbf{0}' & \cdot & \cdot & \cdot & \cdot & \mathbf{0}' & \boldsymbol{\rho}' \\ 0 & -\gamma & 1 & \cdot & \cdot & \cdot & \mathbf{0}' & \boldsymbol{\beta}' & \cdot & \cdot & \cdot & \cdot & \cdot & \boldsymbol{\rho}' \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \boldsymbol{\rho}' \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ 0 & 0 & 0 & \cdot & -\gamma & 1 & \mathbf{0}' & \mathbf{0}' & \cdot & \cdot & \cdot & \cdot & \boldsymbol{\beta}' & \boldsymbol{\rho}' \end{bmatrix}, \tag{4.3.27}$$

The variance–covariance matrix of U is block-diagonal, with the diagonal block equal to V [equation (3.3.4)]. In case IVd′, where y_{i0} are treated as stochastic, the structural form coefficient matrix A is a $(T+1) \times [(T+1) + TK_1 + K_2]$ matrix of the form

$$A = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 & \boldsymbol{\pi}'_{01} & \boldsymbol{\pi}'_{02} & \cdot & \cdot & \cdot & \boldsymbol{\pi}'_{0T} & \boldsymbol{\rho}^{*\prime} \\ -\gamma & 1 & \cdot & \cdot & \cdot & \cdot & \boldsymbol{\beta}' & \mathbf{0}' & \cdot & \cdot & \cdot & \mathbf{0}' & \boldsymbol{\rho}' \\ 0 & -\gamma & \cdot & \cdot & \cdot & \cdot & \mathbf{0}' & \boldsymbol{\beta}' & \cdot & \cdot & \cdot & \mathbf{0}' & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ 0 & \cdot & \cdot & \cdot & -\gamma & 1 & \mathbf{0}' & \mathbf{0}' & \cdot & \cdot & \cdot & \boldsymbol{\beta}' & \boldsymbol{\rho}' \end{bmatrix}, \tag{4.3.28}$$

and the variance–covariance matrix of U is block-diagonal, with the diagonal block a $(T+1) \times (T+1)$ matrix of the form

$$\tilde{V} = \begin{bmatrix} \sigma^2_{w0} & \sigma^2_\tau \mathbf{e}' \\ \sigma^2_\tau \mathbf{e} & V \end{bmatrix}. \tag{4.3.29}$$

Bhargava and Sargan (1983) suggest maximizing the likelihood function of (4.3.26) by directly substituting the restrictions into the structural form coefficient matrix A and the variance–covariance matrix of U'.

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.[10] Or we can use the Chamberlain (1982, 1984) minimum-distance estimator by first obtaining the

[10] For the formula of the constrained estimator, see Theil 1971, p. 285, equation (8.5).

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

unconstrained reduced form coefficients matrix Π, then solving for the structural form parameters (see Section 3.9). The Chamberlain minimum-distance estimator has the same limiting distribution as the constrained generalized 3SLS estimator (see Chapter 5). 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.[11] But if the restrictions on the variance–covariance matrix are not true, the (full information) MLE imposing the wrong restrictions will in general 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.

4.3.3.2 Generalized Least-Squares Estimator

We note that except for Cases III, IVc, and IVd, the likelihood function depends only on a fixed number of parameters. Furthermore, conditional on Ω or σ_u^2, σ_α^2, σ_{y0}^2, and ϕ, the MLE is equivalent to the generalized least-squares estimator. For instance, under Case I, the covariance matrix of $(y_{i1}, \ldots, y_{iT})$ is the usual error-components form (3.3.4). Under Case IIa, b and Case IVa, b or Case IVc and IVd when the conditional mean of θ_{i0} can be represented in the form of (4.3.22), the covariance matrix of $\mathbf{v}_i = (v_{i0}, v_{i1}, \ldots, v_{iT})$, $\tilde{V}$, is of similar form to (4.3.29). Therefore, a GLS estimator of $\boldsymbol{\delta}' = (\boldsymbol{\pi}', \boldsymbol{\rho}^{*\prime}, \gamma, \boldsymbol{\beta}', \boldsymbol{\rho}')$, can be applied,

$$\hat{\boldsymbol{\delta}}_{\text{GLS}} = \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{V}^{-1} \tilde{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{V}^{-1} \tilde{\mathbf{y}}_i\right), \tag{4.3.30}$$

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

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

The estimator is consistent and asymptotically normally distributed as $N \to \infty$.

[11] See Chapter 5.

Blundell and Smith (1991) suggest a conditional GLS procedure by conditioning $(y_{i1}, \dots, y_{iT})$ on $v_{i0} = y_{i0} - E(y_{i0} \mid \mathbf{x}_i', \mathbf{z}_i)$,[12]

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\gamma + Z_i\boldsymbol{\rho} + X_i\boldsymbol{\beta} + \boldsymbol{\tau} v_{i0} + \mathbf{v}_i^*, \tag{4.3.31}$$

where $\mathbf{v}_i^* = (v_{i1}^*, \dots, v_{iT}^*)'$, 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. For Case IIa, $\boldsymbol{\tau} = \mathbf{0}$, Case IIb, $\boldsymbol{\tau} = \mathbf{e}_T \cdot \phi$. When the covariances between y_{i0} and $(y_{i1}, \dots, y_{iT})$ are arbitrary, $\boldsymbol{\tau}$ is a $T \times 1$ vector of unrestricted constants. Application of the GLS to (4.3.31) is consistent as $N \to \infty$.

When the covariance matrix of $\mathbf{v}_i$ or $\mathbf{v}_i^*$ 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 $\mathbf{v}_i$ or $\mathbf{v}_i^*$. For instance, we can use the IV estimator to be discussed in Section 4.3.3.3 to obtain consistent estimators of γ and $\boldsymbol{\beta}$, then substitute them into $y_{it} - \gamma y_{i,t-1} - \boldsymbol{\beta}'\mathbf{x}_{it}$, and regress the resulting value on $\mathbf{z}_i$ across individuals to obtain a consistent estimate of $\boldsymbol{\rho}$. Substituting estimated γ, $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ into (4.3.2), we obtain estimates of v_{it} for $t = 1, \dots, T$. The estimates of v_{i0} can be obtained as the residuals of the cross-section regression of (4.3.22). The covariance matrix of $\mathbf{v}_i$ can then be estimated using the procedures discussed in Chapter 3. The estimated $\mathbf{v}_i^*$ can also be obtained as the residuals of the cross-sectional regression of $\mathbf{y}_i - \mathbf{y}_{i,-1}\gamma - X_i\boldsymbol{\beta}$ on Z_i and $\mathbf{e}\hat{v}_{i0}$. In the second step, we treat the estimated covariance matrix of $\mathbf{v}_i$ or $\mathbf{v}_i^*$ as if they were known, apply the GLS to the system composed of (4.3.2) and (4.3.22) or the conditional system (4.3.31).

It should be noted that if Cov $(y_{i0}, \alpha_i) \neq 0$, the GLS applied to the system (4.3.2) 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 (4.3.31). Applying GLS to (4.3.2) is therefore subject to omitted variable bias. However, the asymptotic bias of the GLS of (4.3.2) is still smaller than that of the OLS or the within estimator of (4.3.2) (Sevestre and Trognon 1982). When T tends to infinity, GLS of (4.3.2) is again consistent because 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 (2002); or Appendix 4A).

[12] 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 mean different from 0 and the mean varies with i will give rise the incidental parameters problem.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

4.3.3.3 Instrumental–Variable Estimator

Because the likelihood functions under different initial conditions are different when dealing with panels involving large numbers of individuals over a short period of time, erroneous choices of initial conditions will yield estimators that are not asymptotically equivalent to the correct one, and hence may not be consistent. Sometimes we have little information to rely on in making a correct choice about the initial conditions. A simple consistent estimator that is independent of the initial conditions is appealing in its own right and in addition can be used to obtain initial values for the iterative process that yields the MLE. One estimation method consists of the following procedure.[13]

Step 1: Taking the first difference of (4.3.7), we obtain

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + u_{it} - u_{i,t-1}. \quad (4.3.32)$$

Because $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$ 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 $(y_{i,t-1} - y_{i,t-2})$ and estimate γ and $\boldsymbol{\beta}$ by the instrumental-variable method. Both

$$\begin{pmatrix} \hat{\gamma}_{iv} \\ \hat{\boldsymbol{\beta}}_{iv} \end{pmatrix} = \left[\sum_{i=1}^{N}\sum_{t=3}^{T} \cdot \begin{pmatrix} (y_{i,t-1} - y_{i,t-2})(y_{i,t-2} - y_{i,t-3}) & (y_{i,t-2} - y_{it-3})(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \\ (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})(y_{i,t-1} - y_{i,t-2}) & (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \end{pmatrix}\right]^{-1} \cdot \left[\sum_{i=1}^{N}\sum_{t=3}^{T} \begin{pmatrix} y_{i,t-2} - y_{i,t-3} \\ \mathbf{x}_{it} - \mathbf{x}_{i,t-1} \end{pmatrix} (y_{it} - y_{i,t-1})\right], \quad (4.3.33)$$

and

$$\begin{pmatrix} \tilde{\boldsymbol{\gamma}}_{iv} \\ \boldsymbol{\beta}_{iv} \end{pmatrix} = \left[\sum_{i=1}^{N}\sum_{t=2}^{T} \cdot \begin{pmatrix} y_{i,t-2}(y_{i,t-1} - y_{i,t-2}) & y_{i,t-2}(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \\ (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})(y_{i,t-1} - y_{i,t-2}) & (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})' \end{pmatrix}\right]^{-1} \cdot \left[\sum_{i=1}^{N}\sum_{t=2}^{T} \begin{pmatrix} y_{i,t-2} \\ \mathbf{x}_{it} - \mathbf{x}_{i,t-1} \end{pmatrix} (y_{it} - y_{i,t-1})\right], \quad (4.3.34)$$

are consistent.

Both (4.3.33) and (4.3.34) are derived using the sample moments $\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbf{q}_{it}(u_{it} - u_{i,t-1}) = 0$ to approximate the population moments $E[\mathbf{q}_{it}(u_{it} - u_{i,t-1})] = \mathbf{0}$, where $\mathbf{q}_{it} = [(y_{i,t-2} - y_{i,t-3}), (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})']'$ for (4.3.33) and $\mathbf{q}_{it} = [y_{i,t-2}, (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})']$ for (4.3.34). Therefore

[13] See Chapter 3, Section 3.5 for another approach.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

(4.3.33) or (4.3.34) is a consistent estimator and $\sqrt{NT}[(\hat{\gamma}_{iv} - \gamma), (\hat{\boldsymbol{\beta}}_{iv} - \boldsymbol{\beta})']'$ is asymptotically normally distributed with mean 0, either N or T or both tend to infinity (in other words, there is no asymptotic bias).

Estimator (4.3.34) has an advantage over (4.3.33) in the sense that the minimum number of time periods required is 2, whereas (4.3.33) requires $T \geq 3$. In practice, if $T \geq 3$, the choice between (4.3.34) and (4.3.33) 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 Anderson and Hsiao (1981).

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

$$\bar{y}_i - \gamma \bar{y}_{i,-1} - \boldsymbol{\beta}'\bar{\mathbf{x}}_i = \boldsymbol{\rho}'\mathbf{z}_i + \alpha_i + \bar{u}_i \qquad i = 1, \ldots, N, \tag{4.3.35}$$

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

Step 3: Estimate σ_u^2 and σ_α^2 by

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} \left[(y_{it} - y_{i,t-1}) - \hat{\gamma}(y_{i,t-1} - y_{i,t-2}) - \hat{\boldsymbol{\beta}}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})\right]^2}{2N(T-1)}, \tag{4.3.36}$$

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{N} \left(\bar{y}_i - \hat{\gamma}\bar{y}_{i,-1} - \hat{\boldsymbol{\rho}}'\mathbf{z}_i - \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_i\right)^2}{N} - \frac{1}{T}\hat{\sigma}_u^2. \tag{4.3.37}$$

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. The instrumental-variable method is simple to implement. But if we also wish to test the maintained hypothesis on initial conditions in the random-effects model, it would seem more appropriate to rely on maximum-likelihood methods.

4.3.3.4 Generalized Method of Moments Estimator

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–Bond (1991); Breusch, Mizon, and Schmidt (1989), etc. 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 $(y_{i,t-1} - y_{i,t-2})$. Letting $\mathbf{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2}, \mathbf{x}_i')'$, where $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, we have

$$E\mathbf{q}_{it}\Delta u_{it} = 0, \quad t = 2, \ldots, T. \tag{4.3.38}$$

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Stacking the $(T-1)$ first differenced equation of (4.3.32) in matrix form we have

$$\Delta \mathbf{y}_i = \Delta \mathbf{y}_{i,-1}\gamma + \Delta X_i \boldsymbol{\beta} + \Delta \mathbf{u}_i, \quad i = 1, \ldots, N \tag{4.3.39}$$

where $\Delta \mathbf{y}_i$, $\Delta \mathbf{y}_{i,-1}$ and $\Delta \mathbf{u}_i$ 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 $(\mathbf{x}_{i2} - \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT} - \mathbf{x}_{i,T-1})'$. The $T(T-1)[K_1 + \frac{1}{2}]$ orthogonality (or moment) conditions of (4.3.38) can be represented as

$$EW_i \Delta \mathbf{u}_i = \mathbf{0}, \tag{4.3.40}$$

where

$$W_i = \begin{pmatrix} \mathbf{q}_{i2} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{q}_{i3} & \ldots & \mathbf{0} \\ \vdots & & \vdots & \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{q}_{iT} \end{pmatrix}, \tag{4.3.41}$$

is of dimension $[T(T-1)(K_1 + \frac{1}{2})] \times (T-1)$. The dimension of (4.3.41) in general is much larger than $K_1 + 1$. Thus, Arellano–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 $\mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})$ denote some population moments of $\mathbf{y}$ and/or $\mathbf{x}$, say the first and second moments of $\mathbf{y}$ and/or $\mathbf{x}$, which are functions of the unknown parameter vector $\boldsymbol{\theta}$ and are supposed to equal some known constants, say 0. Let $\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})$ be their sample estimates based on N independent samples of $(\mathbf{y}_i, \mathbf{x}_i)$. Then the method of moments estimator $\boldsymbol{\theta}$ is the $\hat{\boldsymbol{\theta}}_{mm}$, such that

$$\mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = \hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}_{mm}) = \mathbf{0}. \tag{4.3.42}$$

For instance, the orthogonality conditions between QX_i and $Q\mathbf{u}_i$ for the fixed-effects linear static model (3.2.2), $E(X_i'Q\mathbf{u}_i) = E[X_i'Q(\mathbf{y}_i - \mathbf{e}\alpha_i^* - X_i\boldsymbol{\beta})] = \mathbf{0}$, lead to the LSDV estimator (3.2.8). In this sense, the IV method is a method of moments estimator.

If the number of equations in (4.3.42) 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 $\boldsymbol{\theta}$, (4.3.42) in general has no solution. It is then necessary to minimize some norm (or distance measure) of $\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})$, say

$$[\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})]' A [\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})], \tag{4.3.43}$$

where A is some positive definite matrix.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

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

$$A^* = \left\{E[\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})][\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})]'\right\}^{-1} \tag{4.3.44}$$

(Hansen 1982). The GMM estimator of $\boldsymbol{\theta}$ is to choose $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ such that it minimizes (4.3.43) when $A = A^*$.

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

$$\left(\frac{1}{N}\sum_{i=1}^{N} \Delta\mathbf{u}_i' W_i'\right) \Psi^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \Delta\mathbf{u}_i\right), \tag{4.3.45}$$

with respect to $\boldsymbol{\theta}$, where $\Psi = E[\frac{1}{N^2}\sum_{i=1}^{N} W_i \Delta\mathbf{u}_i \Delta\mathbf{u}_i' W_i']$. Under the assumption that u_{it} is i.i.d. with mean 0 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

$$\underset{(T-1)\times(T-1)}{\tilde{A}} = \begin{bmatrix} 2 & -1 & 0 & . & 0 \\ -1 & 2 & -1 & . & 0 \\ 0 & \ddots & \ddots & & \\ 0 & \ddots & \ddots & . & -1 \\ 0 & & . & -1 & 2 \end{bmatrix}. \tag{4.3.46}$$

Thus, the Arellano and Bond GMM estimator takes the form

$$\begin{aligned} &\hat{\boldsymbol{\theta}}_{\text{GMM,AB}} \\ &= \left\{\left[\sum_{i=1}^{N} \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_i'\right]\left[\sum_{i=1}^{N} W_i \tilde{A} W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i(\Delta\mathbf{y}_{i,-1}, \Delta X_i)\right]\right\}^{-1} \\ &\quad \cdot \left\{\left[\sum_{i=1}^{N} \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_i'\right]\left[\sum_{i=1}^{N} W_i \tilde{A} W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i \Delta\mathbf{y}_i\right]\right\}, \end{aligned} \tag{4.3.47}$$

with asymptotic covariance matrix

$$\begin{aligned} &\text{Cov}(\hat{\boldsymbol{\theta}}_{\text{GMM,AB}}) \\ &= \sigma_u^2 \left\{\left[\sum_{i=1}^{N} \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_i'\right]\left[\sum_{i=1}^{N} W_i \tilde{A} W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i(\Delta\mathbf{y}_{i,-1}, \Delta X_i)\right]\right\}^{-1}. \end{aligned} \tag{4.3.48}$$

In addition to the moment conditions (4.3.38), Arellano and Bover (1995) also note that $E\bar{v}_i = 0$, where $\bar{v}_i = \bar{y}_i - \bar{y}_{i,-1}\gamma - \bar{\mathbf{x}}_i'\boldsymbol{\beta} - \boldsymbol{\rho}'\mathbf{z}_i$.[14] Therefore, if instruments $\tilde{\mathbf{q}}_i$ exist (for instance, the constant 1 is a valid instrument) such that

$$E\tilde{\mathbf{q}}_i\bar{v}_i = \mathbf{0}, \tag{4.3.49}$$

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

Apart from the linear moment conditions (4.3.40), and (4.3.49), Ahn and Schmidt (1995) note that the homoscedasticity condition of $E(v_{it}^2)$ implies the following $T-2$ linear conditions:

$$E(y_{it}\Delta u_{i,t+1} - y_{i,t+1}\Delta u_{i,t+2}) = 0, \quad t = 1, \ldots, T-2. \tag{4.3.50}$$

Combining (4.3.40), (4.3.49), and (4.3.50), a more efficient GMM estimator can be derived by minimizing[15]

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i^{+\prime}W_i^{+\prime}\right)\Psi^{+-1}\left(\frac{1}{N}\sum_{i=1}^{N}W_i^{+}\mathbf{u}_i^{+}\right) \tag{4.3.51}$$

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

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

where

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

However, because the covariance matrix (4.3.50) 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

$$\hat{\Psi}^{+} = \left(\frac{1}{N^2}\sum_{i=1}^{N} W_i^{+}\hat{\mathbf{u}}_i^{+}\hat{\mathbf{u}}_i^{+\prime}W_i^{+^{\prime}}\right) \tag{4.3.52}$$

[14] Note that we let $\mathbf{z}_i = 0$ for ease of exposition. When $\mathbf{z}_i$ is present, the first differencing step of (4.3.38) eliminates $\mathbf{z}_i$ from the specification; hence the moment conditions (4.3.39) remain valid. However, for $Ev_i = 0$ to hold, it requires the assumption of stationarity in mean (Blundell and Bond 1998).

[15] For ease of notation, we again assume that $\mathbf{z}_i = \mathbf{0}$.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

into the objective function (4.3.51) to derive a linear estimator of form (4.3.47) where $\hat{\mathbf{u}}_i^+$ is derived by using some simple consistent estimator of γ and $\boldsymbol{\beta}$, say the IV discussed in Section 4.3.3.3, into (4.3.39) 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 (4.3.40), (4.3.49), and (4.3.50), there exist $(T-1)$ nonlinear moment conditions of the form $E((\bar{y}_i - \boldsymbol{\beta}'\bar{\mathbf{x}}_i)\Delta u_{it}) = 0, t = 2, \ldots, T$, 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 (4.3.50) can be transformed into the $(2T-2)$ linear moment conditions

$$E[(y_{iT} - \boldsymbol{\beta}'\mathbf{x}_{iT})\Delta y_{it}] = 0, \quad t = 1, \ldots, T-1, \tag{4.3.53}$$

and

$$E[(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it})y_{it} - (y_{i,t-1} - \boldsymbol{\beta}'\mathbf{x}_{i,t-1})y_{i,t-1}] = 0, \quad t = 2, \ldots, T. \tag{4.3.54}$$

Though 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 finite sample, owing mainly to bias. The bias is proportional to the number of instruments for each equation. In addition, when γ is close to 1, the lagged instruments $y_{i,t-2-j}$, $j \geq 0$ are weak instruments. There is also a bias-variance tradeoff in the number of moment conditions used for estimation. Koenken and Machado (1999) show that the usual asymptotic theory holds only if the number of moments used is less than the cubic root of the sample size. Okui (2009) proposes a moment selection method based on minimizing (Nagar 1959) the approximate mean square error. In general, when T is small, it is optimal to use all moment conditions. When T is not very small ($\frac{T^2}{N \log T} \longrightarrow \infty$), the optimal number of moment conditions chosen is $O((NT)^{1/3})$ 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. (Actually, Okui (2009) derives his selection method using the forward orthogonal deviation operator of Arellano and Bover (1995), $\Delta u_{it}^* = \sqrt{\frac{T-t}{T-t+1}}\left[u_{it} - \frac{1}{T-t}(u_{i,t+1} + \ldots + u_{iT})\right]$.) 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).

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

To improve the efficiency of GMM when γ is close to 1, Hahn, Hausman, and Kuersteiner (2007) suggest not using the first difference equation as in (4.3.32), but to use the long difference $y_{iT} - y_{i1}$. In the case of first-order autoregressive process (4.3.3),

$$y_{iT} - y_{i1} = \gamma(y_{i,T-1} - y_{io}) + (u_{iT} - u_{i1}). \tag{4.3.55}$$

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

$$E\begin{pmatrix} y_{io} \\ y_{iT-1} - \gamma y_{i,T-2} \\ \cdot \\ \cdot \\ \cdot \\ y_{i2} - \gamma y_{i1} \end{pmatrix} [(y_{iT} - y_{i1}) - \gamma(y_{i,T-1} - y_{io})] = \mathbf{0}. \tag{4.3.56}$$

The instruments $y_{it} - \gamma y_{i,t-1}$ for $t = 2, \ldots, T-1$ require knowledge of γ. A feasible LD estimator could be to use the Arellano–Bond GMM estimator (4.3.47) to obtain a preliminary consistent estimator $\hat{\gamma}_{\text{GMM,AB}}$, then use $(y_{io}, y_{i,T-1} - y_{i,T-2}\hat{\gamma}_{\text{GMM,AB}}, \ldots, y_{i2} - y_{i1}\hat{\gamma}_{\text{GMM,AB}})$ as instruments.

The reason that the LD estimator can improve the efficiency of the GMM based on the first difference equation of (4.3.3) 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,

$$E\,[\hat{\gamma}_{2\text{SLS}} - \gamma] \simeq \frac{1}{n}a, \tag{4.3.57}$$

where n denotes the sample size and

$$a = \frac{\text{(number of instruments)} \times \text{("covariance")}}{\text{"Explained" variance of the first stage reduced form equation}} \tag{4.3.58}$$

Based on this formula, Hahn, Hausman, and Kuersteiner (2007) show that $a = -\frac{1+\gamma}{1-\gamma}$ 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 4.3.1: We derive the MLE (or GLS) or the GMM estimator (4.3.47) assuming that u_{it} is independently distributed across i and over t. If u_{it} is serially correlated, $E(y_{i,t-2}\Delta u_{it}) \neq 0$ for $j \geq 2$. Then neither (4.3.30) nor (4.3.47) is a consistent estimator. On the other hand, the estimator $\hat{\boldsymbol{\theta}}^*$ that replaces W_i in (4.3.47) by the block diagonal instrument matrix $\tilde{W}_i^*$ whose

tth block is given by $\mathbf{x}_i$ if $\mathbf{x}_{it}$ is strictly exogenous (i.e., $E\mathbf{x}_{it}u_{is} = 0$ for all s) or $(\mathbf{x}'_{it}, \mathbf{x}'_{i,t-1}, \dots, \mathbf{x}'_{i1})'$ if $\mathbf{x}_{it}$ is weakly exogenous (i.e., $E(\mathbf{x}_{i,t+j+1}u_{it}) \neq 0$ and $E(u_{it}\mathbf{x}_{i,t-j}) = 0$ for $j \geq 0$)) remains consistent. Therefore, a Hausman (1978) type test statistic can be constructed to test if u_{it} are serially uncorrelated by comparing the difference of $(\hat{\boldsymbol{\theta}}_{\text{GMM,AB}} - \hat{\boldsymbol{\theta}}^*)$.

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

$$\frac{\sum_{i=1}^{N}\sum_{t=4}^{T} \Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}}{\hat{s}} \tag{4.3.59}$$

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

$$\begin{aligned}
\hat{s}^2 = {} & \sum_{i=1}^{N}\left(\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}\right)^2 - 2\left(\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{i,t-2}\Delta\mathbf{x}'_{it}\right) \\
& \cdot\left\{\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}'_{i,-1}\\ \Delta X'_i\end{pmatrix} W'_i\right]\left(\frac{1}{N}\sum_{i=1}^{N}W_i\tilde{A}W'_i\right)^{-1}\left[\sum_{i=1}^{N}W_i(\Delta\mathbf{y}_{i,-1}, \Delta X_i)\right]\right\}^{-1} \\
& \cdot\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}'_{i,-1}\\ \Delta X'_i\end{pmatrix} W'_i\right]\left(\frac{1}{N}\sum_{i=1}^{N}W_i\tilde{A}W'_i\right)^{-1}\left[\sum_{i=1}^{N}W_i\Delta\hat{\mathbf{u}}_i\left(\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}\right)\right] \\
& + \left(\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{i,t-2}\Delta\mathbf{x}'_{it}\right)(\text{Cov }(\hat{\boldsymbol{\theta}}_{\text{GMM,AB}}))\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta\mathbf{x}_{it}\Delta\hat{u}_{i-t-2}\right),
\end{aligned} \tag{4.3.60}$$

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

This statistic in lieu of Hausman-type test statistic can be used to test serial correlation in the case there exist no exogenous variables for model (4.3.7). The statistic (4.3.59) is defined only if $T \geq 5$. When $T < 5$, Arellano–Bond (1991) suggest using the Sargan (1958) test of overidentification,

$$\left(\sum_{i=1}^{N}\Delta\hat{\mathbf{u}}'_iW_i^{*'}\right)\left(\sum_{i=1}^{N}W_i^*\Delta\hat{\mathbf{u}}_i\Delta\hat{\mathbf{u}}'_iW_i^{*'}\right)^{-1}\left(\sum_{i=1}^{N}W_i^*\Delta\hat{\mathbf{u}}_i\right), \tag{4.3.61}$$

where W_i^* could be W_i or any number of instruments that satisfy the orthogonality condition $E(W_i^*\Delta\mathbf{u}_i) = 0$. Under the null of no serial correlation, (4.3.61) is asymptotically χ^2 distributed with $p - (K + 1)$, degrees of freedom for any $p > (K + 1)$, where p denotes the number of rows in W_i^*.

Remark 4.3.2: Because the individual-specific effects α_i are time-invariant, taking the deviation of individual y_{it} equation from any transformation of y_{it} equation that maintains the time-invariance property of α_i can eliminate α_i.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

For instance, Alvarez and Arellano (2003) consider the transformation of y_{it} equation into an equation of the form,

$$c_t[y_{it} - \frac{1}{(T-t)}(y_{i,t+1} + \cdots + y_{iT})], \quad t = 1, \ldots, T-1, \quad (4.3.62)$$

where $c_t^2 = \frac{(T-t)}{(T-t+1)}$. The advantage of considering the equation specified by transformation (4.3.62) is that the residuals $u_{it}^* = c_t[u_t - \frac{1}{(T-t)}(u_{i,t+1} + \ldots + u_{iT})], t = 1, \ldots, T-1$ are orthogonal, that is, $Eu_{it}^* u_{is}^* = 0$ if $t \neq s$ and $Eu_{it}^{*2} = \sigma_u^2$. However, if transformation (4.3.62) is used to remove α_i, the instruments $\mathbf{q}_{it}$ takes the form $(y_{i0}, \ldots, y_{i,t-1}, \mathbf{x}_i')$ in the application of GMM.

4.3.4 Testing Some Maintained Hypotheses on Initial Conditions

As discussed in Sections 4.3.2 and 4.3.3, the interpretation and consistency property 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 exogenous (Case I) we can test the validity of the error-components formulation by maximizing L_1 with or without the restrictions on the covariance matrix V. Let L_1^* denote the maximum of log L_1 subject to the restriction of model (4.3.7), and let L_1^{**} denote the maximum of log L_1 with V being an arbitrary positive definite matrix. Under the null hypothesis, the resulting test statistic $2(L_1^{**} - L_1^*)$ is asymptotically χ^2 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 the maximum of the log likelihood function under Case IVa and Case IVc′ be denoted by $L_{4\mathrm{a}}^*$ and $L_{4c'}^*$, respectively. Let the maximum of the log likelihood function under case IVa or IVc′ without the restriction (4.3.20) or (4.3.24) [namely, the $(T+1) \times (T+1)$ covariance matrix is arbitrary] be denoted by $L_{4\mathrm{a}}^{**}$ or $L_{4c'}^{**}$, respectively. Then, under the null, $2(L_{4a}^{**} - L_{4a}^*)$ and $2(L_{4c'}^{**} - L_{4c'}^*)$ are asymptotically χ^2, with $[(T+1)(T+2)/2 - 2]$ and $[(T+1)(T+2)/2 - 3]$ degrees of freedom, respectively.

To test the stationarity assumption, we denote the maximum of the log likelihood function for Case IVb and Case IVd′ as L_{4b}^* and $L_{4d'}^*$, respectively. Then $2(L_{4b}^* - L_{4a}^*)$ and $2(L_{4d'}^* - L_{4c'}^*)$ are asymptotically χ^2, with 1 degree of freedom. The statistics $2(L_{4a}^{**} - L_{4b}^*)$ and $2(L_{4c'}^{**} - L_{4d'}^*)$ can also be used to test the validity of Case IVb and Case IVd′, respectively. They are asymptotically χ^2 distributed, with $[(T+1)(T+2)/2 - 3]$ and $[(T+1)(T+2)/2 - 4]$ degrees of freedom, respectively.

We can also generalize the Bhargava and Sargan principle to test the assumption that the initial observations have a common mean μ_w or have different means θ_{i0} under various assumptions about the error process. The statistics $2[L_{4c'}^* - L_{4a}^*]$, $2[L_{4c'}^{**} - L_{4a}^{**}]$, or $2[L_{4d'}^* - L_{4b}^*]$ are asymptotically χ^2

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

distributed, with q, $(q-1)$, and $(q-1)$ degrees of freedom, respectively, where q is the number of unknown coefficients in (4.3.22). We can also test the combined assumption of a common mean and a variance-components formulation by using the statistic $2[L_{4c'}^{**} - L_{4a}^{*}]$ or $2[L_{4c'}^{**} - L_{4b}^{*}]$, both of which are asymptotically χ^2 distributed, with $q + (T+1)(T+2)/2 - 3$ and $q + (T+1)(T+2)/2 - 4$ degrees of freedom, respectively.

With regard to testing that y_{i0} are exogenous, unfortunately it is not possible to directly compare L_1 with the likelihood functions of various forms of case IV, 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 (4.3.7) and (4.3.22) under the restriction that v_{i0} are independent of η_i (or α_i) and have variance $\sigma_{\epsilon o}^2$. 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 likelihood function by L_5. Let L_5^{**} denote the maximum of log L_5 with unrestricted variance–covariance matrix for $(v_{i0}, \ldots, v_{iT})$. Then we can test the exogeneity of y_{i0} using $2(L_{4c'}^{**} - L_5^{**})$, which is asymptotically χ^2 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 (3.3.4)]. Let L_5^* denote the maximum of the log likelihood function L_5 under this restriction. Let $L_{4d'}^*$ denote the maximum of the log likelihood function of $(y_{i0}, \ldots, y_{iT})$ under the restriction that $E\mathbf{v}_i\mathbf{v}_i' = \tilde{V}^*$, 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_{w0}^2 *and* σ_τ^2. The statistic $2(L_{4d'}^* - L_5^*)$ is asymptotically χ^2 with 1 degree of freedom if y_{i0} are exogenous. In practice, however, it may not even be necessary to calculate $L_{4d'}^*$, because $L_{4d'}^* \geq L_{4c'}^*$, and if the null is rejected using $2(L_{4c'}^* - L_5^*)$ against the critical value of χ^2 with 1 degree of freedom, then $2(L_{4d'}^* - L_5^{**})$ must also reject the null.

4.3.5 Simulation Evidence

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

$$y_{it} = 1 + 0.5y_{i,t-1} - 0.16z_i + 0.35x_{it} + \alpha_i + u_{it}, \quad i = 1, \ldots, 100, \quad t = 1, \ldots, 20, \tag{4.3.63}$$

where α_i and u_{it} were independently normally distributed, with means 0 and variances 0.09 and 0.4225, respectively. The time-varying exogenous variables

x_{it} were generated by

$$x_{it} = 0.1t + \phi_i x_{i,t-1} + \omega_{it}, \qquad \begin{aligned} i &= 1, \ldots, 100, \\ t &= 1, \ldots, 20, \end{aligned} \tag{4.3.64}$$

with ϕ_i and ω_{it} independently normally distributed, with means 0 and variances 0.01 and 1, respectively. The time-invariant exogenous variables z_i were generated by

$$z_i = -0.2x_{i4} + \omega_i^*, \qquad i = 1, \ldots, 100, \tag{4.3.65}$$

and ω_i^* were independently normally distributed, with mean 0 and variance 1. The z and the x were held fixed over the replications, and the first 10 observations were discarded. Thus, the y_{i0} are in fact stochastic and are correlated with the individual effects α_i. Table 4.2 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 1 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 $2[L_{4c'}^{**} - L_5^{**}]$ and $2[L_{4c'}^{*} - L_5^{*}]$. Under the endogeneity assumption, the likelihood-ratio statistic $2(L_{4c'}^{**} - L_{4c'}^{*})$ rejected the error-components formulation 4 times (out of 50), whereas under the exogeneity assumption, the statistic $2(L_1^{**} - L_1^{*})$ rejected the error-components formulation 7 times.[16]

4.4 AN EXAMPLE

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 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

$$G_t^* = G_t - (1 - r)G_{t-1}. \tag{4.4.1}$$

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

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Table 4.2. *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).

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 (4.4.1), they obtained

$$G_t = \beta_0 + \beta_1 P_t + \beta_2 \Delta N_t + \beta_3 N_{t-1} + \beta_4 \Delta I_t + \beta_5 I_{t-1} + \beta_6 G_{t-1} + v_t, \tag{4.4.2}$$

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 1957–67 to estimate the model for residential and commercial demand for natural gas (4.4.2). 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 (4.4.2) are reported in the second column of Table 4.3. 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 (4.4.2), 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

Table 4.3. *Various estimates of the parameters of Balestra and Nerlove's demand-for-gas model (4.4.2) 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).

third column of Table 4.3. 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 greater than 30 percent – again highly implausible.

Instead of assuming the regional effect to be fixed, they again estimated (4.4.2) by explicitly incorporating individual state-specific effects 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 4.3. The estimated coefficient of lagged consumption is 0.9546. The implied depreciation rate is approximately 4.5 percent, which is in agreement with a priori expectation.

The foregoing results illustrate that by properly taking account of the unobserved heterogeneity in the panel data, Balestra and Nerlove 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 previous sections 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.[17]

4.5 FIXED-EFFECTS MODELS

If individual effects are considered fixed and different across individuals, because of strict multicollinearity between the effects and other time-invariant variables, there is no way one can disentangle the individual-specific effects from the impact of other time-invariant variables. We shall therefore assume $\mathbf{z}_i \equiv \mathbf{0}$. When T tends to infinity, even though lagged y does not satisfy the strict exogeneity condition for the regressors, it does satisfy the weak exogeneity condition of $E(u_{it} \mid y_{i,t-1}, y_{i,t-2}, .; \alpha_i) = 0$; hence the least-squares regression of y_{it} on lagged $y_{i,t-j}$ and $\mathbf{x}_{it}$ and the individual-specific constant yields a consistent estimator. In the case that T is fixed and N tends to infinity, the number of parameters in a fixed-effects specification increases with

[17] 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 mentioned in Chapter 4, Section 4.3, 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.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

the number of cross-sectional observations. This is the classical incidental parameters problem (Neyman and Scott 1948). In a static model with strict exogeneity assumption, the presence of individual specific constants does not affect the consistency of the CV or MLE estimator of the slope coefficients (see Chapter 3). However, the result no longer holds if lagged dependent variables also appear as explanatory variables. The regularity conditions for the consistency of the MLE are violated. In fact, if u_{it} are normally distributed and y_{i0} are given constants, the MLE of (4.2.1) is the CV of (4.2.2) and (4.2.3). The asymptotic bias is given by (4.2.8).

While the MLE is inconsistent when T is fixed and N is large, the IV estimator of (4.3.32) or the GMM estimator (4.3.43) remains consistent and asymptotically normally distributed with fixed α_i^*. The transformed equation (4.3.39) does not involve the incidental parameters α_i^*. The orthogonality condition (4.3.40) remains valid.

In addition to the IV type estimator, a likelihood-based approach based on a transformed likelihood function can also yield a consistent and asymptotically normally distributed estimator.

4.5.1 Transformed Likelihood Approach

The first difference equation (4.3.32) 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 $\mathbf{x}_{i0}$ are available. But (4.3.32) is not defined for $\Delta y_{i1} = (y_{i1} - y_{i0})$ because Δy_{i0} and $\Delta \mathbf{x}_{i0}$ are missing. However, by continuous substitution, we can write Δy_{i1} as

$$\Delta y_{i1} = a_{i1} + \sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j}, \tag{4.5.1}$$

where $a_{i1} = \boldsymbol{\beta}' \sum_{j=0}^{\infty} \Delta \mathbf{x}_{i,1-j} \gamma^j$. Since $\Delta \mathbf{x}_{i,1-j}$, $j = 1, 2, \ldots,$ 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,

$$E(a_{i1} \mid \Delta \mathbf{x}_i) = c^* + \boldsymbol{\pi}' \Delta \mathbf{x}_i, i = 1, \ldots, N, \tag{4.5.2}$$

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

$$\mathbf{x}_{it} = \boldsymbol{\mu}_i + \mathbf{g}t + \sum_{j=0}^{\infty} \mathbf{b}_j' \boldsymbol{\xi}_{i,t-j} \sum_{j=0}^{\infty} \mid b_j \mid < \infty, \tag{4.5.3}$$

where $\boldsymbol{\xi}_{it}$ are assumed to be i.i.d. with mean 0 and constant covariance matrix, then (4.5.2) holds. The data-generating process of the exogenous variables

$\mathbf{x}_{it}$ (4.5.3) 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 $\mathbf{x}_{it}$, then they must be identical across i.

Given (4.5.2), Δy_{i1} can be written as

$$\Delta y_{i1} = c^* + \boldsymbol{\pi}'\Delta \mathbf{x}_i + v_{i1}^*. \tag{4.5.4}$$

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

$$\Omega^* = \sigma_u^2 \begin{bmatrix} h & -1 & 0 & \ldots 0 \\ -1 & 2 & -1 & \\ 0 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \\ 0 & & & -1\ 2 \end{bmatrix} = \sigma_u^2 \tilde{\Omega}^*, \tag{4.5.5}$$

where $h = \frac{\sigma_{v^*}^2}{\sigma_u^2}$.

Combining (4.3.32) and (4.5.4), we can write the likelihood function of $\Delta \mathbf{y}_i^* = (\Delta y_{i1}, \ldots, \Delta y_{iT})'$, $i = 1, \ldots, N$, in the form of

$$(2\pi)^{-\frac{NT}{2}} \mid \Omega^* \mid^{-\frac{N}{2}} \ \exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} \Delta \mathbf{u}_i^{*\prime} \Omega^{*-1} \Delta \mathbf{u}_i^* \right\}, \tag{4.5.6}$$

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

$$\begin{aligned} \Delta \mathbf{u}_i^* = [&\Delta y_{i1} - c^* - \boldsymbol{\pi}'\Delta \mathbf{x}_i, \Delta y_{i2} - \gamma \Delta y_{i1} \\ &- \boldsymbol{\beta}'\Delta \mathbf{x}_{i2}, \ldots, \Delta y_{iT} - \gamma \Delta y_{i,T-1} - \boldsymbol{\beta}'\Delta \mathbf{x}_{iT}]'. \end{aligned} \tag{4.5.7}$$

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

$$\begin{aligned} \tilde{\Omega}^{*-1} &= [1 + T(h-1)]^{-1} \\ &\cdot \begin{bmatrix} T & T-1 & \ldots & 2 & 1 \\ T-1 & (T-1)h & & 2h & h \\ \vdots & \vdots & & \vdots & \vdots \\ 2 & 2h & & 2[(T-2)h-(T-3)] & (T-2)h-(T-3) \\ 1 & h & & (T-2)h-(T-3) & (T-1)h-(T-2) \end{bmatrix}, \end{aligned} \tag{4.5.8}$$

the logarithm of the likelihood function (4.5.6) is

$$\ln L = -\frac{NT}{2}\log 2\pi - \frac{NT}{2}\log \sigma_u^2 - \frac{N}{2}\log\left[1 + T(h-1)\right]$$
$$-\frac{1}{2}\sum_{i=1}^{N}\left[(\Delta \mathbf{y}_i^* - H_i\boldsymbol{\psi})'\Omega^{*-1}(\Delta \mathbf{y}_i^* - H_i\boldsymbol{\psi})\right], \quad (4.5.9)$$

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

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

The MLE is obtained by solving the following equations simultaneously:

$$\boldsymbol{\psi} = \left(\sum_{i=1}^{N} H_i'\hat{\tilde{\Omega}}^{*-1}H_i\right)^{-1}\left(\sum_{i=1}^{N} H_i'\hat{\tilde{\Omega}}^{*-1}\Delta \mathbf{y}_i^*\right), \quad (4.5.10)$$

$$\sigma_u^2 = \frac{1}{NT}\sum_{i=1}^{N}\left[(\Delta \mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})'(\hat{\tilde{\Omega}}^*)^{-1}(\Delta \mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})\right], \quad (4.5.11)$$

$$h = \frac{T-1}{T} + \frac{1}{\hat{\sigma}_u^2 NT^2}\sum_{i=1}^{N}\left[(\Delta \mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})'(\mathbf{J}\,\mathbf{J}')(\Delta \mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})\right], \quad (4.5.12)$$

where $\mathbf{J}' = (T, T-1, \ldots, 2, 1)$. One way to obtain the MLE is to iterate among (4.5.10)–(4.5.12) 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 -1 or explosive, see Han and Phillips (2013) and Kruiniger (2009).

4.5.2 Minimum Distance Estimator

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

$$\text{Min}\sum_{i=1}^{N}\Delta \mathbf{u}_i^{*\prime}\Omega^{*-1}\Delta \mathbf{u}_i^*. \quad (4.5.13)$$

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

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 (4.5.4) across i to obtain the least-squares residuals $\hat{v}_{i1}^*$, and then estimate

$$\hat{\sigma}_{v^*}^2 = \frac{1}{N - TK_1 - 1} \sum_{i=1}^{N} \hat{v}_{i1}^{*2}. \tag{4.5.14}$$

Similarly, we can apply the IV to (4.3.32) and obtain the estimated residuals $\Delta\hat{u}_{it}$ and

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \hat{\Delta\mathbf{u}}_i' \tilde{A}^{-1} \hat{\Delta\mathbf{u}}_i \tag{4.5.15}$$

where $\tilde{A}$ is defined in (4.3.46).

In the second step, we substitute estimated σ_u^2 and $\sigma_{v^*}^2$ into (4.5.5) and treat them as if they were known and use (4.5.10) to obtain the MDE of $\boldsymbol{\psi}$, $\hat{\boldsymbol{\psi}}_{\text{MDE}}$.

The asymptotic covariance matrix of MDE, Var $(\hat{\boldsymbol{\psi}}_{\text{MDE}})$, 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 $(\hat{\boldsymbol{\psi}}_{\text{FMDE}})$, contrary to the static case, is equal to (Hsiao, Pesaran, and Tahmiscioglu 2002)

$$\left(\frac{1}{N}\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1} + \left(\frac{1}{N}\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1} \begin{bmatrix} 0 & \mathbf{0}' & 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0}' & d & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \left(\frac{1}{N}\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1}, \tag{4.5.16}$$

where

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

The second term of (4.5.16) arises because the estimation of $\boldsymbol{\psi}$ and Ω^* are not asymptotically independent when the lagged dependent variables also appear as regressors.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

4.5.3 Relations between the Likelihood-Based Estimator and the GMM

Although normality is assumed to derive the transformed MLE and MDE, it is not required. Both estimators remain consistent and asymptotically normally distributed even the errors are not normally distributed. Under normality, the transformed MLE achieves the Cramér–Rao lower bound, and hence 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.

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

$$\operatorname{Cov}\begin{pmatrix} \boldsymbol{\gamma}_{MDE} \\ \boldsymbol{\beta}_{MDE} \end{pmatrix} = \sigma_u^2 \left[\sum_{i=1}^{N} \begin{pmatrix} \Delta \mathbf{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} \left(\tilde{A} - \frac{1}{h}\mathbf{g}\mathbf{g}' \right)^{-1} (\Delta \mathbf{y}_{i,-1}, \Delta X_i) \right]^{-1} \tag{4.5.17}$$

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

We note that (4.5.17) is smaller than

$$\sigma_u^2 \left[\sum_{i=1}^{N} \begin{pmatrix} \Delta \mathbf{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} \tilde{A}^{-1} (\Delta \mathbf{y}_{i,-1}, \Delta X_i) \right]^{-1}, \tag{4.5.18}$$

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

$$\begin{aligned} &\sum_{i=1}^{N} \begin{pmatrix} \Delta \mathbf{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} \tilde{A}^{-1} (\Delta \mathbf{y}_{i,-1}, \Delta X_i) - \left[\sum_{i=1}^{N} \begin{pmatrix} \Delta \mathbf{y}'_{i,-1} \\ \Delta X'_i \end{pmatrix} W'_i \right] \left(\sum_{i=1}^{N} W_i \tilde{A} W'_i \right)^{-1} \\ &\qquad \cdot \left[\sum_{i=1}^{N} W_i (\Delta y_{i,-1}, \Delta X_i) \right] \qquad\qquad (4.5.19) \\ &\quad = D'[I - Q(Q'Q)^{-1}Q]D, \end{aligned}$$

is a positive semidefinite matrix, where $D = (D'_1, \ldots, D'_N)'$, $Q = (Q'_1, Q'_2, \ldots, Q'_N)'$, $D_i = \Lambda'(\Delta \mathbf{y}_{i,-1}, \Delta X_i)$, $Q_i = \Lambda^{-1} W_i$, and $\Lambda\Lambda' = \tilde{A}^{-1}$. Therefore, the asymptotic covariance matrix of the GMM estimator (4.3.47), (4.3.48), is greater than (4.5.18), which is greater than (4.5.17) in the sense that the difference of the two covariance matrix is a positive semidefinite matrix.

When $\tilde{\Omega}^*$ is unknown, the asymptotic covariance matrix of (4.3.47) remains as (4.3.48). But the asymptotic covariance matrix of the feasible MDE is (4.5.16). Although the first term of (4.5.16) is smaller than (4.3.47), it is not clear that with the addition of the second term, it will remain smaller than (4.3.48). However, it is very likely so because of several factors. First, additional information due to the Δy_{i1} equation is utilized that 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-based approach uses the t instruments $(y_{i0}, y_{i1}, \ldots, y_{i,t-1})$. Third, the likelihood approach uses the condition that $E(H_i'\Omega^{*-1}\Delta\mathbf{u}_i^*) = \mathbf{0}$ and the GMM method uses the condition $E(\frac{1}{N}\sum_{i=1}^N W_i\Delta\mathbf{u}_i) = \mathbf{0}$. The grouping of observations in general will lead to a loss of information.[18]

Although both the GMM and the likelihood-based estimator are consistent, 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 $E(H_i'\Omega^{*-1}\Delta\mathbf{u}_i^*) = \mathbf{0}$, and hence is asymptotically unbiased independent of the way N or $T \longrightarrow \infty$ (Hsiao and Zhang 2013). 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 $\frac{1}{N}\sum_{i=1}^N \mathbf{q}_{it}(y_{it} - y_{i,t-1})$ and $\frac{1}{N}\sum_{i=1}^N q_{it}(u_{it} - u_{i,t-1})$, which is of order $\frac{1}{N}$, $O(\frac{1}{N})$. Therefore, when T is fixed and N is large, the GMM estimator is consistent and $\sqrt{N}(\hat{\gamma}_{\text{GMM}} - \gamma)$ is centered at 0. However, the number of moment conditions for the GMM (say (4.3.40)) 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 $\frac{T}{N} \longrightarrow c, 0 < c < \infty$ as $N \to \infty$, the effects of the correlations due to $\frac{1}{N}\sum_{i=1}^N \mathbf{q}_{it}(y_{it} - y_{i,t-1})$ and $\frac{1}{N}\sum_{i=1}^N \mathbf{q}_{it}(u_{it} - u_{i,t-1})$ get magnified. Alvarez and Arellano (2003) show that $\sqrt{NT}\hat{\gamma}_{\text{GMM}}$ has asymptotic bias equal to $-\sqrt{c}(1+\gamma)$. On the other hand, the likelihood-based estimator is asymptotically unbiased (Hsiao and Zhang 2013). In other words, the GMM estimator multiplied by the scale factor $\sqrt{NT}$ is not centered at $\sqrt{NT}\gamma$, but the likelihood based estimator is.[19] Whether an estimator is asymptotically biased or not has important implications in statistical inference because in hypothesis testing typically we normalize the estimated γ by the inverse of its standard error, which is equivalent to multiplying the estimator by the scale factor $\sqrt{NT}$. The Monte Carlo studies conducted by Hsiao and Zhang (2013) show that there is no size distortion for the MLE or simple IV ((4.3.33), (4.3.34)) but there are significant size distortions for GMM when N and T are of similar magnitude. For a nominal 5% significance level test, the actual size could be 40% when $\gamma = .5$ and 80% when $\gamma = .8$ for cases when N and T are of similar magnitude.

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

[19] 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 for the MLE (see Hsiao and Zhang 2013).

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Table 4.4. *Monte Carlo design*

Design number	γ	β	ϕ	θ	g	$R^2_{\Delta y}$	σ_ϵ
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: Hsiao, Pesaran, and Tahmiscioglu (2002, Table 1).

Hsiao, Pesaran, and Tahmiscioglu (2002) have conducted Monte Carlo studies to compare the performance of the IV of (4.3.34), the GMM of (4.3.47), the MLE, and the MDE when T is small and N is finite. They generate y_{it} by

$$y_{it} = \alpha_i + \gamma y_{i,t-1} + \beta x_{it} + u_{it}, \tag{4.5.20}$$

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 χ^2 with 2 degrees of freedom. The regressor x_{it} is generated according to

$$x_{it} = \mu_i + gt + \xi_{it} \tag{4.5.21}$$

where ξ_{it} follows an autoregressive moving average process

$$\xi_{it} - \phi\xi_{i,t-1} = \epsilon_{it} + \theta\epsilon_{i,t-1} \tag{4.5.22}$$

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 4.4 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 0. But the first 50 observations are discarded. The bias and root mean square error (RMSE) of various estimators of γ and β when $T = 5$ and $N = 50$ based on 2500 replications are reported in Tables 4.5 and 4.6, respectively. The results show that the bias of the MLE of γ as a percentage of the true value is smaller than 1 percent 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 method could run into weak IV problem (for an 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 to 20 percent in many

Table 4.5. *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).

cases and is larger than the bias of the MDE. The MLE also has the smallest RMSE followed by the MDE, then GMM. The IV has the largest RMSE.

4.5.4 Issues of Random versus Fixed-Effects Specification

The GMM or the MLE of the transformed likelihood function (4.5.6) or the MDE (4.5.10) is consistent and asymptotically normally distributed whether α_i are fixed or random. However, when α_i are random and uncorrelated with $\mathbf{x}_{it}$, the likelihood function of the form (4.3.19) uses the level variables whereas (4.5.6) 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 (4.5.6) yields estimators that will not be as efficient as

Table 4.6. *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).

the MLE of (4.3.19) when α_i are indeed random. However, if α_i are fixed or correlated with $\mathbf{x}_{it}$, the MLE of (4.3.19) yields an inconsistent estimator.

The transformed MLE or MDE is consistent under a more general data-generating process of $\mathbf{x}_{it}$ than the MLE of (4.3.19) or the GLS (4.3.30). For the Bhargava and Sargan (1983) MLE of the random effects model to be consistent, we will have to assume that the $\mathbf{x}_{it}$ are strictly exogenous and are generated from the same stationary process with common means ((4.3.21)). Otherwise, $E(y_{i0} \mid \mathbf{x}_i) = \mathbf{c}_i + \boldsymbol{\pi}_i'\mathbf{x}_i$, where $\mathbf{c}_i$ 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 $\mathbf{x}_{it}$ to be correlated with individual specific effects, α_i, and to have different means (or intercepts) (4.5.3). Therefore it appears that if one is not sure about the assumption of the effects, α_i, or the homogeneity assumption about the data-generating process of $\mathbf{x}_{it}$, one should work with the transformed likelihood function (4.5.6) or the MDE (4.5.10) despite the fact 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 $\mathbf{x}_{it}$ process under the assumption that α_i are random. Under the null of random effects and homogeneity of the $\mathbf{x}_{it}$ process, the MLE of the form (4.3.19) is asymptotically efficient. The transformed MLE of (4.5.6) is consistent, but not efficient. On the other hand, if α_i are fixed or $\mathbf{x}_{it}$ is not generated by a homogeneous process but satisfies (4.5.3), the transformed MLE of (4.5.6) is consistent, but the MLE of (4.3.19) is inconsistent. Therefore, a Hausman type test statistics (3.5.2) can be constructed by comparing the difference between the two estimators.

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

In previous sections we discussed estimation of the dynamic model

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i^* + u_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{4.6.1}$$

under the assumption that u_{it} are serially uncorrelated, where we now again let $\mathbf{x}_{it}$ 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 (4.6.1) to eliminate the individual effect α_i^*, and stacking all equations for a single individual, we have a system of $(T-1)$ equations,

$$\begin{aligned} y_{i2} - y_{i1} &= \gamma(y_{i1} - y_{i0}) + \boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1}) + (u_{i2} - u_{i1}), \\ y_{i3} - y_{i2} &= \gamma(y_{i2} - y_{i1}) + \boldsymbol{\beta}'(\mathbf{x}_{i3} - \mathbf{x}_{i2}) + (u_{i3} - u_{i2}), \\ &\vdots \\ y_{iT} - y_{i,T-1} &= \gamma(y_{i,T-1} - y_{i,T-2}) + \boldsymbol{\beta}'(\mathbf{x}_{iT} - \mathbf{x}_{i,T-1}) \\ &\quad + (u_{iT} - u_{i,T-1}), \quad i = 1, \ldots, N, \end{aligned} \tag{4.6.2}$$

We complete the system (4.6.2) with the identities

$$\begin{aligned} y_{i0} &= E^*(y_{i0} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) + [y_{i0} - E^*(y_{i0} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})] \\ &= a_0 + \sum_{t=1}^{T} \boldsymbol{\pi}_{0t}'\mathbf{x}_{it} + \epsilon_{i0} \end{aligned} \tag{4.6.3}$$

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

and

$$\begin{aligned} y_{i1} &= E^*(y_{i1} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) + [y_{i1} - E^*(y_{i1} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})] \\ &= a_1 + \sum_{t=1}^{T} \boldsymbol{\pi}'_{1t}\mathbf{x}_{it} + \epsilon_{i1}, \qquad i = 1, \ldots, N. \end{aligned} \tag{4.6.4}$$

where E^* denotes the projection operator. Because (4.6.3) and (4.6.4) are exactly identified equations, we can ignore them and apply the three-stage least squares (3SLS) or generalized 3SLS (see Chapter 5) to the system (4.6.2) only. With regard to the cross equation constraints in (4.6.2), 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 (4.6.2) 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 $\mathbf{x}_{it}$. However, to implement simultaneous-equations estimation methods to (4.6.2) without imposing restrictions on the serial-correlation structure of u_{it}, there must exist strictly exogenous variables $\mathbf{x}_{it}$ such that

$$E(u_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = 0. \tag{4.6.5}$$

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

4.7 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 $\boldsymbol{\beta} = \mathbf{0}$ in (4.1.2); thus the model becomes

$$y_{it} = \gamma y_{i,t-1} + v_{it}, \tag{4.7.1}$$

$$v_{it} = \alpha_i + \lambda_t + u_{it}, \qquad \begin{aligned} &i = 1, \ldots, N, \\ &t = 1, \ldots, T, \\ &y_{i0} \text{ observable.} \end{aligned} \tag{4.7.2}$$

The panel data estimators discussed in Sections 4.3–4.6 assume no presence of λ_t (i.e., $\lambda_t = 0 \ \forall \ t$). 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 (4.3.47) 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 (4.3.40) of 0. However, if λ_t are also present as in (4.7.2),

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

this condition is likely to be violated. To see this, taking the first difference of (4.7.1) yields

$$\begin{aligned}\Delta y_{it} &= \gamma \Delta y_{i,t-1} + \Delta v_{it}\\ &= \gamma \Delta y_{i,t-1} + \Delta \lambda_t + \Delta u_{it},\\ i &= 1, \ldots, N,\\ t &= 2, \ldots, T.\end{aligned} \tag{4.7.3}$$

Although under the assumption λ_t are independently distributed over t with mean 0,

$$E(y_{i,t-j} \Delta v_{it}) = 0 \quad \text{for } j = 2, \ldots, t, \tag{4.7.4}$$

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

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta v_{it} = \frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta\lambda_t\\ + \frac{1}{N}\sum_{i=1}^{N} y_{i,t-j}\Delta u_{it}\end{aligned} \tag{4.7.5}$$

converges to $\bar{y}_{t-j}\Delta\lambda_t$, which in general is not equal to 0, in particular, if y_{it} has mean different from 0,[20] 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} is independent normal and fixed y_{i0}, the MLE of the FE model (4.7.1) is equal to:

$$\tilde{\gamma}_{cv} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y^*_{i,t-1} y^*_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y^{*2}_{i,t-1}}, \tag{4.7.6}$$

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 $\bar{y}_t$, $\bar{y}_{i,-1}$, $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_t$, $\mathbf{x}^*_{it}$, 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,

$$y^*_{it} = \gamma y^*_{i,t-1} + v^*_{it}, \tag{4.7.7}$$

and then applying the least-squares estimator of (4.7.7).

The probability limit of $\tilde{\gamma}_{cv}$ is identical to the case where $\lambda_t \equiv 0$ for all t (4.2.8) (Hahn and Moon 2006; Hsiao and Tahmiscioglu 2008). The bias

[20] For instance, if y_{it} is also a function of exogenous variables as (4.1.2), where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

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 (4.3.8), because $Ey_{i0}v_{it} \neq 0$, we either have to write (4.7.1) conditional on y_{i0} or to complete the system (4.7.1) by deriving the marginal distribution of y_{i0}. By continuous substitutions, we have

$$\begin{aligned} y_{i0} &= \frac{1-\gamma^m}{1-\gamma}\alpha_i + \sum_{j=0}^{m-1} \lambda_{i,-j}\gamma^j + \sum_{j=0}^{m-1} \epsilon_{i,-j}\gamma^j \\ &= v_{i0}, \end{aligned} \tag{4.7.8}$$

assuming the process started at period $-m$.

Under (4.3.8), $Ey_{i0} = Ev_{i0} = 0$, $\text{var}(y_{i0}) = \sigma_0^2$, $E(v_{i0}v_{it}) = \frac{1-\gamma^m}{1-\gamma}\sigma_\alpha^2 = c^*$, $Ev_{it}v_{jt} = d^*$. Stacking the $T+1$ time series observations for the ith individual into a vector, $\mathbf{y}_i = (y_{i0}, \ldots, y_{i,T})'$ and $\mathbf{y}_{i,-1} = (0, y_{i1}, \ldots, y_{i,T-1})'$, $\mathbf{v}_i = (v_{i0}, \ldots, v_{iT})'$. Let $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$, $\mathbf{y}_{-1} = (\mathbf{y}_{1,-1}', \ldots, \mathbf{y}_{N,-1}')$, $\mathbf{v} = (\mathbf{v}_1', \ldots, \mathbf{v}_N')'$, then

$$\mathbf{y} = \mathbf{y}_{-1}\gamma + \mathbf{v}, \tag{4.7.9}$$

$$\begin{aligned} E\mathbf{v} &= \mathbf{0}, \\ E\mathbf{v}\mathbf{v}' &= \sigma_u^2 I_N \otimes \begin{pmatrix} \omega & \mathbf{0}' \\ \mathbf{0} & I_T \end{pmatrix} + \sigma_\alpha^2 I_N \otimes \begin{pmatrix} 0 & c^*\mathbf{e}_T' \\ c^*\mathbf{e}_T & \mathbf{e}_T\mathbf{e}_T' \end{pmatrix} \\ &\quad + \sigma_\lambda^2 \mathbf{e}_N \mathbf{e}_N \otimes \begin{pmatrix} d^* & \mathbf{0}' \\ \mathbf{0} & I_T \end{pmatrix}, \end{aligned} \tag{4.7.10}$$

where $\otimes$ denotes the Kronecker product, and ω denotes the variance of v_{i0} divided by σ_u^2. The system (4.7.9) has a fixed number of unknowns $(\gamma, \sigma_u^2, \sigma_\alpha^2, \sigma_\lambda^2, \sigma_0^2, c^*, d^*)$ as N and T increase. Therefore, the MLE (or quasi-MLE or GLS of (4.7.9)) 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 (4.7.3). 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

$$\Delta y_t = \gamma \Delta y_{t-1} + \Delta\lambda_t + \Delta u_t, \tag{4.7.11}$$

where $\Delta u_t = \frac{1}{N}\sum_{i=1}^{N} \Delta u_{it}$. Taking the deviation of (4.7.3) from (4.7.11) yields

$$\Delta y^*_{it} = \gamma \Delta y^*_{i,t-1} + \Delta u^*_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 2, \ldots, T, \end{array} \tag{4.7.12}$$

where $\Delta y^*_{it} = (\Delta y_{it} - \Delta y_t)$ and $\Delta u^*_{it} = (\Delta u_{it} - \Delta u_t)$. The system (4.7.12) no longer involves α_i and λ_t.

Since

$$E[y_{i,t-j}\Delta u^*_{it}] = 0 \quad \text{for} \quad \begin{array}{l} j = 2, \ldots, t, \\ t = 2, \ldots, T. \end{array} \tag{4.7.13}$$

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

$$E(W_i \Delta\tilde{\mathbf{u}}^*_i) = \mathbf{0}, \tag{4.7.14}$$

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

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

and $\mathbf{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2})'$, $t = 2, 3, \ldots, T$. Following Arellano and Bond (1991), we can propose a generalized method of moments (GMM) estimator,[21]

$$\begin{aligned} \tilde{\gamma}_{\text{GMM}} = {} & \left\{\left[\frac{1}{N}\sum_{i=1}^{N}\Delta\tilde{\mathbf{y}}^{*'}_{i,-1}W_i'\right]\hat{\Psi}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}W_i\Delta\tilde{\mathbf{y}}^*_{i,-1}\right]\right\}^{-1} \\ & \cdot\left\{\left[\frac{1}{N}\sum_{i=1}^{N}\Delta\tilde{\mathbf{y}}^{*'}_{i,-1}\mathbf{W}_i'\right]\hat{\Psi}^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}W_i\Delta\tilde{\mathbf{y}}^*_i\right]\right\}, \end{aligned} \tag{4.7.15}$$

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

$$\hat{\Psi} = \frac{1}{N^2}\left[\sum_{i=1}^{N}W_i\Delta\hat{\tilde{\mathbf{u}}}^*_i\right]\left[\sum_{i=1}^{N}W_i\Delta\hat{\tilde{\mathbf{u}}}^*_i\right]' \tag{4.7.16}$$

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

[21] 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), Blundell and Bond (1998).

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

$$\text{asym. Cov}\,(\tilde{\gamma}_{\text{GMM}}) = \left\{ \left[\sum_{i=1}^{N} \Delta \tilde{\mathbf{y}}^{*'}_{i,-1} W_i' \right] \hat{\Psi}^{-1} \left[\sum_{i=1}^{N} W_i \Delta \tilde{\mathbf{y}}^{*}_{i,-1} \right] \right\}^{-1}. \quad (4.7.17)$$

To implement the likelihood approach, we need to complete the system (4.7.12) by deriving the marginal distribution of Δy^*_{i1} through continuous substitution,

$$\begin{aligned} \Delta y^*_{i1} &= \sum_{j=0}^{m-1} \Delta u^*_{i,1-j} \gamma^j \\ &= \Delta u^*_{i1}, \qquad i = 1, \ldots, N. \end{aligned} \quad (4.7.18)$$

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

$$\Delta \mathbf{y}^*_i = \Delta \mathbf{y}^*_{i,-1} \gamma + \Delta u^*_i, \quad (4.7.19)$$

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:

$$\hat{\gamma}_{\text{GLS}} = \left[\sum_{i=1}^{N} \Delta \mathbf{y}^{*'}_{i,-1} \tilde{A}^{*-1} \Delta \mathbf{y}^*_{i,-1} \right]^{-1} \left[\sum_{i=1}^{N} \Delta \mathbf{y}^{*'}_{i,-1} \tilde{A}^{*-1} \Delta \mathbf{y}^*_i \right], \quad (4.7.20)$$

where $\tilde{A}^*$ is a $T \times T$ matrix of the form,

$$\tilde{A}^* = \begin{bmatrix} \omega & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & . & . \\ 0 & -1 & 2 & -1 & \ldots & . & . \\ . & . & . & . & . & 2 & -1 \\ 0 & . & . & . & & -1 & 2 \end{bmatrix}. \quad (4.7.21)$$

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

$$\text{Var}\,(\hat{\gamma}_{\text{GLS}}) = \sigma_u^2 \left[\sum_{i=1}^{N} \Delta \mathbf{y}^{*'}_{i,-1} \tilde{A}^{*-1} \Delta \mathbf{y}^*_{i,-1} \right]^{-1}. \quad (4.7.22)$$

Remark 4.7.1: The GLS with $\Delta \boldsymbol{\lambda}$ present is basically of the same form as the GLS without the time-specific effects (i.e., $\Delta \boldsymbol{\lambda} = \mathbf{0}$) (Hsiao, Pesaran, and Tahmiscioglu 2002), (4.5.10). However,there is an important difference between the two. The estimator (4.7.20) uses $\Delta y^*_{i,t-1}$ as the regressor for the equation Δy^*_{it} (4.7.19), does not use $\Delta y_{i,t-1}$ as the regressor for the equation Δy_{it} ((4.7.3)). If there are indeed common shocks that affect all the cross-sectional units, then the estimator (4.5.10) is inconsistent while (4.7.20) is consistent (for details, see Hsiao and Tahmiscioglu 2008). Note also that even though when there are no time-specific effects, (4.7.20) remains consistent, although it will not be as efficient as (4.5.10).

Remark 4.7.2: The estimator (4.7.20) and the estimator (4.7.15) remain consistent and asymptotically normally distributed when the effects are random because the transformation (4.7.11) effectively removes the individual- and time-specific effects from the specification. However, if the effects are indeed random, and uncorrelated with $\mathbf{x}_{it}$ then the MLE or GLS of (4.7.7) is more efficient.

Remark 4.7.3: The GLS (4.7.20) assumes known ω. If ω is unknown, one may substitute it by a consistent estimator $\hat{\omega}$, and 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 $\lim\,(\frac{N}{T}) = c > 0$, then the FGLS will be asymptotically equivalent to the GLS (Hsiao and Tahmiscioglu 2008).

Remark 4.7.4: The MLE or GLS of (4.7.20) can also be derived by treating $\Delta\lambda_t$ as fixed parameters in the system (4.7.3). Through continuous substitution, we have

$$\Delta y_{i1} = \lambda_1^* + \Delta\tilde{u}_{i1}, \tag{4.7.23}$$

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

$$\begin{aligned} \underset{NT\times 1}{\Delta\mathbf{y}} = \begin{pmatrix} \Delta\mathbf{y}_1 \\ \vdots \\ \Delta\mathbf{y}_N \end{pmatrix} &= \begin{pmatrix} \Delta\mathbf{y}_{1,-1} \\ \vdots \\ \Delta\mathbf{y}_{N,-1} \end{pmatrix} \gamma + (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda} + \begin{pmatrix} \Delta\mathbf{u}_1 \\ \vdots \\ \Delta\mathbf{u}_N \end{pmatrix} \\ &= \Delta\mathbf{y}_{-1}\gamma + (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda} + \Delta\mathbf{u}, \end{aligned} \tag{4.7.24}$$

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

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

$$\begin{aligned} \log L = &-\frac{NT}{2}\log\sigma_u^2 - \frac{N}{2}\log|\tilde{A}^*| - \frac{1}{2\sigma_u^2}[\Delta\mathbf{y} - \Delta\mathbf{y}_{-1}\gamma - (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda}]' \\ &\cdot (I_N \otimes \tilde{A}^{*-1})[\Delta\mathbf{y} - \Delta\mathbf{y}_{-1}\gamma - (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda}]. \end{aligned} \tag{4.7.25}$$

Taking the partial derivative of (4.7.25) with respect to $\Delta\boldsymbol{\lambda}$ and solving for $\Delta\boldsymbol{\lambda}$ yields

$$\Delta\hat{\boldsymbol{\lambda}} = (N^{-1}\mathbf{e}_N' \otimes I_T)(\Delta\mathbf{y} - \Delta\mathbf{y}_{-1}\gamma). \tag{4.7.26}$$

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Substituting (4.7.26) into (4.7.25) yields the concentrated log-likelihood function.

$$\begin{aligned}\log L_c = &-\frac{NT}{2}\log\sigma_\epsilon^2 - \frac{N}{2}\log\ \mid \tilde{A}^* \mid \\ &- \frac{1}{2\sigma_\epsilon^2}(\Delta\mathbf{y}^* - \Delta\mathbf{y}_{-1}^*\gamma)'(I_N\otimes\tilde{A}^{*-1})(\Delta\mathbf{y}^* - \Delta y_{-1}^*\gamma).\end{aligned} \tag{4.7.27}$$

Maximizing (4.7.27) conditional on ω yields (4.7.20).

Remark 4.7.5: When γ approaches 1 and σ_α^2 is large relative to σ_u^2, the GMM estimator of the form (4.3.47) 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 4.7.6: Hahn and Moon (2006) propose a bias-corrected estimator as

$$\tilde{\gamma}_b = \tilde{\gamma}_{cv} + \frac{1}{T}(1+\tilde{\gamma}_{cv}). \tag{4.7.28}$$

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

$$\sqrt{NT}(\tilde{\gamma}_b - \gamma) \Longrightarrow N(0, 1-\gamma^2). \tag{4.7.29}$$

The limited Monte Carlo studies conducted by Hsiao and Tahmiscioglu (2008) to investigate the finite sample properties of the feasible GLS (FGLS), GMM, and bias-corrected (BC) estimator of Hahn and Moon (2006) have shown that in terms of bias and RMSEs, 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 $\gamma = 0.8$). However, if N and T are of similar magnitude, $\frac{T}{N} = c \neq 0$, there are significant size distortion (e.g., Hsiao and Zhang 2013). The BC has significant size distortion, presumably because of the use of asymptotic covariance matrix, which is significantly downward biased in the finite sample.

Remark 4.7.7: Hsiao and Tahmiscioglu (2008) also compared 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 RMSEs 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 account both individual- and time-specific effects when T is finite.

APPENDIX 4A: DERIVATION OF THE ASYMPTOTIC COVARIANCE MATRIX OF FEASIBLE MDE

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

$$\sqrt{N}(\hat{\boldsymbol{\psi}}_{\text{MDE}} - \boldsymbol{\psi}) = \left(\frac{1}{N}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} H_i\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} \Delta\mathbf{u}_i^*\right). \quad (4A.1)$$

When $N \longrightarrow \infty$

$$\frac{1}{N}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} H_i \longrightarrow \frac{1}{N}\sum_{i=1}^{N} H_i' \tilde{\Omega}^{*-1} H_i \quad (4A.2)$$

but

$$\begin{aligned} \frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i' \hat{\tilde{\Omega}}^{*-1} \Delta\mathbf{u}_i^* \simeq{}& \frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i' \tilde{\Omega}^{*-1} \Delta\mathbf{u}_i^* \\ &+ \left[\frac{1}{N}\sum_{i=1}^{N} H_i' \left(\frac{\partial}{\partial h}\tilde{\Omega}^{*-1}\right) \Delta\mathbf{u}_i^*\right] \cdot \sqrt{N}(\hat{h} - h), \end{aligned} \quad (4A.3)$$

where the right-hand side follows from taking a Taylor series expansion of $\hat{\tilde{\Omega}}^{*-1}$ around $\tilde{\Omega}^{*-1}$. By (4.5.8),

$$\begin{aligned} \frac{\partial}{\partial h}\tilde{\Omega}^{*-1} ={}& \frac{-T}{[1 + T(h-1)]^2}\tilde{\Omega}^{*-1} \\ &+ \frac{1}{1 + T(h-1)]} \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & T-1 & \dots & 2 & 1 \\ \ddots & \ddots & & \ddots & \ddots \\ . & 2 & \dots & 2(T-2) & T-2 \\ 0 & 1 & \dots & T-2 & T-1 \end{bmatrix}. \end{aligned} \quad (4A.4)$$

We have

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

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

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

$$\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{y}'_{i,-1}\begin{bmatrix} T-1 & \dots & 1 \\ & \ddots & \vdots \\ 2 & & T-2 \\ 1 & & T-1 \end{bmatrix}\Delta\mathbf{u}_i^*$$

$$\longrightarrow [\gamma^{T-2}+2\gamma^{T-3}+\cdots+(T-1)]\sigma_u^2.$$

Since 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

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

$$-\begin{bmatrix}0\\ \mathbf{0}\\ 1\\ \mathbf{0}\end{bmatrix}\frac{[\gamma^{T-2}+2\gamma^{T-3}+\cdots+(T-1)]}{[1+T(h-1)]\sigma_u^2} \tag{4A.5}$$

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

with the asymptotic covariance matrix equal to (4.5.16).

APPENDIX 4B: 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-dimensional 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 such as 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-sectional 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.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

(b) *Diagonal Path Limits*. This approach allows the two indexes, N and T, to pass to infinity along a specific diagonal path in the two-dimensional 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 to get definitive results.

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

$$X_{N,T} = \frac{1}{k_N} \sum_{i=1}^{N} Y_{i,T}, \tag{4B.1}$$

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

$$Y_{i,T} = \frac{1}{d_T} \sum_{t=1}^{T} f(Z_{i,t}), \tag{4B.2}$$

for the h-component independently, identically distributed random vectors, $Z_{i,t}$ with finite 4th 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, which will allow for uniformity in the convergence arguments. Phillips and Moon (1999) give the 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

$$\limsup_{T\to\infty} \sup_{N} P\{\| X_{N,T} - X_N \| > \epsilon\} = 0 \text{ for every } \epsilon > 0. \tag{4B.3}$$

(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

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

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

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

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) $\lim\sup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_{i,T} \parallel < \infty$;
(ii) $\lim\sup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} \parallel EY_{i,T} - EY_i \parallel = 0$;
(iii) $\lim\sup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_{i,T} \parallel 1 \left\{\parallel Y_{i,T} \parallel > N\epsilon\right\} = 0 \; \forall \epsilon > 0$;
(iv) $\lim\sup_{N} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_i \parallel 1 \{\parallel Y_i \parallel > N\epsilon\} = 0 \; \forall \epsilon > 0$,

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 (4B.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 N and T tend to infinity. Moreover, even in the case that an estimator is inconsistent for fixed T and large N case, say, the CV estimator for the fixed effects dynamic model (4.2.1), 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

$$X_{N,T} = \frac{1}{N} \sum_{i=1}^{N} Y_{i,T}. \tag{4B.5}$$

Suppose $Y_{i,T}$ is independently, identically distributed across i for each T with $E(Y_{i,T}) = \frac{1}{\sqrt{T}} b$ and 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 Var $(Y_{i,T})$ is bounded, by a law of large numbers, X_N converges to 0 in probability as $N \to \infty$. Since (4B.5) satisfies (4B.3), the sequential limit is equal to the joint limit as $N, T \to \infty$. This can be clearly seen by writing

$$\begin{aligned} X_{N,T} &= \frac{1}{N} \sum_{i=1}^{N} [Y_{i,T} - E(Y_{i,T})] + \frac{1}{N} \sum_{i=1}^{N} E(Y_{i,T}) \\ &= \frac{1}{N} \sum_{i=1}^{N} [Y_{i,T} - E(Y_{i,T})] + \frac{b}{\sqrt{T}}. \end{aligned} \tag{4B.6}$$

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

$$E(X_{N,T}^2) = \frac{1}{N} \text{ Var } (Y_{i,T}) + \frac{b^2}{T} \longrightarrow 0 \tag{4B.7}$$

as $N, T \longrightarrow \infty$. Equation (4B.7) implies that $X_{N,T}$ converges to 0 jointly as $N, T \longrightarrow \infty$.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Alternatively, if we let

$$X_{N,T} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Y_{i,T}. \tag{4B.8}$$

The sequential limit would imply $X_{N,T}$ is asymptotically normally distributed with $E(X_{N,T}) = 0$. However, under (4B.8) the condition (4B.3) is violated. The joint limit would have

$$EX_{N,T} = \frac{\sqrt{c}}{N} \sum_{i=1}^{N} b \longrightarrow \sqrt{c}b, \tag{4B.9}$$

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 2002).

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

$$y = E(y \mid x) + v = \beta x + v. \tag{4B.10}$$

If v_t is stationary (or $I(0)$ process), the least-squares estimator of β, $\hat{\beta}$, 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 the y and x is bivariate normally distributed as $N(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}, \tag{4B.11}$$

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

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{yt} \\ u_{xt} \end{pmatrix}, \tag{4B.12}$$

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

$$\text{plim}\ \hat{\beta} = \Omega_{yx}\Omega_{xx}^{-1}, \tag{4B.13}$$

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

which can be viewed as 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 x_t defined by

$$\begin{aligned}\Omega &= \lim_{T\to\infty} E\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{u}_t\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{u}_t'\right)\right] \\ &= \sum_{\ell=-\infty}^{\infty} E(\mathbf{u}_0\mathbf{u}_\ell') = \begin{pmatrix}\Omega_{yy} & \Omega_{yx}\\ \Omega_{xy} & \Omega_{xx}\end{pmatrix}.\end{aligned} \tag{4B.14}$$

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) extend this concept of a long-run average relation among cross-sectional units further

$$\beta = E(\Omega_{yx,i})(E\Omega_{xx,i})^{-1} = \Omega_{yx}\Omega_{xx}^{-1}. \tag{4B.15}$$

and show that the least-squares estimator converges to (4B.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:

$$\begin{aligned} y_t &= f_t + w_t, \\ x_t &= f_t, \end{aligned} \tag{4B.16}$$

with

$$\begin{pmatrix} w_t \\ f_t \end{pmatrix} = \begin{pmatrix} w_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} u_{wt} \\ u_{ft} \end{pmatrix}, \tag{4B.17}$$

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 w_t 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 a long-run correlation between y_t and x_t, and this is what is measured by the regression coefficient β in (4B.13).

Phillips and Moon (1999, 2000) show that for large N and T panels, the regression coefficient β coverages to such a 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 (4B.13) if only $T \to \infty$.

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press

Therefore, if we define spurious regression as yielding nonzero β for the two independent variables, then contrary to the case of time series regression of involving two linearly independent I(1) variables (Phillips 1986) the issue of spurious regression will not arise for the panel estimates of $N \rightarrow \infty$ (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 2010).

https://doi.org/10.1017/CBO9781139839327.007 Published online by Cambridge University Press