CHAPTER 13

Variable Coefficients Models

13.1 INTRODUCTION

So far our discussion has been based on the assumption that through decomposing the impacts of omitted variables into individual-specific and time-specific effects, whether in additive form (Chapters 2–9) or in multiplicative form (Chapter 10), or in some other form (Chapter 11), we are able to control the correlations between the included explanatory variables and omitted variables. But there are instances in which these approaches are not able to completely control the correlations between the included explanatory variables and omitted (observed or unobserved) variables representing changing economic structures or unobserved different socioeconomic and demographic background factors. Under these circumstances, the parameters of the included variables may vary over time and/or may differ for different cross-sectional units. For instance, in farm production, it is likely that variables not included in the specification could also impact the marginal productivity of fertilizer used, such as soil characteristics (e.g., slope, soil fertility, water reserve, etc.) or climate conditions. So are empirical studies of economic growth. The per capita output growth rates are assumed to depend over a common horizon on two sets of variables. One set of variables consists of initial per capita output, savings, and population growth rates, variables that are suggested by the Solow growth model. The second set of variables consists of control variables that correspond to whatever additional determinants of growth a researcher wishes to examine (e.g., Durlauf 2001; Durlauf and Quah 1999). However, there is nothing in growth theory which would lead one to think that the marginal effect of a change in high school enrollment percentages on the per capita growth of the U.S. should be the same as the marginal effect on a country in sub-Saharan Africa. Had all these factors characterizing the differences in growth been taken into account in the specification, a common slope coefficients model may seem reasonable. However, these variables could be unavailable or could be difficult to observe with precision. Moreover, a model is not a mirror; it is a simplification of the real world to capture the relationships among the essential variables. As a matter of fact, any parsimonious regression will necessarily leave out many factors that would from the perspective of economic theory be likely to affect the parameters of the included variables (e.g., Canova 1999; Durlauf and Johnson 1995). In these situations, varying parameter models appear to be more capable of capturing the unobserved heterogeneity than a model that puts all the impact of omitted variables in the error terms.

In a study on investment expenditures of 60 small and middle-sized firms in capital-goods-producing industries from 1935 to 1955, excluding the war years (1942–1945), Kuh (1963) reported that in a majority of cases he investigated, the common intercept and common slope coefficients for all firms, as well as the variable-intercept common-slope hypotheses, were rejected (see Hsiao 2014, Tables 2.3 and 2.4). Similar results were found by Swamy (1970), who used the annual data of 11 U.S. corporations from 1935 to 1954 to fit the Grunfeld (1958) investment functions. His preliminary test of variable-intercept-common slope coefficients against the variable-intercept and variable slope coefficients for the value of a firm's outstanding shares at the beginning of the year and its beginning-of-year capital stock yielded an F value of 14.4521. That is well above the 5% value of an F distribution with 27 and 187 degrees of freedom.[1]

When an investigator is mainly interested in the fundamental relationship between the outcome variable and a set of primary conditional variables, either for ease of analysis or because of the unavailability of the secondary conditional variables, it would seem reasonable to allow variations in parameters across cross-sectional units and/or over time as a means to take account of the interindividual and/or interperiod heterogeneity. A single-equation model in its most general form can be written as

$$y_{it} = \sum_{k=1}^{K} \beta_{kit} x_{kit} + u_{it}, \quad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{13.1.1}$$

where, in contrast to previous chapters, we no longer treat the intercept differently from other explanatory variables and let $x_{1it} = 1$. However, if all the coefficients are treated as fixed and different for different cross-sectional units in different time periods, there are NKT parameters with only NT observations. Obviously, there is no way we can obtain any meaningful estimates of β_{kit}. We are thus led to search for an approach that allows β_{kit} to vary across individuals and over time but can still make statistical inference based from a panel data set.

One possibility would be to introduce dummy variables into the model that would indicate differences in the coefficients across individual units and/or over time, that is, to develop an approach similar to the least squares dummy-variable approach. In the case in which only cross-sectional differences are present and there is no variation of the parameters in time, this approach is equivalent to postulating a separate regression for each cross-sectional unit

$$y_{it} = \boldsymbol{x}_{it}' \boldsymbol{\beta}_i + u_{it}, \quad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{13.1.2}$$

where $\boldsymbol{\beta}_i$ and $\boldsymbol{x}_{it}$ are $K \times 1$ vectors of parameters and explanatory variables.[2]

Alternatively, each regression coefficient can be viewed as a random variable with a probability distribution (e.g., Hurwicz 1950; Klein 1953; Theil and Mennes 1959; Zellner 1966). The random-coefficients specification reduces the number of parameters to be estimated substantially, while still allowing the coefficients to differ from unit to unit and/or from time to time. Depending on the type of assumption about the parameter variation, it can be further classified into one of two categories: stationary and nonstationary random-coefficients models.

[1] See Mehta, Narasimham, and Swamy (1978) for another example showing that using error-components formulation to account for heterogeneity does not always yield economically meaningful results.

[2] Alternatively, we can postulate a separate regression for each time period; so $y_{it} = \boldsymbol{x}_{it}' \boldsymbol{\beta}_t + u_{it}$.

Stationary random-coefficients models regard the coefficients as having constant means and variance–covariances. Namely, the $K \times 1$ vector of parameters $\boldsymbol{\beta}_{it}$ is specified as

$$\boldsymbol{\beta}_{it} = \bar{\boldsymbol{\beta}} + \boldsymbol{\xi}_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{13.1.3}$$

where $\bar{\boldsymbol{\beta}}$ is a $K \times 1$ vector of constants, and $\boldsymbol{\xi}_{it}$ is a $K \times 1$ vector of stationary random variables with zero means and constant variance–covariances. For this type of model, Arellano and Bonhomme (2012) provided general conditions for identifying distributional characteristics. In this chapter, we assume the mean and covariance matrix of the random-coefficients model are identifiable and discuss only (i) estimating the mean coefficient vector $\bar{\boldsymbol{\beta}}$, (ii) predicting each individual component $\boldsymbol{\xi}_{it}$, (iii) estimating the dispersion of the individual-parameter vector, and (iv) testing the hypothesis that the variances of $\boldsymbol{\xi}_{it}$ are zero.

The nonstationary random-coefficients models do not regard the coefficient vector as having constant mean or variance. Changes in coefficients from one observation to the next can be the result of the realization of a nonstationary stochastic process or can be a function of exogenous variables. In this case we are interested in (i) estimating the parameters characterizing the time-evolving process, (ii) estimating the initial value and the history of parameter realizations, (iii) predicting future evolutions, and (iv) testing the hypothesis of random variation.

Because of the computational complexities of random-coefficients models or incidental parameter issues when varying parameters are treated as unknown constants, variable coefficients models have not gained as wide an acceptance in empirical work as has the variable-intercept model. However, that does not mean that there is less need for taking account of parameter heterogeneity in pooling the data. In this chapter we shall introduce some of the popular single-equation varying coefficients models. We shall first discuss models in which the variation of coefficients is independent of the variation of exogenous explanatory variables. Single-equation models with coefficients varying over individuals are discussed in Section 13.2. In Section 13.3 we discuss models with coefficients varying over individuals and time. Section 13.4 proposes a mixed fixed and random-coefficients model as a unifying framework to various approaches of controlling unobserved heterogeneity. Section 13.5 discusses issues of dynamic models. Section 13.6 provides an analysis of liquidity constraints and firm investment expenditure. Section 13.7 concerns models with time-evolving coefficients. Models with coefficients that are functions of other exogenous variables will be discussed in Section 13.8. Models in which the variation of coefficients are correlated with the explanatory variables are covered in Section 13.9. For random-coefficients models with heteroscedasticity, see Bresson et al. (2011); with cross-correlated residuals, see Bresson and Hsiao (2011); and for simultaneous-equations models with random coefficients, see Chow (1983), Kelejian (1977), and Raj and Ullah (1981). Further discussion of this subject can also be found in Amemiya (1983), Chow (1983), Hsiao and Pesaran (2008), Judge et al. (1980), and Raj and Ullah (1981).

13.2 COEFFICIENTS THAT VARY OVER CROSS-SECTIONAL UNITS

When regression coefficients are viewed as invariant over time but varying from one unit to another, we can write the model as

$$\begin{aligned} y_{it} &= \sum_{k=1}^{K} \beta_{ki} x_{kit} + u_{it} \\ &= \sum_{k=1}^{K} (\bar{\beta}_k + \alpha_{ki}) x_{kit} + u_{it}, \qquad \begin{array}{l} i = 1, \dots, N, \\ t = 1, \dots, T, \end{array} \end{aligned} \tag{13.2.1}$$

where $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \dots, \bar{\beta}_K)'$ can be viewed as the common-mean-coefficient vector and $\boldsymbol{\alpha}_i = (\alpha_{1i}, \dots, \alpha_{Ki})'$ as the individual deviation from the common mean $\bar{\boldsymbol{\beta}}$. If individual observations are heterogeneous or the performance of individual units from the database is of interest, then $\boldsymbol{\alpha}_i$ are treated as fixed constants. If conditional on x_{kit}, individual units can be viewed as random draws from a common population, or if the population characteristics are of interest, then α_{ki} are generally treated as random variables having zero means and constant variances and covariances.

13.2.1 Fixed-Coefficient Model

13.2.1.1 Complete Heterogeneity

When $\boldsymbol{\beta}_i$ are treated as fixed and different constants, we can stack the NT observations in the form of the Zellner (1962) seemingly unrelated regression model

$$\begin{aligned} \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix} &= \begin{bmatrix} X_1 & & & \mathbf{0} \\ & X_2 & & \\ & & \ddots & \\ \mathbf{0} & & & X_N \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_N \end{bmatrix}. \\ &= \tilde{X}\boldsymbol{\beta} + \boldsymbol{u} \end{aligned} \tag{13.2.2}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ are $T \times 1$ vectors of $(y_{i1}, \dots, y_{iT})'$ and $(u_{i1}, \dots, u_{iT})'$, X_i is the $T \times K$ matrix of the time-series observations of the ith individual's explanatory variables with the tth row equal to $\boldsymbol{x}_{it}'$, $\tilde{X}$ is the $NT \times NK$ block diagonal matrix with the ith block being X_i, and $\boldsymbol{\beta}$ is an $NK \times 1$ vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_N')'$. If the covariances between different cross-sectional units are not zero, $E\boldsymbol{u}_i\boldsymbol{u}_j' \neq \mathbf{0}$, the GLS estimator of $(\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_N')$ is more efficient than the single-equation estimator of $\boldsymbol{\beta}_i$ for each cross-sectional unit. If X_i are identical for all i or $E\boldsymbol{u}_i\boldsymbol{u}_i' = \sigma_i^2 I$ and $E\boldsymbol{u}_i\boldsymbol{u}_j' = \mathbf{0}$ for $i \neq j$, the GLS estimator for $(\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_N')$ is the same as applying the least squares separately to the time-series observations of each cross-sectional unit.

13.2.1.2 Group Heterogeneity

When N is large, it is neither feasible nor desirable to let $\boldsymbol{\beta}_i$ be different for different i. An alternative to individual heterogeneity is to assume group heterogeneity in place of complete heterogeneity. In other words, the sample is assumed to compose of G heterogeneous groups. Individuals belonging to a particular group all have the same response function (e.g., Lin and Ng 2012; Su, Shi, and Phillips 2016),

$$y_{it,g} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_g + u_{it,g}, \text{ for } i \in \text{ group } g. \tag{13.2.3}$$

If the grouping is known from some external consideration (e.g., Bester and Hansen 2012), estimation of (13.2.3) can proceed following the Zellner (1962) seemingly unrealted regression framework. However, if such external information is not available, two issues arise: (i) how to determine the number of groups, G; and (ii) how to identify the group an individual belongs. Following the idea of LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani 1996) under the assumption that the number of groups, G, is known, Su, Shi, and Phillips (SSP) (2016) suggest a modified penalized least squares approach,

$$\text{Min } Q^G = Q + \frac{a}{N}\sum_{i=1}^{N}\prod_{g=1}^{G} \parallel \boldsymbol{\beta}_i - \boldsymbol{\beta}_g \parallel , \tag{13.2.4}$$

to simultaneously classify individuals into groups and estimate $\boldsymbol{\beta}_g$, where $\parallel \cdot \parallel$ denotes the Frobenius norm, $\parallel A \parallel = [tr\ AA']^{1/2}$,

$$Q = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x}_{it}'\boldsymbol{\beta}_i)^2$$

and a is a tuning constant. SSP show that minimizing (13.2.4) achieves classification of individuals into groups and consistent estimation of $\boldsymbol{\beta}_g$ in a single step when N and T are large. SSP also propose to select the number of groups, G, by minimizing

$$\log \hat{\sigma}_G^2 + cGK, \tag{13.2.5}$$

where

$$\hat{\sigma}_G^2 = \frac{1}{NT}\sum_{g=1}^{G}\sum_{i\in g}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}_{g,G})^2$$

is the estimated average residual sum of squares based on G-group estimates of (13.2.3), $\hat{\boldsymbol{\beta}}_{g,G}$, and c is a tuning constant.

One can use information criterion such as AIC (Akaike 1973) or BIC (Schwarz 1978) to select the number of groups. If time series dimension is not small, one can also use Lu and Su (2020) and Lu et al. (2020) cross-validation to determine the number of groups. Suppose there are M groups being considered. For the model based on G groups, let $\hat{\boldsymbol{\beta}}_{gG}$ be the estimator $\hat{\boldsymbol{\beta}}_{gG,t}$ with the (i,t)th observation deleted from the sample for $g = 1, \ldots, G$, and $\hat{y}_{it,G} = \boldsymbol{x}_{it}'\ \hat{\boldsymbol{\beta}}_{gG,t}$. Then the cross-validation function for models based on G groups is

$$CV(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{y}_{it,G})^2. \tag{13.2.6}$$

Then choose the number of groups, G^*, as

$$G^* = \min_G CV(G), \quad G = 1, \ldots, M. \tag{13.2.7}$$

13.2.2 Random-Coefficients Model

13.2.2.1 The Model

When $\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i$ are treated as random with a common mean $\bar{\boldsymbol{\beta}}$, we assume that the distribution of $\boldsymbol{\alpha}_i$ conditional on the explanatory variables stays constant over time.

Hence, its distribution, in particular its variance–covariance, is identifiable.[3] Swamy (1970) assumed that[4]

$$
\begin{aligned}
&E\boldsymbol{\alpha}_i = \mathbf{0}, \\
&\underset{K\times K}{E\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j'} = \begin{cases} \Delta & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \\
&E\boldsymbol{x}_{it}\boldsymbol{\alpha}_j' = \mathbf{0}, \quad E\boldsymbol{\alpha}_i\boldsymbol{u}_j' = \mathbf{0}, \\
&E\boldsymbol{u}_i\boldsymbol{u}_j' = \begin{cases} \sigma_i^2 I_T & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases}
\end{aligned}
\tag{13.2.8}
$$

Stacking all NT observations, we have

$$
\boldsymbol{y} = X\bar{\boldsymbol{\beta}} + \tilde{X}\boldsymbol{\alpha} + \boldsymbol{u}, \tag{13.2.9}
$$

where

$$
\underset{NT\times 1}{\boldsymbol{y}} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')',
$$

$$
\underset{NT\times K}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \quad \underset{NT\times NK}{\tilde{X}} = \begin{bmatrix} X_1 & & & \mathbf{0} \\ & X_2 & & \\ & & \ddots & \\ \mathbf{0} & & & X_N \end{bmatrix} = \operatorname{diag}\,(X_1, \ldots, X_N),
$$

$\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_N')'$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_N')'$. The covariance matrix for the composite disturbance term $\tilde{X}\boldsymbol{\alpha} + \boldsymbol{u}$ is block-diagonal, with the ith diagonal block given by

$$
\Phi_i = X_i \Delta X_i' + \sigma_i^2 I_T. \tag{13.2.10}
$$

13.2.2.2 Estimation

Under Swamy's (1970) assumption, the simple regression of $\boldsymbol{y}$ on X will yield an unbiased and consistent estimator of $\bar{\boldsymbol{\beta}}$ if $(1/NT)\ X'X$ converges to a nonsingular constant matrix. But the estimator is inefficient, and the usual least squares formula for computing the variance–covariance matrix of the estimator is incorrect, often leading to misleading statistical inferences. Moreover, when the pattern of parameter variation is of interest in its own right, an estimator ignoring parameter variation is incapable of shedding light on this aspect of the economic process.

[3] When the distribution of random coefficients conditional on the explanatory variables is unrestricted and the errors of the equation have arbitrary dependence at all lags, the conditional distribution of the random coefficients may not be identifiable (Arellano and Bonhomme 2012; Graham and Powell 2012). However, Arellano and Bonhomme (2012) show that if the errors of the equation have limited time dependence, it remains identifiable.

[4] See Chamberlain (1992) for an extension of the Mundlak–Chamberlain approach of conditioning the individual effects on the conditioning variables to models with individual-specific slopes that may be correlated with conditioning variables. An instrumental variable estimator is proposed within a finite dimensional, method of moments framework.

The best linear unbiased estimator of $\bar{\boldsymbol{\beta}}$ for (13.2.9) is the GLS estimator,[5]

$$\begin{aligned}\hat{\bar{\boldsymbol{\beta}}}_{GLS} &= \left(\sum_{i=1}^{N} X_i'\Phi_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'\Phi_i^{-1}\boldsymbol{y}_i\right)\\ &= \sum_{i=1}^{N} W_i\hat{\boldsymbol{\beta}}_i,\end{aligned} \tag{13.2.11}$$

where

$$W_i = \left\{\sum_{i=1}^{N}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}\right\}^{-1}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1},$$

and

$$\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\boldsymbol{y}_i.$$

The last expression of (13.2.11) shows that the GLS estimator is a matrix-weighted average of the least squares estimator for each cross-sectional unit, with the weights inversely proportional to their covariance matrices. It also shows that the GLS estimator requires only a matrix inversion of order K, and so it is not much more complicated to compute than the simple least squares estimator.

The covariance matrix for the GLS estimator is

$$\begin{aligned}\text{Var}\,(\hat{\bar{\boldsymbol{\beta}}}_{GLS}) &= \left(\sum_{i=1}^{N} X_i'\Phi_i^{-1}X_i\right)^{-1}\\ &= \left\{\sum_{i=1}^{N}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}\right\}^{-1}.\end{aligned} \tag{13.2.12}$$

[5] Repeatedly using the formula that $(A + BDB')^{-1} = A^{-1} - A^{-1}B(B'A^{-1}B + D^{-1})^{-1}B'A^{-1}$ (Rao 1973, Chapter 1), we have

$$\begin{aligned}X_i'\Phi_i^{-1}X_i &= X_i'[\sigma_i^2 I + X_i\Delta X_i']^{-1}X_i\\ &= X_i'\left\{\frac{1}{\sigma_i^2}I_T - \frac{1}{\sigma_i^2}X_i[X_i'X_i + \sigma_i^2\Delta^{-1}]^{-1}X_i'\right\}X_i\\ &= \frac{1}{\sigma_i^2}\Bigg[X_i'X_i - X_i'X_i\Bigg\{(X_i'X_i)^{-1}\\ &\quad - (X_i'X_i)^{-1}\left[(X_i'X_i)^{-1} + \frac{1}{\sigma_i^2}\Delta\right]^{-1}(X_i'X_i)^{-1}\Bigg\}X_i'X_i\Bigg]\\ &= [\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}.\end{aligned}$$

Swamy (1970) proposed using the least squares estimators $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\boldsymbol{y}_i$ and their residuals $\hat{\boldsymbol{u}}_i = \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i$ to obtain unbiased estimators of σ_i^2 and Δ,[6]

$$\begin{aligned}\hat{\sigma}_i^2 &= \frac{\hat{\boldsymbol{u}}_i'\hat{\boldsymbol{u}}_i}{T-K}\\ &= \frac{1}{T-K}\boldsymbol{y}_i'[I - X_i(X_i'X_i)^{-1}X_i']\boldsymbol{y}_i,\end{aligned} \tag{13.2.13}$$

$$\begin{aligned}\hat{\Delta} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right)\\ \left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right)' - \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2(X_i'X_i)^{-1}.\end{aligned} \tag{13.2.14}$$

Again, just as in the error-component model, the estimator (13.2.14) is not necessarily nonnegative definite. In this situation, Swamy (also see Judge et al. 1980) has suggested replacing (13.2.14) by

$$\hat{\Delta} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right)\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right)'. \tag{13.2.15}$$

This estimator, although not unbiased, is nonnegative definite and is consistent when both N and T tend to infinity. Alternatively, we can use the Bayes mode estimator suggested by Lindley and Smith (1972) and Smith (1973),

$$\Delta^* = \{R + (N-1)\hat{\Delta}\}/(N + \rho - K - 2), \tag{13.2.16}$$

where R and ρ are prior parameters, assuming that Δ^{-1} has a Wishart distribution with ρ degrees of freedom and matrix R. For instance, we may let $R = \hat{\Delta}$ and $\rho = 2$ as in Hsiao, Pesaran, and Tahmiscioglu (1999).

Swamy (1970) proved that substituting $\hat{\sigma}_i^2$ and $\hat{\Delta}$ for σ_i^2 and Δ in (13.2.11) yields an asymptotically normal and efficient estimator of $\bar{\boldsymbol{\beta}}$. The speed of convergence of the GLS estimator is $N^{1/2}$. This can be seen by noting that the inverse of the covariance matrix for the GLS estimator equation (13.2.12) is[7]

$$\begin{aligned}\text{Var}\,(\hat{\bar{\boldsymbol{\beta}}}_{GLS})^{-1} &= N\Delta^{-1} - \Delta^{-1}\left[\sum_{i=1}^{N}\left(\Delta^{-1} + \frac{1}{\sigma_i^2}X_i'X_i\right)^{-1}\right]\Delta^{-1}\\ &= O(N) - O(N/T).\end{aligned} \tag{13.2.17}$$

Swamy (1970) used the model (13.2.8) and (13.2.9) to reestimate the Grunfeld investment function with the annual data of 11 U.S. corporations. His GLS estimates of the common-mean coefficients of the firms' beginning-of-year value of outstanding shares and capital stock are 0.0843 and 0.1961, with asymptotic standard errors 0.014 and 0.0412, respectively. The estimated dispersion measure of these coefficients is

[6] Equation (13.2.11) follows from the relation that $\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i + (X_i'X_i)^{-1}X'\boldsymbol{u}_i$ and $E(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})' = \Delta + \sigma_i^2(X_i'X_i)^{-1}$.

[7] We use the notation $O(N)$ to denote that the sequence $N^{-1}a_N$ is bounded (Theil 1971, p. 358).

$$\hat{\Delta} = \begin{bmatrix} 0.0011 & -0.0002 \\ & \\ & 0.0187 \end{bmatrix}. \tag{13.2.18}$$

Zellner (1966) has shown that when each $\boldsymbol{\beta}_i$ can be viewed as a random variable with a constant mean, and $\boldsymbol{\beta}_i$ and $\boldsymbol{x}_i$ are uncorrelated, thereby satisfying Swamy's (1970) assumption, the model will not possess an aggregation bias. In this sense, Swamy's estimate can also be interpreted as an average relationship, indicating that in general the value of a firm's outstanding shares is an important variable explaining the investment.

13.2.2.3 Predicting Individual Coefficients

Sometimes one may wish to predict the individual component $\boldsymbol{\beta}_i$, because it provides information on the behavior of each individual and also because it provides a basis for predicting future values of the dependent variable for a given individual. Swamy (1970, 1971) has shown that the best linear unbiased predictor, conditional on given $\boldsymbol{\beta}_i$, is the least squares estimator $\hat{\boldsymbol{\beta}}_i$. However, if the sampling properties of the class of predictors are considered in terms of repeated sampling over both time and individuals, Lee and Griffiths (1979) (also see Lindley and Smith 1972 or Section 13.4) have suggested predicting $\boldsymbol{\beta}_i$ by

$$\hat{\boldsymbol{\beta}}_i^* = \hat{\hat{\boldsymbol{\beta}}}_{GLS} + \Delta X_i'(X_i \Delta X_i' + \sigma_i^2 I_T)^{-1}(\boldsymbol{y}_i - X_i \hat{\hat{\boldsymbol{\beta}}}_{GLS}). \tag{13.2.19}$$

This predictor is the best linear unbiased estimator in the sense that $E(\hat{\boldsymbol{\beta}}_i^* - \boldsymbol{\beta}_i) = \mathbf{0}$, where the expectation is an unconditional one.

13.2.2.4 Testing for Coefficient Variation

An important question in empirical investigation is whether or not the regression coefficients are indeed varying across cross-sectional units. Because the effect of introducing random-coefficients variation is to give the dependent variable a different variance at each observation, models with this feature can be transformed into a particular heteroscedastic formulation, and likelihood-ratio tests can be used to detect departure from the constant-parameter assumption. However, computation of the likelihood-ratio test statistic can be complicated. To avoid the iterative calculations necessary to obtain maximum likelihood estimates of the parameters in the full model, Breusch and Pagan (1979) proposed a Lagrangian multiplier test for heteroscedasticity. Their test has the same asymptotic properties as the likelihood-ratio test in standard situations, but it is computationally much simpler. It can be computed simply by repeatedly applying least square regressions.

Dividing the individual-mean-over-time equation by σ_i, we have

$$\frac{1}{\sigma_i}\bar{y}_i = \frac{1}{\sigma_i}\bar{\boldsymbol{x}}_i'\bar{\boldsymbol{\beta}} + \omega_i, \quad i = 1, \ldots, N, \tag{13.2.20}$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \bar{\boldsymbol{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{it}, \bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}, \omega_i = \frac{1}{\sigma_i}\bar{\boldsymbol{x}}_i'\boldsymbol{\alpha}_i + \frac{1}{\sigma_i}\bar{u}_i.$$

When the assumption (13.2.8) holds, model (13.2.20) is a model with heteroscedastic variances, Var $(\omega_i) = (1/T) + (1/\sigma_i^2)\bar{\boldsymbol{x}}_i'\Delta\bar{\boldsymbol{x}}_i, i = 1, \ldots, N$. Under the null hypothesis that $\Delta = \mathbf{0}$, (13.2.20) has homoscedastic variances, Var $(\omega_i) = 1/T, i = 1, \ldots, N$. Thus, we can generalize the Breusch and Pagan (1979) test of heteroscedasticity to test for random-coefficients variation here.

Following the procedures of Rao (1973, p. 418–419) it can be shown that the transformed Lagrangian-multiplier statistic[8] for testing the null hypothesis leads to computing one-half the predicted sum of squares in a regression of

$$(T\omega_i^2 - 1) = \frac{1}{\sigma_i^2}\left[\sum_{k=1}^{K}\sum_{k'=1}^{K} \bar{x}_{ki}\bar{x}_{k'i}\sigma^2_{\alpha_{kk'}}\right] + \epsilon_i, \quad i = 1, \ldots, N, \tag{13.2.21}$$

where $\sigma^2_{\alpha_{kk'}} = E(\alpha_{ki}\alpha_{k'i})$.[9] Because ω_i and σ_i^2 usually are unknown, we can substitute them by their estimated values $\hat{\omega}_i$ and $\hat{\sigma}_i^2$, where $\hat{\omega}_i$ is the least squares residual of (13.2.20) and $\hat{\sigma}_i^2$ is given by (13.2.13). When both N and T tend to infinity, the transformed Lagrangian-multiplier statistic has the same limiting distribution as chi-square with $[K(K+1)]/2$ degrees of freedom under the null hypothesis of $\Delta = \mathbf{0}$.

The Breusch and Pagan (1979) Lagrangian multiplier test can be put into the White (1980) information matrix test framework. Chesher (1984) has shown that the many variants of varying parameters of the same general type of model under consideration can be tested using the statistic

$$\begin{aligned} D_N(\hat{\boldsymbol{\theta}}_N) = {} & \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \frac{\partial^2 \log f(y_{it} \mid \boldsymbol{x}_{it}, \hat{\boldsymbol{\theta}}_N)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \\ & + \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T} \frac{\partial \log f(y_{it} \mid \boldsymbol{x}_{it}, \hat{\boldsymbol{\theta}}_N)}{\partial\boldsymbol{\theta}}\right]\left[\sum_{t=1}^{T} \frac{\partial \log f(y_{it} \mid \boldsymbol{x}_{it}; \hat{\boldsymbol{\theta}}_N)}{\partial\boldsymbol{\theta}'}\right], \end{aligned} \tag{13.2.22}$$

where $f(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{\theta})$ denotes the conditional density of y_{it} given $\boldsymbol{x}_{it}$ and $\boldsymbol{\theta}$ under the null of no parameter variation, and $\hat{\boldsymbol{\theta}}_N$ denotes the maximum likelihood estimator of $\boldsymbol{\theta}$. Under the null, $E\left(\frac{\partial^2 \log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = -E\left(\frac{\partial \log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \cdot \frac{\partial \log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)$. Therefore, elements of $\sqrt{N}D_N(\hat{\boldsymbol{\theta}}_N)$ are asymptotically jointly normal with mean zero and covariance matrix given by White (1980) and simplified by Chesher (1983) and Lancaster (1984).

Alternatively, because for a given i, $\boldsymbol{\alpha}_i$ is fixed, we can test for random variation indirectly by testing whether or not the fixed-coefficient vectors $\boldsymbol{\beta}_i$ are all equal. That is, we form the null hypothesis:

$$H_0\colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N = \bar{\boldsymbol{\beta}}.$$

[8] We call this a transformed Lagrangian multiplier test because it is derived by maximizing the log-likelihood function of $\bar{y}_i/\sigma_i$ rather than that of y_{it}/σ_{it}.

[9] Let

$$(T\hat{\omega}_i^2 - 1) = \frac{1}{\sigma_i^2}\left[\sum_{k=1}^{K}\sum_{k'=1}^{K} \bar{x}_{ki}\bar{x}_{k'i}\hat{\sigma}^2_{\alpha kk'}\right]$$

be the least squares predicted value of $(T\hat{\omega}_i^2 - 1)$; then the predicted sum of squares is

$$\sum_{i=1}^{N}(T\hat{\omega}_i^2 - 1)^2.$$

If different cross-sectional units have the same variance, $\sigma_i^2 = \sigma^2, i = 1, \ldots, N$, the conventional analysis-of-covariance test for homogeneity discussed in Section 2.7 (F_3) can be applied. If σ_i^2 are assumed different, as postulated by Swamy (1970, 1971), we can apply the modified test statistic

$$F_3^* = \sum_{i=1}^{N} \frac{(\hat{\boldsymbol{\beta}}_i - \hat{\hat{\boldsymbol{\beta}}}^*)' X_i' X_i (\hat{\boldsymbol{\beta}}_i - \hat{\hat{\boldsymbol{\beta}}}^*)}{\hat{\sigma}_i^2}, \tag{13.2.23}$$

where

$$\hat{\hat{\boldsymbol{\beta}}}^* = \left[\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} X_i' X_i\right]^{-1} \left[\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} X_i' \boldsymbol{y}_i\right].$$

Under H_0, (13.2.23) is asymptotically chi-square distributed, with $K(N-1)$ degrees of freedom, as T tends to infinity and N is fixed.

Similarly, one can test for slope homogeneity conditional on individual-specific effects. Let $X_i = (\boldsymbol{e}_T, \tilde{X}_i)$ and $\boldsymbol{\beta}_i' = (\beta_{1i}, \boldsymbol{\beta}_{2i}')$, where $\tilde{X}_i = (\boldsymbol{x}_{2,it})$ denotes the $T \times (K-1)$ time-varying exogenous variables, $\tilde{\boldsymbol{x}}_{2,it}$, and $\boldsymbol{\beta}_{2i}$ denotes the $(K-1) \times 1$ coefficients of $\boldsymbol{x}_{2,it}$. Then

$$\tilde{F}_3^* = \sum_{i=1}^{N} \left(\hat{\boldsymbol{\beta}}_{2i} - \hat{\hat{\boldsymbol{\beta}}}_2^*\right)' \left[\frac{1}{\hat{\sigma}_i^2} \tilde{X}_i' Q \tilde{X}_i\right] (\hat{\boldsymbol{\beta}}_{2i} - \hat{\hat{\boldsymbol{\beta}}}_2^*), \tag{13.2.24}$$

where $Q = I_T - \frac{1}{T}\boldsymbol{e}\boldsymbol{e}'$,

$$\hat{\boldsymbol{\beta}}_{2i} = (\tilde{X}_i' Q \tilde{X}_i)^{-1} (\tilde{X}_i' Q \boldsymbol{y}_i), \tag{13.2.25}$$

$$\hat{\hat{\boldsymbol{\beta}}}_2^* = \left(\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} \tilde{X}_i' Q \tilde{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} \tilde{X}_i' Q \boldsymbol{y}_i\right) \tag{13.2.26}$$

and

$$\hat{\sigma}_i^2 = \frac{1}{T-K} (\boldsymbol{y}_i - \tilde{X}_i' \hat{\boldsymbol{\beta}}_{2i})' Q (\boldsymbol{y}_i - \tilde{X}_i' \hat{\boldsymbol{\beta}}_{2i}). \tag{13.2.27}$$

The statistic $\tilde{F}_3^*$ is asymptotically chi-square distributed with $(K-1)(N-1)$ degress of freedom when N is fixed and $T \to \infty$. When both N and T go to infinity, Pesaran and Yamagata (2008) show that $\frac{1}{\sqrt{N}} F_3^*$ or $\frac{1}{\sqrt{N}} \tilde{F}_3^*$ is asymptotically normally distributed with mean 0 and variance 1 provided $\frac{\sqrt{N}}{T} \to 0$ as $N \to \infty$. Furthermore, they show that if $\hat{\sigma}_i^2$ (Equation 13.2.27) is replaced by the estimator

$$\tilde{\sigma}_i^2 = \frac{1}{T-1} (\boldsymbol{y}_i - \tilde{X}_i \hat{\boldsymbol{\beta}}_{2i})' Q (\boldsymbol{y}_i - \tilde{X}_i \hat{\boldsymbol{\beta}}_{2i}), \tag{13.2.28}$$

$\frac{1}{\sqrt{N}} F_3^*$ or $\frac{1}{\sqrt{N}} \tilde{F}_3^*$ possesses better finite sample properties than using $\hat{\sigma}_i^2$ in (13.2.23) or (13.2.24).

13.2.2.5 Fixed or Random Coefficients

The question of whether $\boldsymbol{\beta}_i$ should be assumed fixed and different or random and different depends on whether $\boldsymbol{\beta}_i$ can be viewed as from heterogeneous population or random draws from a common population, or whether we are making inferences conditional on the individual characteristics or making unconditional inferences on the population characteristics. If $\boldsymbol{\beta}_i$ are heterogeneous or we are making inferences conditional on the individual characteristics, the fixed-coefficients model should be used. If $\boldsymbol{\beta}_i$ can be viewed as random draws from a common population and inference is on the population characteristics, the random-coefficients model should be used. However, extending his work on the variable-intercept model, Mundlak (1978b) has raised the issue of whether or not the variable coefficients are correlated with the explanatory variables. If they are, the assumptions of the Swamy random-coefficients model are unreasonable, and the GLS estimator of the mean coefficient vector will be biased. To correct this bias, Mundlak (1978b) suggested that the inferences of $f(\boldsymbol{y}_i \mid X_i, \boldsymbol{\beta})$ be viewed as $\int f(\boldsymbol{y}_i \mid X_i, \bar{\boldsymbol{\beta}}, \boldsymbol{\alpha}_i) f(\boldsymbol{\alpha}_i \mid X_i)\, d\boldsymbol{\alpha}_i$, where $f(\boldsymbol{y}_i \mid X_i, \bar{\boldsymbol{\beta}}, \boldsymbol{\alpha}_i)$ denotes the conditional density of $\boldsymbol{y}_i$ given $X_i, \bar{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha}_i$, and $f(\boldsymbol{\alpha}_i \mid X_i)$ denotes the conditional density of $\boldsymbol{\alpha}_i$ given X_i, which provides auxiliary equations for the coefficient vector $\boldsymbol{\alpha}_i$ as a function of the ith individual's observed explanatory variables. Because this framework can be viewed as a special case of a random-coefficients model with the coefficients being functions of other explanatory variables, we shall maintain the assumption that the random coefficients are not correlated with the explanatory variables, and we shall discuss estimation of the random coefficients that are functions of other explanatory variables in Section 13.8.

13.2.2.6 An Example

To illustrate the specific issues involved in estimating a behavioral equation using temporal cross-sectional observations when the data do not support the hypothesis that the coefficients are the same for all cross-sectional units, we report a study conducted by Barth, Kraft, and Kraft (1979). They used quarterly observations on output prices, wages, materials prices, inventories, and sales for 17 manufacturing industries for the period 1959 (I) to 1971 (II) to estimate a price equation for the U.S. manufacturing sector. Assuming heteroscedastic disturbance, but common intercept and slope coefficients across industries, and using the two-step Aitken estimator, Barth et al. (1979) obtained

$$\begin{array}{cccccccc} \hat{y} = & 0.0005 & + & 0.2853x_2 & + & 0.0068x_3 & + & 0.0024x_4, \\ & (0.0003) & & (0.0304) & & (0.005) & & (0.0017) \end{array} \tag{13.2.29}$$

where y is the quarterly change in output price, x_2 is labor costs, x_3 is materials input prices, and x_4 is a proxy variable for demand, constructed from the ratio of finished inventory to sales. The standard errors of the estimates are in parentheses.

The findings of (13.2.29) are somewhat unsettling. The contribution of materials input costs is extremely small, less than 1%. Furthermore, the proxy variable has the wrong sign. As the inventory-to-sales ratio increases, one would expect the resulting inventory buildup to exert a downward pressure on prices.

There are many reasons that (13.2.29) can go wrong. For instance, pricing behavior across industries is likely to vary, because input combinations are different, labor markets are not homogeneous, and demand may be more elastic or inelastic in one industry than another. In fact, a modified one-way analysis-of-covariance test for the common intercept and slope coefficients,

$$H_0\colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N, \quad N = 17,$$

using the statistic (13.2.23), has a value of 449.28. That is well above the chi-square-critical value of 92.841 for the 1% significance level with 64 $((N-1)K)$ degrees of freedom.

The rejection of the hypothesis of homogeneous price behavior across industries suggests a need to modify the model to allow for heterogeneous behavior across industries. However, previous studies have found that output prices are affected mainly by unit labor and materials input costs, and secondly, if at all, by demand factors. Thus, to account for heterogeneous behavior, one can assume that the relationships among variables are proper, but that the coefficients are different across industries. But if these coefficients are treated as fixed and different, this will imply a complicated aggregation problem for the price behavior of the U.S. manufacturing sector (e.g., Theil 1954). On the other hand, if the coefficients are treated as random, with common means, there is no aggregation bias (Zellner 1966). The random-coefficients formulation will provide a microeconomic foundation to aggregation, as well as permit the aggregate-price equation to capture more fully the disaggregated industry behavior. Therefore, Barth et al. (1979) used the Swamy random-coefficients formulation, (13.2.8) and (13.2.9), to reestimate the price equation. Their new estimates, with standard errors in parentheses, are

$$\begin{array}{lccccccc} \hat{y} = & -0.0006 & + & 0.3093x_2 & + & 0.2687x_3 & - & 0.0082x_4. \\ & (0.0005) & & (0.0432) & & (0.0577) & & (0.0101) \end{array} \tag{13.2.30}$$

The estimated dispersion of these coefficients is

$$\hat{\Delta} = \begin{array}{c} \begin{array}{cccc} \quad\beta_1 & \quad\beta_2 & \quad\beta_3 & \quad\beta_4 \end{array} \\ \begin{bmatrix} 0.0000 & -0.0002 & 0.0000 & -0.0001 \\ & 0.0020 & 0.0003 & 0.0081 \\ & & 0.0320 & 0.0030 \\ & & & 0.0014 \end{bmatrix} \end{array}. \tag{13.2.31}$$

The results of the Swamy random-coefficients formulation appear more plausible than the previous aggregate price specification (Equation 13.2.29), which ignores variation across industries from several points of view: (i) both labor costs and materials costs are now dominant in determining output prices; (ii) the proxy variable for demand has the correct sign, although it plays only a small and insignificant role in the determination of manufacturing prices; and (iii) productivity, as captured in the intercept term, appears to be increasing.

This example suggests that one must be careful about drawing conclusions on the basis of aggregate data or pooled estimates that do not allow for individual heterogeneity. Such estimates can be misleading in terms of both the size of coefficients and the significance of variables.

13.3 COEFFICIENTS THAT VARY OVER TIME AND CROSS-SECTIONAL UNITS

13.3.1 The Model

Just as in the variable-intercept models, it is possible to assume that the coefficient of the explanatory variable has a component specific to an individual unit and a component specific to a given time period such that

$$y_{it} = \sum_{k=1}^{K} (\bar{\beta}_k + \alpha_{ki} + \lambda_{kt}) x_{kit} + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{13.3.1}$$

Stacking all NT observations, we can rewrite (13.3.1) as

$$\boldsymbol{y} = X\bar{\boldsymbol{\beta}} + \tilde{X}\boldsymbol{\alpha} + \underline{\boldsymbol{X}}\boldsymbol{\lambda} + \boldsymbol{u}, \tag{13.3.2}$$

where $\boldsymbol{y}, X, \tilde{X}, \boldsymbol{u}$, and $\boldsymbol{\alpha}$ are defined in Section 13.2,

$$\underset{NT\times TK}{\underline{\boldsymbol{X}}} = \begin{bmatrix} \underline{\boldsymbol{X}}_1 \\ \underline{\boldsymbol{X}}_2 \\ \vdots \\ \underline{\boldsymbol{X}}_N \end{bmatrix}, \quad \underset{T\times TK}{\underline{\boldsymbol{X}}_i} = \begin{bmatrix} \boldsymbol{x}_{i1}' & & & \boldsymbol{0}' \\ & \boldsymbol{x}_{i2}' & & \\ & & \ddots & \\ \boldsymbol{0} & & & \boldsymbol{x}_{iT}' \end{bmatrix},$$

and

$$\underset{KT\times 1}{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_T')', \ \underset{K\times 1}{\boldsymbol{\lambda}_t} = (\lambda_{1t}, \ldots, \lambda_{Kt})'.$$

We can also rewrite (13.3.2) as

$$\boldsymbol{y} = X\bar{\boldsymbol{\beta}} + U_1\boldsymbol{\alpha}_1 + U_2\boldsymbol{\alpha}_2 + \cdots + U_K\boldsymbol{\alpha}_K + U_{K+1}\boldsymbol{\lambda}_1 + \cdots + U_{2K}\boldsymbol{\lambda}_K + U_{2K+1}\boldsymbol{u}, \tag{13.3.3}$$

where

$$\underset{NT\times N}{U_k} = \begin{bmatrix} x_{k11} & & & \\ \vdots & & & \boldsymbol{0} \\ x_{k1T} & & & \\ & x_{k21} & & \\ & \vdots & & \\ & x_{k2T} & & \\ & & \ddots & \\ & & & x_{kN1} \\ \boldsymbol{0} & & & \vdots \\ & & & x_{kNT} \end{bmatrix}, \quad k = 1, \ldots, K,$$

$$
\underset{NT\times T}{U_{K+k}} = \begin{bmatrix} x_{k11} & & & \mathbf{0} \\ & x_{k12} & & \\ & & \ddots & \\ \mathbf{0} & & & x_{k1T} \\ x_{k21} & & & \mathbf{0} \\ & x_{k22} & & \\ & & \ddots & \\ \mathbf{0} & & & x_{k2T} \\ x_{kN1} & & & \mathbf{0} \\ & \ddots & & \\ \mathbf{0} & & & x_{kNT} \end{bmatrix}, \quad k = 1, \ldots, K, \tag{13.3.4}
$$

$$
U_{2K+1} = I_{NT},
$$

$$
\underset{N\times 1}{\boldsymbol{\alpha}_k} = (\alpha_{k1}, \ldots, \alpha_{kN})', \quad \underset{T\times 1}{\boldsymbol{\lambda}_k} = (\lambda_{k1}, \ldots, \lambda_{kT})'.
$$

When $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ as well as $\bar{\boldsymbol{\beta}}$ are considered fixed, it is a fixed-effects model; when $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ are considered random, with $\bar{\boldsymbol{\beta}}$ fixed, Equation (13.3.3) corresponds to the mixed analysis-of-variance model (Hartley and Rao 1967). Thus, model (13.3.1) and its special case, model (13.2.1), fall within the general analysis-of-variance framework.

13.3.2 Fixed-Coefficient Model

When $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ are treated as fixed, as mentioned earlier, (13.3.1) can be viewed as a fixed-effects analysis-of-variance model. However, the matrix of explanatory variables is $NT \times (T+N+1)K$; but its rank is only $(T+N-1)K$; so we must impose $2K$ independent linear restrictions on the coefficients $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ for estimation of $\bar{\boldsymbol{\beta}}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\lambda}$. A natural way of imposing the constraints in this case is to let[10]

$$
\sum_{i=1}^{N} \alpha_{ik} = 0, \tag{13.3.5}
$$

[10] We did not impose similar restrictions in Section 13.2.1 because we did not separate $\bar{\boldsymbol{\beta}}$ from $\boldsymbol{\alpha}_i$.

and

$$\sum_{t=1}^{T} \lambda_{kt} = 0, \quad k = 1, \ldots, K. \tag{13.3.6}$$

Then the best linear unbiased estimators (BLUEs) of $\boldsymbol{\beta}, \boldsymbol{\alpha}$, and $\boldsymbol{\lambda}$ are the solutions of

$$\min\ (\boldsymbol{y} - X\bar{\boldsymbol{\beta}} - \tilde{X}\boldsymbol{\alpha} - \underline{X}\boldsymbol{\lambda})'(\boldsymbol{y} - X\bar{\boldsymbol{\beta}} - \tilde{X}\boldsymbol{\alpha} - \underline{X}\boldsymbol{\lambda}) \tag{13.3.7}$$

subject to (13.3.5) and (13.3.6) when u_{it} are i.i.d. over i and t.

13.3.3 Random-Coefficients Model

When $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_t$ are treated as random, Hsiao (1974a, 1975) assumes that

$$\begin{aligned}
\underset{K\times K}{E\boldsymbol{\alpha}_i\boldsymbol{\alpha}'_j} &= \begin{cases} \Delta & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \\
\underset{K\times K}{E\boldsymbol{\lambda}_t\boldsymbol{\lambda}'_s} &= \begin{cases} \Lambda & \text{if } t = s, \\ \mathbf{0} & \text{if } t \neq s, \end{cases}
\end{aligned} \tag{13.3.8}$$

$$E\boldsymbol{\alpha}_i\boldsymbol{\lambda}'_t = \mathbf{0}, \quad E\boldsymbol{\alpha}_i\boldsymbol{x}'_{it} = \mathbf{0}, \quad E\boldsymbol{\lambda}_t\boldsymbol{x}'_{it} = \mathbf{0},$$

and

$$E\boldsymbol{u}_i\boldsymbol{u}'_j = \begin{cases} \sigma_u^2 I_T & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases}$$

Then the composite error term,

$$\boldsymbol{v} = \tilde{X}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\lambda} + \boldsymbol{u}, \tag{13.3.9}$$

has a variance–covariance matrix of

$$\begin{aligned}
\Omega = E\boldsymbol{v}\boldsymbol{v}' &= \begin{bmatrix} X_1\Delta X'_1 & & & \mathbf{0} \\ & X_2\Delta X'_2 & & \\ & & \ddots & \\ \mathbf{0} & & & X_N\Delta X'_N \end{bmatrix} \\
&+ \begin{bmatrix} D(X_1\Lambda X'_1) & D(X_1\Lambda X'_2) & \ldots & D(X_1\Lambda X'_N) \\ D(X_2\Lambda X'_1) & D(X_2\Lambda X'_2) & & \\ & & \ddots & \\ D(X_N\Lambda X'_1) & & & D(X_N\Lambda X'_N) \end{bmatrix} + \sigma_u^2 I_{NT},
\end{aligned} \tag{13.3.10}$$

where

$$\underset{T\times T}{D(X_i\Lambda X'_j)} = \begin{bmatrix} \boldsymbol{x}'_{i1}\Lambda\boldsymbol{x}_{j1} & & & \mathbf{0} \\ & \boldsymbol{x}'_{i2}\Lambda\boldsymbol{x}_{j2} & & \\ & & \ddots & \\ \mathbf{0} & & & \boldsymbol{x}'_{iT}\Lambda\boldsymbol{x}_{jT} \end{bmatrix}.$$

We can estimate $\bar{\boldsymbol{\beta}}$ by the least squares method, but as discussed in Section 13.2.2.2, it is not efficient. Moreover, the conventional formula for the covariance matrix of the least squares esimator is misleading. If Ω is known, the BLUE of $\bar{\boldsymbol{\beta}}$ is the GLS estimator,

$$\hat{\bar{\boldsymbol{\beta}}}_{GLS} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}\boldsymbol{y}). \tag{13.3.11}$$

The variance–covariance matrix of the GLS estimator is

$$\mathrm{Var}\,(\hat{\bar{\boldsymbol{\beta}}}_{GLS}) = (X'\Omega^{-1}X)^{-1}. \tag{13.3.12}$$

Without knowledge of Ω, we can estimate $\bar{\boldsymbol{\beta}}$ and Ω simultaneously by the maximum likelihood method. However, because of the computational difficulty, a natural alternative is to first estimate Ω, then substitute the estimated Ω in (13.3.11).

When Δ and Λ are diagonal, it is easy to see from (13.3.3) that Ω is a linear combination of known matrices with unknown weights. So the problem of estimating the unknown covariance matrix is actually the problem of estimating the variance components. Statistical methods developed for estimating the variance (and covariance) components can be applied here (e.g., Anderson 1969, 1970; Rao 1970, 1972). In this section we shall describe only a method due to Hildreth and Houck (1968).[11]

Consider the time-series equation for the ith individual,

$$\boldsymbol{y}_i = X_i(\bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i) + \underline{\boldsymbol{X}}_i\boldsymbol{\lambda} + \boldsymbol{u}_i. \tag{13.3.13}$$

We can treat $\boldsymbol{\alpha}_i$ as if it is a vector of constants. Then (13.3.13) is a linear model with heteroscedastic variance. The variance of the error term $r_{it} = \sum_{k=1}^{K} \lambda_{kt}x_{kit} + u_{it}$ is

$$\theta_{it} = E[r_{it}^2] = \sum_{k=1}^{K} \sigma_{\lambda k}^2 x_{kit}^2 + \sigma_u^2. \tag{13.3.14}$$

Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT})'$; then

$$\boldsymbol{\theta}_i = \dot{X}_i\boldsymbol{\sigma}_\lambda^2, \tag{13.3.15}$$

where the first element of $\boldsymbol{x}_{it} = 1$, $\dot{X}_i$ is X_i with each of its elements squared, and $\boldsymbol{\sigma}_\lambda^2 = (\sigma_{\lambda 1}^2 + \sigma_u^2, \sigma_{\lambda 2}^2, \ldots, \sigma_{\lambda K}^2)'$.

An estimate of $\boldsymbol{r}_i$ can be obtained as the least squares residual, $\hat{\boldsymbol{r}}_i = \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i = M_i\boldsymbol{y}_i$, where $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\boldsymbol{y}_i$ and $M_i = I_T - X_i(X_i'X_i)^{-1}X_i'$. Squaring each element of $\hat{\boldsymbol{r}}_i$ and denoting it by $\dot{\hat{\boldsymbol{r}}}_i$, we have

$$E(\dot{\hat{\boldsymbol{r}}}_i) = \dot{M}_i\boldsymbol{\theta}_i = F_i\boldsymbol{\sigma}_\lambda^2, \tag{13.3.16}$$

where $\dot{M}_i$ is M_i with each of its elements squared, and $F_i = \dot{M}_i\dot{X}_i$.

Repeating the foregoing process for all i gives

$$E(\dot{\hat{\boldsymbol{r}}}) = F\boldsymbol{\sigma}_\lambda^2, \tag{13.3.17}$$

where $\dot{\hat{\boldsymbol{r}}} = (\dot{\hat{\boldsymbol{r}}}_1, \ldots, \dot{\hat{\boldsymbol{r}}}_N)'$, and $F = (F_1', \ldots, F_N')'$. Application of least squares to (13.3.17) yields a consistent estimator of $\boldsymbol{\sigma}_\lambda^2$,

$$\hat{\boldsymbol{\sigma}}_\lambda^2 = (F'F)^{-1}F'\dot{\hat{\boldsymbol{r}}}. \tag{13.3.18}$$

Similarly, we can apply the same procedure with respect to each time period to yield a consistent estimator of $\boldsymbol{\sigma}_\alpha^2 = (\sigma_{\alpha_1}^2 + \sigma_u^2, \sigma_{\alpha_2}^2, \ldots, \sigma_{\alpha_K}^2)'$. To obtain separate estimates of σ_u^2, $\sigma_{\alpha_1}^2$, and $\sigma_{\lambda 1}^2$, we note that $E(\boldsymbol{x}_{it}'\boldsymbol{\alpha}_i + u_{it})(\boldsymbol{x}_{it}'\boldsymbol{\lambda}_t + u_{it}) = \sigma_u^2$. So, letting $\hat{s}_{it}$ denote

[11] It has been shown (Hsiao 1975) that the Hildreth–Houck estimator is the minimum-norm quadratic unbiased estimator of Rao (1970).

the residual obtained by applying least squares separately to each time period, we can consistently estimate σ_u^2 by

$$\hat{\sigma}_u^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{r}_{it} \hat{s}_{it}. \tag{13.3.19}$$

Subtracting (13.3.19) from an estimated $\sigma_{\alpha_1}^2 + \sigma_u^2$ and $\sigma_{\lambda_1}^2 + \sigma_u^2$, we obtain consistent estimates of σ_α^2 and $\sigma_{\lambda_1}^2$, respectively.

Substituting consistently estimated values of $\boldsymbol{\sigma}_\alpha^2$, $\boldsymbol{\sigma}_\lambda^2$, and σ_u^2 into (13.3.11), one can show that when N and T both tend to infinity and N/T tends to a nonzero constant, the two-stage Aitken estimator is asymptotically as efficient as if one knew the true Ω. Also, Kelejian and Stephan (1983) have pointed out that contrary to the conventional regression model, the speed of convergence of $\hat{\boldsymbol{\beta}}_{GLS}$ here is not $(NT)^{1/2}$, but max $(N^{1/2}, T^{1/2})$.

If one is interested in predicting the random components associated with an individual, Lee and Griffiths (1979) have shown that the predictor

$$\hat{\boldsymbol{\alpha}} = (I_N \otimes \Delta) X' \Omega^{-1} (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{GLS}) \tag{13.3.20}$$

is the BLUE.

To test for the random variation of the coefficients, we can again apply the Breusch and Pagan (1980) Lagrangian-multiplier test for heteroscedasticity. Because for a given i, $\boldsymbol{\alpha}_i$ is fixed, the error term $\boldsymbol{x}_{it}'\boldsymbol{\lambda}_t + u_{it}$ will be homoscedastic if the coefficients are not varying over time. Therefore, under the null, one-half the explained sum of squares in a regression[12]

$$\frac{\hat{u}_{it}^2}{\hat{\sigma}_u^2} = \dot{\boldsymbol{x}}_{it}'\boldsymbol{\sigma}_\lambda^2 + \epsilon_{it}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{13.3.21}$$

is distributed asymptotically as a chi-square, with $K - 1$ degrees of freedom, where $\hat{u}_{it} = y_{it} - \hat{\boldsymbol{\beta}}_i'\boldsymbol{x}_{it}, \hat{\sigma}_u^2 = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{\boldsymbol{\beta}}_i'\boldsymbol{x}_{it})^2/NT$, and $\dot{\boldsymbol{x}}_{it}$ is $\boldsymbol{x}_{it}$ with each element squared.[13]

Similarly, we can test for random variation across cross-sectional units by regressing

$$\frac{\hat{u}_{it}^{*2}}{\hat{\sigma}_u^{*2}} = \dot{\boldsymbol{x}}_{it}'\boldsymbol{\sigma}_\alpha^2 + \epsilon_{it}^*, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{13.3.22}$$

where $\hat{u}_{it}^* = y_{it} - \hat{\boldsymbol{\beta}}_t'\boldsymbol{x}_{it}, \hat{\sigma}_u^{*2} = \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^{*2}/NT$, and $\hat{\boldsymbol{\beta}}_t$ is the least square estimate of $\boldsymbol{\beta}_t = \boldsymbol{\beta} + \boldsymbol{\lambda}_t$ across cross-sectional units for a given t. Under the null hypothesis of no random variation across cross-sectional units, one-half of the explained sum of squares of (13.3.22) is asymptotically chi-square-distributed, with $K - 1$ degrees of freedom.

We can also test the random variation indirectly by applying the classic analysis-of-covariance test. For details, see Hsiao (1974a).

Swamy and Mehta (1977) have proposed a more general type of time-varying-component model by allowing $E\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t' = \Lambda_t$ to vary over t. However, models with the

[12] Let $(y_{it} - \bar{y})$ be the deviation of the sample mean, and let $(\widehat{y_{it} - \bar{y}})$ be its least-squares prediction. Then the explained sum of squares is $\sum(\widehat{y_{it} - \bar{y}})^2$.

[13] Note here that the first term $\dot{x}_{1it} = 1$. So the null hypothesis is $(\sigma_{\lambda 2}^2, \ldots, \sigma_{\lambda K}^2) = (0, \ldots, 0)$.

coefficients varying randomly across cross-sectional units and over time have not gained much acceptance in empirical investigation. Part of the reason is because the inversion of Ω is at least of order max (NK, TK) (Hsiao 1974a). For any panel data of reasonable size, this would be a computationally demanding problem.

13.4 A MIXED FIXED- AND RANDOM-COEFFICIENTS MODEL

13.4.1 Model Formulation

Many of the previously discussed models can be put as special cases of a general mixed fixed- and random-coefficients model. For ease of exposition, we shall assume that only time-invariant cross-sectional heterogeneity exists.

Suppose that each cross-sectional unit is postulated to be different, so that

$$y_{it} = \sum_{k=1}^{K} \beta_{ki} x_{kit} + \sum_{\ell=1}^{m} \gamma_{\ell i} w_{\ell it} + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{13.4.1}$$

where $\boldsymbol{x}_{it}$ and $\boldsymbol{w}_{it}$ are each $K \times 1$ and $m \times 1$ vectors of explanatory variables that are independent of the error of the equation, u_{it}. Stacking the NT observations together, we have

$$\boldsymbol{y} = X\boldsymbol{\beta} + W\boldsymbol{\gamma} + \boldsymbol{u}, \tag{13.4.2}$$

where

$$\underset{NT \times NK}{X} = \begin{pmatrix} X_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & X_2 & \ldots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & & X_N \end{pmatrix},$$

$$\underset{NT \times Nm}{W} = \begin{pmatrix} W_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & W_2 & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & W_N \end{pmatrix},$$

$$\underset{NT \times 1}{\boldsymbol{u}} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_N')',$$

$$\underset{NK \times 1}{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N')' \text{ and } \underset{Nm \times 1}{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_N')'.$$

Equation (13.4.1), just like (13.4.2), assumes a different behavioral equation relation for each cross-sectional unit. In this situation, the only advantage of pooling is to put the model (13.4.2) in Zellner's (1962) seemingly unrelated regression framework to gain efficiency of the estimates of the individual behavioral equation.

The motivation of a mixed fixed- and random-coefficients model is that while there may be fundamental differences among cross-sectional units, conditioning on these individual specific effects, one may still be able to draw inferences on certain population characteristics through the imposition of a priori constraints on the coefficients of $\boldsymbol{x}_{it}$ and $\boldsymbol{w}_{it}$. We assume that there exist two kinds of restrictions, stochastic and fixed restrictions (e.g., Hsiao 1991a; Hsiao, Appelbe, and Dineen 1993) in the form:

A.13.4.1 The coefficients of $\boldsymbol{x}_{it}$ are assumed to be subject to stochastic restrictions of the form:

$$\boldsymbol{\beta} = A_1\bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}, \tag{13.4.3}$$

where A_1 is an $NK \times L$ matrix with known element, $\bar{\boldsymbol{\beta}}$ is an $L \times 1$ vector of constants, and $\boldsymbol{\alpha}$ is assumed to be (normally distributed) random variables with mean $\mathbf{0}$ and nonsingular constant covariance matrix C and is independent of $\boldsymbol{x}_{it}$.

A.13.4.2 The coefficients of $\boldsymbol{w}_{it}$ are assumed to be subject to

$$\boldsymbol{\gamma} = A_2\bar{\boldsymbol{\gamma}}, \tag{13.4.4}$$

where A_2 is an $Nm \times n$ matrix with known elements, and $\bar{\boldsymbol{\gamma}}$ is an $n \times 1$ vector of constants.

Since A_2 is known, we may substitute (13.4.4) into (13.4.2) and write the model as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \tilde{W}\bar{\boldsymbol{\gamma}} + \boldsymbol{u} \tag{13.4.5}$$

subject to (13.4.3), where $\tilde{W} = WA_2$.

A.13.4.2 allows for various possible fixed parameter configurations. For instance, if $\boldsymbol{\gamma}$ is different across cross-sectional units, we can let $A_2 = I_N \otimes I_m$. On the other hand, if we wish to constrain $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j$, we can let $A_2 = \boldsymbol{e}_N \otimes I_m$.

Many of the linear panel data models with unobserved individual specific but time-invariant heterogeneity can be treated as the special case of models (13.4.2)–(13.4.4). These include:

1. A common model for all cross-sectional units. If there is no interindividual difference in behavioral patterns, we may let $X = \mathbf{0}$, $A_2 = \boldsymbol{e}_N \otimes I_m$, so (13.4.2) becomes

$$y_{it} = \boldsymbol{w}_{it}'\bar{\boldsymbol{\gamma}} + u_{it}. \tag{13.4.6}$$

2. Different models for different cross-sectional units. When each individual is considered different, then $X = \mathbf{0}$, $A_2 = I_N \otimes I_m$, and (13.4.2) becomes

$$y_{it} = \boldsymbol{w}_{it}'\boldsymbol{\gamma}_i + u_{it}. \tag{13.4.7}$$

3. Variable intercept model (e.g., Kuh 1963; or Chapter 2.2). If conditional on the observed exogenous variables, the interindividual differences stay constant through time; let $X = \mathbf{0}$,

$$A_2 = (I_N \otimes \boldsymbol{i}_m \vdots \boldsymbol{e}_N \otimes I^*_{m-1}), \bar{\boldsymbol{\gamma}} = (\gamma_{11}, \ldots, \gamma_{N1}, \bar{\gamma}_2, \ldots, \bar{\gamma}_m)',$$

where we arrange $W_i = (\boldsymbol{e}_T, \boldsymbol{w}_{i2}, \ldots, \boldsymbol{w}_{im}), i = 1, \ldots, N$. $\boldsymbol{i}_m = (1, 0, \ldots, 0)'$, $\underset{m\times(m-1)}{I^*_{m-1}} = (\mathbf{0} \vdots I_{m-1})'$, then (13.4.2) becomes

$$y_{it} = \gamma_{i1} + \bar{\gamma}_2 w_{it2} + \cdots + \bar{\gamma}_m w_{itm} + u_{it}. \tag{13.4.8}$$

4. Error components model (e.g., Balestra and Nerlove 1966; Wallace and Hussain 1969; or Section 2.3). When the effects of the individual-specific, time-invariant omitted variables are treated as random variables just like the assumption on the effects of other omitted variables, we can let $X_i = \boldsymbol{e}_T, \boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_N)$, $A_1 = \boldsymbol{e}_N, C = \sigma^2_\alpha I_N$, $\bar{\beta}$ is an unknown constant, and $\boldsymbol{w}_{it}$ does not contain an intercept term. Then (13.4.2) becomes

$$y_{it} = \bar{\beta} + \bar{\boldsymbol{\gamma}}'\boldsymbol{w}_{it} + \alpha_i + u_{it}. \tag{13.4.9}$$

5. Random-Coefficients Model (Swamy 1970; or Chapter 13.2.2). Let $Z = \mathbf{0}$, $A_1 = \boldsymbol{e}_N \otimes I_K, C = I_N \otimes \Delta$, we have model (13.2.7).

13.4.2 A Bayes Solution

The formulation of (13.4.5) subject to (13.4.3) can be viewed from a Bayesian perspective as there exist informative prior on $\boldsymbol{\beta}$ (13.4.3), but not on $\bar{\boldsymbol{\gamma}}$. In the classical sampling approach, inferences are made by typically assuming that the probability law generating the observations, $\boldsymbol{y}$, $f(\boldsymbol{y}, \boldsymbol{\theta})$, is known, but not the vector of constant parameter $\boldsymbol{\theta}$. Estimators $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ of the parameters $\boldsymbol{\theta}$ are chosen as functions of $\boldsymbol{y}$ so that their sampling distributions, in repeated experiments, are in some sense concentrated as closely as possible about the true values of $\boldsymbol{\theta}$. In the Bayesian approach, a different line is taken. First, all quantities, including the parameters, are considered random variables. Second, all probability statements are conditional, so that in making a probability statement, it is necessary to refer to the conditioning event as well as the event whose probability is being discussed. Therefore, as part of the model, a prior distribution of the parameter $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, is introduced. The prior distribution is supposed to express a state of knowledge (or ignorance) about $\boldsymbol{\theta}$ before the data are obtained. Given the probability model $f(\boldsymbol{y}; \boldsymbol{\theta})$, the prior distribution, and the data $\boldsymbol{y}$, the probability distribution of $\boldsymbol{\theta}$ is revised to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, which is called the posterior distribution of $\boldsymbol{\theta}$, according to Bayes's theorem (e.g., Intriligator, Bodkin, and Hsiao 1996).[14]

$$P(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto P(\boldsymbol{\theta}) f(\boldsymbol{y} \mid \boldsymbol{\theta}), \tag{13.4.10}$$

where the sign "$\propto$" denoting "is proportional to," with the factor of proportionality being a normalizing constant.

Under the assumption that

A.13.4.3 $\boldsymbol{u} \sim N(\mathbf{0}, \Omega)$,

we may write the model (13.4.5) as

A.1 Conditional on $X, \tilde{W}, \boldsymbol{\beta}$, and $\bar{\boldsymbol{\gamma}}$

$$\boldsymbol{y} \sim N(X\boldsymbol{\beta} + \tilde{W}\bar{\boldsymbol{\gamma}}, \Omega). \tag{13.4.11}$$

A.2 The prior distribution of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\gamma}}$ are independent,

$$P(\boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}) = P(\boldsymbol{\beta}) \cdot P(\bar{\boldsymbol{\gamma}}). \tag{13.4.12}$$

A.3 $P(\boldsymbol{\beta}) \sim N(A_1\bar{\boldsymbol{\beta}}, C)$.

A.4 There is no information about $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$; therefore, $P(\bar{\boldsymbol{\beta}})$ and $P(\bar{\boldsymbol{\gamma}})$ are independent and

$$P(\bar{\boldsymbol{\beta}}) \propto \text{ constant},$$

$$P(\bar{\boldsymbol{\gamma}}) \propto \text{ constant}.$$

Conditional on Ω and C, repeatedly applying the formulas in Appendix 13, yields (Hsiao et al. 1993)

[14] According to Bayes's theorem, the probability of B given A, written as $P(B \mid A)$, equals $P(B \mid A) = \frac{P(A|B)P(B)}{P(A)}$, which is proportional to $P(A \mid B)P(B)$.

1. The posterior distribution of $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$ given $\boldsymbol{y}$ is

$$N\left(\begin{pmatrix}\bar{\boldsymbol{\beta}}^* \\ \bar{\boldsymbol{\gamma}}^*\end{pmatrix}, D_1\right), \tag{13.4.13}$$

where

$$D_1 = \left[\begin{pmatrix}A_1'X' \\ \tilde{W}'\end{pmatrix}(\Omega + XCX')^{-1}(XA_1, \tilde{W})\right]^{-1}, \tag{13.4.14}$$

and

$$\begin{pmatrix}\bar{\boldsymbol{\beta}}^* \\ \bar{\boldsymbol{\gamma}}^*\end{pmatrix} = D_1 \begin{bmatrix}A_1'X' \\ \tilde{W}'\end{bmatrix}(\Omega + XCX')^{-1}\boldsymbol{y} \tag{13.4.15}$$

2. The posterior distribution of $\boldsymbol{\beta}$ given $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{y}$ is $N(\boldsymbol{\beta}^*, D_2)$, where

$$D_2 = \{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X + C^{-1}\}^{-1}, \tag{13.4.16}$$

$$\boldsymbol{\beta}^* = D_2\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]\boldsymbol{y} + C^{-1}A_1\bar{\boldsymbol{\beta}}\}. \tag{13.4.17}$$

3. The (unconditional) posterior distribution of $\boldsymbol{\beta}$ is $N(\boldsymbol{\beta}^{**}, D_3)$, where

$$\begin{aligned} D_3 = \{&X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X + C^{-1} \\ &- C^{-1}A_1(A_1'C^{-1}A_1)^{-1}A_1'C^{-1}\}^{-1}, \end{aligned} \tag{13.4.18}$$

$$\begin{aligned} \boldsymbol{\beta}^{**} &= D_3\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]\boldsymbol{y}\} \\ &= D_2\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X\hat{\boldsymbol{\beta}} + C^{-1}A_1\bar{\boldsymbol{\beta}}^*\}, \end{aligned} \tag{13.4.19}$$

where $\hat{\boldsymbol{\beta}}$ is the GLS estimate of (13.4.5),

$$\begin{aligned} \hat{\boldsymbol{\beta}} = &\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X\}^{-1} \\ &\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]\boldsymbol{y}\}. \end{aligned} \tag{13.4.20}$$

Given a quadratic loss function of the error of the estimation, a Bayes point estimate is the posterior mean. The posterior mean of $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$ (13.4.15) is the GLS estimator of the model (13.4.5) after substituting the restriction (13.4.3),

$$\boldsymbol{y} = XA_1\bar{\boldsymbol{\beta}} + \tilde{W}\bar{\boldsymbol{\gamma}} + \boldsymbol{v}, \tag{13.4.21}$$

where $\boldsymbol{v} = X\boldsymbol{\alpha} + \boldsymbol{u}$. However, the posterior mean of $\boldsymbol{\beta}$ is not the GLS estimator of (13.4.5). It is the weighted average between the GLS estimator of $\boldsymbol{\beta}$ and the overall mean $\bar{\boldsymbol{\beta}}$ (13.4.17) or $\bar{\boldsymbol{\beta}}^*$ (13.4.19) with the weights proportional to the inverse of the precision of respective estimates. The reason is that, although both (13.4.2) and (13.4.5) allow the coefficients to be different across cross-sectional units, (13.4.3) has imposed additional prior information that $\boldsymbol{\beta}$ are randomly distributed with mean $A_1\bar{\boldsymbol{\beta}}$. For (13.4.2), the best linear predictor for an individual outcome is to substitute the best linear unbiased estimator of the individual coefficients into the individual equation. For models (13.4.5) and (13.4.3), because the expected $\boldsymbol{\beta}_i$ is the same across i and the actual difference can be attributed to a chance outcome, additional information about $\boldsymbol{\beta}_i$ may be obtained by examining the behavior of others; hence, (13.4.17) or (13.4.19).

In the special case of the error-components model (13.4.9), $X = I_N \otimes \boldsymbol{e}_T$. Under the assumption that $\boldsymbol{w}_{it}$ contains an intercept term (i.e., $\bar{\beta} = 0$) and u_{it} is i.i.d., the Bayes estimators (13.4.15) of $\bar{\boldsymbol{\gamma}}$ are simply the GLS estimator of (13.4.21), $\bar{\boldsymbol{\gamma}}^*$. The Bayes estimator of α_i is

$$\alpha_i^{**} = \left(\frac{T\sigma_\alpha^2}{T\sigma_\alpha^2 + \sigma_u^2} \right) \hat{\bar{v}}_i, \tag{13.4.22}$$

where $\hat{\bar{v}}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{v}_{it}$ and $\hat{v}_{it} = y_{it} - \bar{\boldsymbol{\gamma}}^* \boldsymbol{w}_{it}$. Substituting $\bar{\boldsymbol{\gamma}}^*$ and α_i^{**} for the unknown $\bar{\boldsymbol{\gamma}}$ and α_i in (13.4.9), Wansbeek and Kapteyn (1978) and Taub (1979) show that

$$\hat{y}_{i,T+S} = \bar{\boldsymbol{\gamma}}^{*'} \boldsymbol{w}_{i,t+s} + \alpha_i^{**} \tag{13.4.23}$$

is the best linear predictor (BLUP) for the ith individual s periods ahead.[15]

13.4.3 Random or Fixed Differences?

13.4.3.1 An Example of the Contrast Between Individual and Pooled Parameter Estimates

In a classical framework, it makes no sense to predict the independently drawn random variable $\boldsymbol{\beta}_i$ (or $\boldsymbol{\alpha}_i$). However, in panel data, we actually operate with two dimensions – a cross-sectional dimension and a time series dimension. Even though $\boldsymbol{\beta}_i$ is an independently distributed random variable across i, once a particular $\boldsymbol{\beta}_i$ is drawn, it stays constant over time. Therefore, it makes sense to predict $\boldsymbol{\beta}_i$. The classical predictor of $\boldsymbol{\beta}_i$ is the generalized least squares estimator of the model (13.4.5). The Bayes predictor (13.4.19) is the weighted average between the generalized least squares estimator of $\boldsymbol{\beta}$ for the model (13.4.5) and the overall mean $A_1\bar{\boldsymbol{\beta}}$ if $\bar{\boldsymbol{\beta}}$ is known or $A_1\bar{\boldsymbol{\beta}}^*$ if $\bar{\boldsymbol{\beta}}$ is unknown with the weights proportional to the inverse of the precisions of respective weights. The Bayes estimator of the individual coefficients, $\boldsymbol{\beta}_i$, "shrinks" the GLS estimator of $\boldsymbol{\beta}_i$ toward the grand mean $\bar{\boldsymbol{\beta}}$ or $\bar{\boldsymbol{\beta}}^*$. The reason for doing so stems from de Finetti's (1964) exchangeability assumption. When there are not enough time series observations to allow for precise estimation of individual $\boldsymbol{\beta}_i$ (namely, T is small), additional information about $\boldsymbol{\beta}_i$ may be obtained by examining the behavior of others because the expected response is assumed to be the same and the actual differences in response among individuals are the work of a chance mechanism.

Table 13.1 presents the Canadian route-specific estimates of the demand for customer-dialed long distance service over 920 miles (long-haul) based on quarterly data from 1980.I to 1989.IV (Hsiao et al. 1993). Some of the point-to-point individual route estimates (unconstrained model) of the price and income coefficients have the wrong signs, (Table 13.1, Column 2) perhaps because of multicollinearity. However, when one invokes the representative consumer argument by assuming that consumers respond in more or less the same way to price and income changes, thus assuming the coefficients of these variables across routes are considered random draws from a common population with a constant mean and variance–covariance matrix, but also allows the route-specific effects to exist by assuming that the coefficients of the intercept and seasonal dummies are fixed and different for different routes, all the estimated route specific price and income coefficients have the correct signs (Table 13.1, column 3).

[15] When u_{it} is serially correlated, see Baltagi and Li (1992). For the asymptotic mean square error when the coefficients and error-components parameters are estimated, see Baillie and Baltagi (1999).

Table 13.1. *Long-haul regression coefficients*[a]

	Price coefficient	
Route	unconstrained	Mixed coefficients
1	−0.0712(−0.15)	−0.2875(N/A)
2	0.1694(0.44)	−0.0220(N/A)
3	−1.0142(−5.22)	−0.7743(N/A)
4	−0.4874(−2.29)	−0.1686(N/A)
5	−0.3190(−2.71)	−0.2925(N/A)
6	0.0365(0.20)	−0.0568(N/A)
7	−0.3996(−3.92)	−0.3881(N/A)
8	−0.1033(−0.95)	−0.2504(N/A)
9	−0.3965(−4.22)	−0.2821(N/A)
10	−0.6187(−4.82)	−0.5934(N/A)
Average	N/A	−0.3116
Route	Income coefficient	
1	1.4301(3.07)	0.4740(N/A)
2	−0.348(−0.09)	0.2679(N/A)
3	0.3698(1.95)	0.3394(N/A)
4	0.2497(0.70)	0.3145(N/A)
5	0.5556(2.71)	0.3501(N/A)
6	0.1119(0.95)	0.1344(N/A)
7	0.9197(8.10)	0.5342(N/A)
8	0.3886(3.88)	0.5255(N/A)
9	0.6688(6.16)	0.5648(N/A)
10	0.1928(2.39)	0.2574(N/A)
Average	N/A	0.3762

[a] t-statistics in parentheses.
Source: Hsiao, Appelbe, and Dineen (1993, Table 3).

13.4.3.2 Model Selection

The above example demonstrates that the way in which individual heterogeneity is taken into account makes a difference in the accuracy of inference. The various estimation methods discussed so far presuppose that we know which coefficients should be treated as fixed (and different) and which coefficients should be treated as random. In practice, we have very little prior information on selecting the appropriate specifications. Various statistical tests have been suggested to select an appropriate formulation (e.g., Breusch and Pagan 1979; Hausman 1978; or Chapter 13.2.2.4). However, all these tests essentially exploit the implication of certain formulation in a specific framework. They are indirect in nature. The distribution of a test statistic is derived under a specific null, but the alternative is composite. The rejection of a null does not automatically imply the acceptance of a specific alternative. It would appear more appropriate to treat the fixed coefficients, random coefficients, or various forms of mixed fixed- and random-coefficients models as different models and use model selection criteria to select an appropriate specification (Hsiao and Sun 2000). For instance, a well-known model selection criterion such as Akaike's (1973) information criteria or Schwarz's (1978) Bayesian information criteria that selects the model H_j among $j = 1, \ldots, J$ different specifications if it yields the smallest value of

$$-2 \log f(\mathbf{y} \mid H_j) + 2m_j, \quad j = 1, \ldots, J, \tag{13.4.24}$$

or

$$-2 \log f(\mathbf{y} \mid H_j) + m_j \log NT, \quad j = 1, \ldots, J, \tag{13.4.25}$$

can be used, where $\log f(\boldsymbol{y} \mid H_j)$ and m_j denote the log-likelihood values of $\boldsymbol{y}$ and the number of unknown parameters of model H_j. Alternatively, Hsiao (1995) and Min and Zellner (1993) suggest selecting the model that yields the highest predictive density. In this framework, time series observations are divided into two periods, 1 to T_1, denoted by $\boldsymbol{y}^1$, and $T_1 + 1$ to T, denoted by $\boldsymbol{y}^2$. The first T_1 observations are used to obtain the probability distribution of the parameters associated with H_j, say $\boldsymbol{\theta}^j$, $P(\boldsymbol{\theta}^j \mid \boldsymbol{y}^1)$. The predictive density is then evaluated as

$$\int f(\boldsymbol{y}^2 \mid \boldsymbol{\theta}^j) p(\boldsymbol{\theta}^j \mid \boldsymbol{y}^1) d\boldsymbol{\theta}^j, \tag{13.4.26}$$

where $f(\boldsymbol{y}^2 \mid \boldsymbol{\theta}^j)$ is the density of $\boldsymbol{y}^2$ conditional on $\boldsymbol{\theta}^j$. Given the sensitivity of the Bayesian approach to the choice of prior, the advantage of using (13.4.26) is that the choice of a model does not have to depend on the prior. One can use the noninformative (or diffuse) prior to derive $P(\boldsymbol{\theta}^j \mid \boldsymbol{y}^1)$. It is also consistent with the theme that "a severe test for an economic theory, the only test and the ultimate test is its ability to predict" (Klein 1988, p. 21; also see Friedman 1953).

When $\boldsymbol{y}^2$ contains only a limited number of observations, the choice of model in terms of predictive density may become heavily sample dependent. If too many observations are put in $\boldsymbol{y}^2$, then a lot of sample information is not utilized to estimate unknown parameters. One compromise is to modify (13.4.26) by recursively updating the estimates,

$$\begin{aligned} &\int f(\boldsymbol{y}_T \mid \boldsymbol{\theta}^j, \boldsymbol{y}^{T-1}) P(\boldsymbol{\theta}^j \mid \boldsymbol{y}^{T-1}) d\boldsymbol{\theta}^j \\ &\cdot \int f(\boldsymbol{y}_{T-1} \mid \boldsymbol{\theta}^j, \boldsymbol{y}^{T-2}) P(\boldsymbol{\theta}^j \mid \boldsymbol{y}^{T-2}) d\boldsymbol{\theta}^j \\ &\cdots \int f(\boldsymbol{y}_{T_1+1} \mid \boldsymbol{\theta}^j, \boldsymbol{y}^{1}) P(\boldsymbol{\theta}^j \mid \boldsymbol{y}^{1}) d\boldsymbol{\theta}^j, \end{aligned} \tag{13.4.27}$$

where $P(\boldsymbol{\theta}^j \mid \boldsymbol{y}^T)$ denotes the posterior distribution of $\boldsymbol{\theta}$ given observations from 1 to T. While the formula may look formidable, it turns out that the Bayes updating formula is fairly straightforward to compute. For instance, consider the model (13.4.5). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \bar{\boldsymbol{\gamma}})$ and $\boldsymbol{\theta}_t$ and V_t denote the posterior mean and variance of $\boldsymbol{\theta}$ based on the first t-observations. Then

$$\boldsymbol{\theta}_t = V_{t-1}(Q_t' \Omega^{-1} \boldsymbol{y}_t + V_{t-1}^{-1} \boldsymbol{\theta}_{t-1}), \tag{13.4.28}$$

$$V_t = (Q_t' \Omega^{-1} Q_t + V_{t-1}^{-1})^{-1}, \quad t = T_1 + 1, \ldots, T, \tag{13.4.29}$$

and

$$\begin{aligned} P(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}^t) &= \int P(\boldsymbol{y}_{t+1} \mid \theta, \boldsymbol{y}^t) P(\boldsymbol{\theta} \mid \boldsymbol{y}^t) d\boldsymbol{\theta} \\ &\sim N(Q_{t+1}\boldsymbol{\theta}_t, \Omega + Q_{t+1} V_t Q_{t+1}'), \end{aligned} \tag{13.4.30}$$

where $\boldsymbol{y}_t' = (y_{1t}, y_{2t}, \ldots, y_{Nt})$, $Q_t = (\boldsymbol{x}_t', \boldsymbol{w}_t')$, $\boldsymbol{x}_t = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt})$, $\boldsymbol{w}_t = (\boldsymbol{w}_{1t}, \ldots, \boldsymbol{w}_{Nt})$, $\Omega = E\boldsymbol{u}_t\boldsymbol{u}_t'$, and $\boldsymbol{u}_t' = (u_{1t}, \ldots, u_{Nt})$ (Hsiao et al. 1993).

Hsiao and Sun (2000) conducted limited Monte Carlo studies to evaluate the performance of these model selection criteria in selecting the random, fixed, and mixed random-fixed-coefficients specification. They all appear to have a very high percentage in selecting the correct specification.

13.5 DYNAMIC RANDOM-COEFFICIENTS MODELS

For ease of exposition and without loss of the essentials, instead of considering generalizing (13.4.5) into the dynamic model, in this section we consider the generalization of the random-coefficients model (13.2.1) to the dynamic model of the form[16]

$$y_{it} = \gamma_i y_{i,t-1} + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + u_{it}, \qquad \mid \gamma_i \mid < 1, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{13.5.1}$$

where $\boldsymbol{x}_{it}$ is a $k \times 1$ vector of exogenous variables, and the error term u_{it} is assumed to be independently identically distributed (i.i.d.) over t with mean zero and variance $\sigma_{u_i}^2$ and is independent across i. The coefficients $\boldsymbol{\theta}_i = (\gamma_i, \boldsymbol{\beta}_i')'$ are assumed to be independently distributed across i with mean $\bar{\boldsymbol{\theta}} = (\bar{\gamma}, \bar{\boldsymbol{\beta}}')'$ and covariance matrix Δ. Let

$$\boldsymbol{\theta}_i = \bar{\boldsymbol{\theta}} + \boldsymbol{\alpha}_i, \tag{13.5.2}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \boldsymbol{\alpha}_{i2}')$. We have

$$E\boldsymbol{\alpha}_i = \mathbf{0}, E\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j' = \Delta \text{ if } i = j, \text{ and } \mathbf{0} \text{ otherwise,} \tag{13.5.3}$$

and[17]

$$E\boldsymbol{\alpha}_i \boldsymbol{x}_{jt}' = \mathbf{0}. \tag{13.5.4}$$

Stacking the T time series observations of the ith individuals in matrix form yields

$$\underset{T\times 1}{\boldsymbol{y}_i} = Q_i\boldsymbol{\theta}_i + \boldsymbol{u}_i, \quad i = 1, \ldots, N. \tag{13.5.5}$$

[16] We are concerned only with the estimation of the short-run adjustment coefficient $\bar{\gamma}$. For a discussion of estimating the long-run coefficient, see Pesaran and Smith (1995), Pesaran and Zhao (1999), Pesaran, Shin, and Smith (1999), and Phillips and Moon (1999, 2000).

[17] The strict exogeneity condition (13.5.4) of $\boldsymbol{x}_{it}$ is crucial in the identification of the dynamic random-coefficients model. Chamberlain (1993) has given an example of the lack of identification of γ in a model of the form

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

where x_{it} takes either 0 or 1. Because $E(\alpha_i \mid \boldsymbol{x}_i, \boldsymbol{y}_{i,-1})$ is unrestricted, the only moments that are relevant for the identification of γ are

$$E(\Delta y_{it} - \gamma\Delta y_{i,t-1} \mid \boldsymbol{x}_i^{t-1}, \boldsymbol{y}_i^{t-2}) = E(\beta_i \Delta x_{it} \mid \boldsymbol{x}_i^{t-1}, \boldsymbol{y}_i^{t-2}), t = 2, \ldots, T,$$

where $\boldsymbol{x}_i^t = (x_{i1}, \ldots, \boldsymbol{x}_{it}), \boldsymbol{y}_i^t = (y_{i0}, \ldots, y_{it})$. Let $\boldsymbol{w}_i^t = (\boldsymbol{x}_i^t, \boldsymbol{y}_i^t)$; the above expression is equivalent to the following two conditions:

$$\begin{aligned} &E(\Delta y_{it} - \gamma\Delta y_{i,t-1} \mid \boldsymbol{w}_i^{t-2}, \boldsymbol{x}_{i,t-1} = 0) \\ &\quad = E(\beta_i \mid \boldsymbol{w}_i^{t-2}, \boldsymbol{x}_{i,t-1} = 0) P_r(x_{it} = 1 \mid \boldsymbol{x}_i^{t-2}, x_{i,t-1} = 0), \end{aligned}$$

and

$$\begin{aligned} &E(\Delta y_{it} - \gamma\Delta y_{i,t-1} \mid \boldsymbol{x}_i^{t-2}, x_{i,t-1} = 1) \\ &\quad = -E(\boldsymbol{\beta}_i \mid \boldsymbol{w}_i^{t-2}, x_{i,t-1} = 1) P_r(x_{it} = 0 \mid \boldsymbol{w}_i^{t-2}, x_{i,t-1} = 1) \end{aligned}$$

If $E(\beta_i \mid \boldsymbol{w}_i^{t-2}, x_{i,t-1} = 0)$ and $E(\beta_i \mid \boldsymbol{w}_i^{t-2}, x_{i,t-1} = 1)$ are unrestricted and T is fixed, the autoregressive parameter γ cannot be identified from the above two equations.

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$, $Q_i = (\boldsymbol{y}_{i,-1}, X_i)$, $\boldsymbol{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $X_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iT})'$, and for ease of exposition, we assume that y_{i0} are observable.[18]

We note that because $y_{i,t-1}$ depends on γ_i, $EQ_i\boldsymbol{\alpha}_i' \neq \mathbf{0}$, i.e., the independence between the explanatory variables and $\boldsymbol{\alpha}_i$ (13.2.8) is violated. Substituting $\boldsymbol{\theta}_i = \bar{\boldsymbol{\theta}} + \boldsymbol{\alpha}_i$ into (13.5.5) yields

$$\boldsymbol{y}_i = Q_i\bar{\boldsymbol{\theta}} + \boldsymbol{v}_i, \quad i = 1, \ldots, N, \tag{13.5.6}$$

where

$$\boldsymbol{v}_i = Q_i\boldsymbol{\alpha}_i + \boldsymbol{u}_i. \tag{13.5.7}$$

Since

$$y_{i,t-1} = \sum_{j=0}^{\infty}(\bar{\gamma} + \alpha_{i1})^j \boldsymbol{x}'_{i,t-j-1}(\bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_{i2}) + \sum_{j=0}^{\infty}(\bar{\gamma} + \alpha_{i1})^j u_{i,t-j-1}, \tag{13.5.8}$$

it follows that $E(\boldsymbol{v}_i \mid Q_i) \neq \mathbf{0}$. Therefore, contrary to the static case, the least squares estimator of the common mean $\bar{\boldsymbol{\theta}}$ is inconsistent.

Equations (13.5.7) and (13.5.8) also demonstrate that the covariance matrix of $\boldsymbol{v}_i$, V, is not easily derivable. Thus, the procedure of premultiplying (13.5.6) by $V^{-1/2}$ to transform the model into the one with serially uncorrelated error is not implementable. Neither does the instrumental variable method appear implementable because the instruments that are uncorrelated with $\boldsymbol{v}_i$ are most likely uncorrelated with Q_i as well.

Pesaran and Smith (1995) have noted that as $T \to \infty$, the least squares regression of $\boldsymbol{y}_i$ on Q_i yields a consistent estimator of $\boldsymbol{\theta}_i$, $\hat{\boldsymbol{\theta}}_i$. They suggest a mean group estimator of $\bar{\boldsymbol{\theta}}$ by taking the average of $\hat{\boldsymbol{\theta}}_i$ across i,

$$\hat{\bar{\boldsymbol{\theta}}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_i. \tag{13.5.9}$$

The mean group estimator (13.5.9) is consistent and asymptotically normally distributed so long as $\sqrt{N}/T \to 0$ as both N and $T \to \infty$ (Hsiao, Pesaran, and Tahmiscioglu 1999). Hsiao et al. (2019) further show that the group mean estimator achieves the semiparametric efficiency bound (Chamberlain 1982).

However, panels with large T are typically the exception in economics. Nevertheless, under the assumption that y_{i0} are fixed and known and $\boldsymbol{\alpha}_i$ and u_{it} are independently normally distributed, we can implement the Bayes estimator of $\bar{\boldsymbol{\theta}}$ conditional on σ_i^2 and Δ using the formula (13.4.13), just as in the mixed-model case discussed in Section 13.4. The Bayes estimator condition on Δ and σ_i^2 is equal to

$$\hat{\bar{\boldsymbol{\theta}}}_B = \left\{\sum_{i=1}^{N}[\sigma_i^2(Q_i'Q_i)^{-1} + \Delta]^{-1}\right\}^{-1} \sum_{i=1}^{N}\left[\sigma_i^2(Q_i'Q_i)^{-1} + \Delta\right]^{-1}\hat{\boldsymbol{\theta}}_i, \tag{13.5.10}$$

which is a weighted average of the least squares estimator of individual units with the weights being inversely proportional to individual variances. When $T \to \infty, N \to \infty$, and $\sqrt{N}/T^{3/2} \to 0$, the Bayes estimator is asymptotically equivalent to the mean group estimator (13.5.9).

[18] We assume that $T (> 3)$ is large enough to identify γ and β. For an example of lack of identification when $T = 3$ and y_{it} is binary, see Chamberlain (1993) or Arellano and Honoré (2001); see also Chapter 6.

In practice, the variance components, σ_i^2 and Δ, are rarely known, so the Bayes estimator (13.5.10) is rarely feasible. One approach is to substitute the consistently estimated σ_i^2 and Δ, say (13.2.13) and (13.2.14), into the formula (13.5.10) and treat them as if they were known. For ease of reference, we shall call (13.5.10) with known σ_i^2 and Δ the infeasible Bayes estimator. We shall call the estimator obtained by substituting σ_i^2 and Δ in (13.5.10) by their consistent estimates, say (13.2.13) and (13.2.14), the empirical Bayes estimator.

The other approach is to follow Lindley and Smith (1972) by assuming that the prior distribution of σ_i^2 and Δ are independent and are distributed as

$$P(\Delta^{-1}, \sigma_1^2, \ldots, \sigma_N^2) = W(\Delta^{-1} \mid (\rho R)^{-1}, \rho) \prod_{i=1}^{N} \sigma_i^{-1}, \tag{13.5.11}$$

where W represents the Wishart distribution with scale matrix (ρR) and degrees of freedom ρ (e.g., Anderson 1985). Incorporating this prior into the models (13.5.1) and (13.5.2), we can obtain the marginal posterior densities of the parameters of interest by integrating out σ_i^2 and Δ from the joint posterior density. However, the required integrations do not yield closed form solutions. Hsiao, Pesaran, and Tahmiscioglu (1999) have suggested using the Gibbs sampler to calculate marginal densities.

The Gibbs sampler is an iterative Markov Chain Monte Carlo method which requires only the knowledge of the full conditional densities of the parameter vector (e.g., Gelfand and Smith 1990). Starting from some arbitrary initial values, say $(\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \ldots, \boldsymbol{\theta}_k^{(0)})$ for a parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, it samples alternatively from the conditional density of each component of the parameter vector conditional on the values of other components sampled in the latest iteration. That is:

(1) Sample $\boldsymbol{\theta}_1^{(j+1)}$ from $P(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \boldsymbol{y})$

(2) Sample $\boldsymbol{\theta}_2^{(j+1)}$ from $P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \boldsymbol{y})$

$\vdots$

(k) Sample $\boldsymbol{\theta}_k^{(j+1)}$ from $P(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(j+1)}, \boldsymbol{y})$

The vectors $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}$ form a Markov Chain, with transition probability from stage $\boldsymbol{\theta}^{(j)}$ to the next stage $\boldsymbol{\theta}^{(j+1)}$ being

$$\begin{aligned} K(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(j+1)}) = {} & P(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \boldsymbol{y}) P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \mathbf{y}) \\ & \ldots P(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(j+1)}, \boldsymbol{y}). \end{aligned}$$

As the number of iterations j approaches infinity, the sampled values in effect can be regarded as drawing from true joint and marginal posterior densities. Moreover, the ergodic averages of functions of the sample values will be a consistent estimation of their expected values.

Under the assumption that the prior of $\bar{\theta}$ is $N(\bar{\boldsymbol{\theta}}^*, \Psi)$, the relevant conditional distributions that are needed to implement the Gibbs sampler for (13.5.1) and (13.5.2) are easily obtained from

$$P(\boldsymbol{\theta}_i \mid \boldsymbol{y}, \bar{\boldsymbol{\theta}}, \Delta^{-1}, \sigma_1^2, \ldots, \sigma_N^2) \sim N\left\{A_i(\sigma_i^{-2} Q_i' \boldsymbol{y}_i + \Delta^{-1}\bar{\theta}), A_i\right\}, i = 1, \ldots, N,$$

$$P(\bar{\boldsymbol{\theta}} \mid \boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \Delta^{-1}, \sigma_1^2, \ldots, \sigma_N^2) \sim N\left\{D(N\Delta^{-1}\hat{\bar{\boldsymbol{\theta}}} + \Psi^{-1}\boldsymbol{\theta}^*), B\right\},$$

$$
\begin{aligned}
P(\Delta^{-1} \mid \boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \bar{\boldsymbol{\theta}}, \sigma_1^2, \ldots, \sigma_N^2) \\
\sim W\left[\left(\sum_{i=1}^{N} (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})' + \rho R\right)^{-1}, \rho + N\right], \\
P(\sigma_i^2 \mid \boldsymbol{y}_i, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \bar{\boldsymbol{\theta}}, \Delta^{-1}) \\
\sim IG[T/2, (\boldsymbol{y}_i - Q_i\boldsymbol{\theta}_i)'(\boldsymbol{y}_i - Q_i\boldsymbol{\theta}_i)/2], i = 1 \ldots, N,
\end{aligned}
$$

where $A_i = (\sigma_i^{-2} Q_i' Q_i + \Delta^{-1})^{-1}, D = (N\Delta^{-1} + \Psi^{-1})^{-1}$, $\hat{\bar{\boldsymbol{\theta}}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\theta}_i$, and IG denotes the inverse gamma distribution.

Hsiao, Pesaran, and Tahmiscioglu (1999) conducted Monte Carlo experiments to study the finite sample properties of (13.5.10), referred to as the infeasible Bayes estimator; the Bayes estimator obtained through the Gibbs sampler, referred to as the hierarchical Bayes estimator; the empirical Bayes estimator; the group mean estimator (13.5.9); the bias-corrected mean group estimator, obtained by directly correcting the finite T bias of the least squares estimator, $\hat{\boldsymbol{\theta}}_i$, using the formula of Kiviet (1995) and Kiviet and Phillips (1993), then taking the average; and the pooled least squares estimator. Table 13.2 presents the bias of the different estimators of $\bar{\gamma}$ for $N = 50$ and $T = 5$ or 20. The infeasible Bayes estimator performs very well. It has small bias even for $T = 5$. For $T = 5$, its bias falls within the range of 3%–17%. For $T = 20$, the bias is at most about 2%. The hierarchical Bayes estimator also performs well,[19] followed by the empirical when Bayes estimator T is small but improves quickly as T increases. The empirical Bayes estimator gives very good results even for $T = 5$ in some cases but the bias also appears to be quite high in certain other cases. As T gets larger, its bias decreases considerably. The mean group and the bias-corrected mean group estimator both have large bias when T is small, with the bias-

Table 13.2. *Bias of the short-run coefficient $\bar{\gamma}$*

			Bias					
T	$\bar{\gamma}$		Pooled OLS	Mean group	Bias-corrected mean group	Infeasible Bayes	Empirical Bayes	Hierarchicall Bayes
5	1	0.3	0.36859	−0.23613	−0.14068	0.05120	−0.12054	−0.02500
	2	0.3	0.41116	−0.23564	−0.14007	0.04740	−0.11151	−0.01500
	3	0.6	1.28029	−0.17924	−0.10969	0.05751	−0.02874	0.02884
	4	0.6	1.29490	−0.18339	−0.10830	0.06879	−0.00704	0.06465
	5	0.3	0.06347	−0.26087	−0.15550	0.01016	−0.18724	−0.10068
	6	0.3	0.08352	−0.26039	−0.15486	0.01141	−0.18073	−0.09544
	7	0.6	0.54756	−0.28781	−0.17283	0.05441	−0.12731	−0.02997
	8	0.6	0.57606	−0.28198	−0.16935	0.06258	−0.10366	−0.01012
20	9	0.3	0.44268	−0.07174	−0.01365	0.00340	−0.00238	0.00621
	10	0.3	0.49006	−0.06910	−0.01230	0.00498	−0.00106	0.00694
	11	0.35	0.25755	−0.06847	−0.01209	−0.00172	−0.01004	−0.00011
	12	0.35	0.25869	−0.06644	−0.01189	−0.00229	−0.00842	0.00116
	13	0.3	0.07199	−0.07966	−0.01508	−0.00054	−0.01637	−0.00494
	14	0.3	0.09342	−0.07659	−0.01282	0.00244	−0.01262	−0.00107
	15	0.55	0.26997	−0.09700	−0.02224	−0.00062	−0.01630	0.00011
	16	0.55	0.29863	−0.09448	−0.02174	−0.00053	−0.01352	0.00198

Source: Hsiao, Pesaran, and Tahmiscioglu (1999).

19 The $\Psi^{-1} = 0, \rho = 2$, and R equal to the Swamy estimate of Δ are used to implement the hierarchical Bayes estimator.

corrected mean group estimator performs slightly better. However, the performance of both improves as T increases, and both are still much better than the least squares estimator. The least squares estimator yields significant bias, and its bias persists as T increases.

The Bayes estimator is derived under the assumption that the initial observations, y_{i0}, are fixed constants. As discussed in Chapter 3 and by Anderson and Hsiao (1981, 1982), this assumption is clearly unjustifiable for a panel with finite T. However, contrary to the sampling approach where the correct modeling of initial observations is quite important, the Bayesian approach appears to perform fairly well in the estimation of the mean coefficients for dynamic random-coefficients models even the initial observations are treated as fixed constants. The Monte Carlo study also cautions against the practice of justifying the use of certain estimators based on their asymptotic properties here. Both the mean group and the corrected mean group estimators perform poorly in panels with very small T. The hierarchical Bayes estimator appears preferable to the other consistent estimators unless the time dimension of the panel is sufficiently large.

13.6 AN EXAMPLE: LIQUIDITY CONSTRAINTS AND FIRM INVESTMENT EXPENDITURE

The effects of financial constraints on company investment have been subject to intensive debate by economists. At one extreme, Jorgenson (1971) claims that "the evidence clearly favors the Modigliani–Miller theory (1958, 61). Internal liquidity is not an important determinant of the investment, given the level of output and external funds." At the other extreme, Stiglitz and Weiss (1981) argue that because of imperfections in the capital markets, costs of internal and external funds generally will diverge, and internal and external funds generally will not be perfect substitutes for each other. Fazzari, Hubbard, and Petersen (1988), Bond and Meghir (1994), and others tested for the importance of internal finance by studying the effects of cash flow across different groups of companies like identifying groups of firms according to company retention practices. If the null hypothesis of the perfect capital market is correct, then no difference should be found in the coefficient of cash flow variable across groups. However, these authors find that cash flow coefficient is large for companies with low dividend payout rates.

However, there is no sound theoretical basis for assuming that only low dividend payout companies are subject to financial constraints. The finding that larger companies have larger cash flow coefficients is inconsistent with both the transaction costs and asymmetric information explanations of liquidity constraints. Whether firm heterogeneity can be captured by grouping firms according to some indicators remains open to question.

Hsiao and Tahmiscioglu (1997) use COMPUSTAT annual industrial files of 561 firms in the manufacturing sector for the period 1971–1992 to estimate the following five different investment expenditure models with and without using liquidity models:

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i1}\left(\frac{LIQ}{K}\right)_{i,t-1} + \epsilon_{it}, \tag{13.6.1}$$

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i1}\left(\frac{LIQ}{K}\right)_{i,t-1} + \beta_{i2} q_{it} + \epsilon_{it}, \tag{13.6.2}$$

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i1}\left(\frac{LIQ}{K}\right)_{i,t-1} + \beta_{i2}\left(\frac{S}{K}\right)_{i,t-1} + \epsilon_{it}, \tag{13.6.3}$$

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i2} q_{it} + \epsilon_{it}, \tag{13.6.4}$$

Table 13.3. *Individual firm regressions (percentage of firms with significant coefficients)*

	Percentage of firms				
	Model 1	2	3	4	5
Coefficient for:					
$(LIQ/K)_{t-1}$	46	36	31		
q		31		38	
(S/K_{t-1})			27		44
Percentage of firms with significant autocorrelation	14	12	13	20	15
Actual F	2.47	2.98	2.01	2.66	2.11
Critical F	1.08	1.08	1.08	1.06	1.06

Note: The number of firms is 561. The significance level is 5 percent for a one-tailed test. Actual F is the F statistic for testing the equality of slope coefficients across firms. For the F test, the 5 percent significance level is chosen. To detect serial correlation, Durbin's t-test at the 5 percent significance level is used.
Source: Hsiao and Tahmiscioglu (1997, Table 1).

and

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i2} \left(\frac{S}{K}\right)_{i,t-1} + \epsilon_{it}. \tag{13.6.5}$$

where I_{it} is firm i's capital investment at time t, LIQ_{it} is a liquidity variable (defined as cash flow minus dividends); S_{it} is sales; q_{it} is Tobin's q (Brainard and Tobin 1968; Tobin 1969), defined as the ratio of the market value of the firm to the replacement value of capital; and K_{it} is the beginning-of-period capital stock. The coefficient β_{i1} measures the short-run impact of the liquidity variable on firm i's investment in each of these three specifications. Models 4 and 5 (Eqations 13.6.4 and 13.6.5) are two popular variants of investment equations that do not use the liquidity variable as an explanatory variable – the Tobin q model (e.g., Hayashi 1982; Summers 1981) and the sales capacity model (e.g., Kuh 1963). The sale variable can be regarded as a proxy for future demand for the firm's output. The q theory relates investment to marginal q, which is defined as the ratio of the market value of new investment goods to their replacement cost. If a firm has unexploited profit opportunities, then an increase of its capital stock of \$1 will increase its market value by more than \$1. Therefore, firm managers can be expected to increase investment until marginal q equals 1. Thus, investment will be an increasing function of marginal q. Because marginal q is unobservable, it is common in empirical work to substitute it with average or Tobin's q.

Tables 13.3 and 13.4 present some summary information from the firm-by-firm regressions of these five models. Table 13.3 shows the percentage of significant coefficients at the 5% significance level for a one-tailed test. Table 13.4 shows the first and third quartiles of the estimated coefficients. The estimated coefficients vary widely from firm to firm. The F-test of slope homogeneity across firms while allowing for firm-specific intercepts is also rejected (see Table 13.3).

The approach of relating the variation of β_{i1} to firm characteristics such as dividend payout rate, company size, sales growth, capital intensity, standard deviation of retained earnings, debt-equity ratio, measures of liquidity stocks from the balance sheet, number of shareholders and industry dummies is unsuccessful.[20] These variables as a whole do not explain the variation of estimated β_{i1} well. The maximum $\bar{R}^2$ is only 0.113. Many of

[20] For a discussion of coefficients as a function of some variables, see Section 13.8.

Table 13.4. *Coefficient heterogeneity: slope estimates at first and third quartiles across a sample of 561 firms*

	Slope estimates			
Model	$(\mathrm{I/K})_{i,t-1}$	$(\mathrm{LIQ/K})_{i,t-1}$	q_{it}	$(\mathrm{S/K})_{i,t-1}$
1	0.026, 0.405	0.127, 0.529		
2	−0.028, 0.359	0.062, 0.464	0, 0.039	
3	0.100, 0.295	0.020, 0.488		−0.005, 0.057
4	0.110, 0.459		0.007, 0.048	
5	−0.935, 0.367			0.012, 0.077

Source: Hsiao and Tahmiscioglu (1997, Table 2).

Table 13.5. *Variable intercept estimation of models for less- and more capital-intensive firms*

	Variable intercept estimate					
Variable	Less-capital-intensive firms			More-capital-intensive firms		
$(\mathrm{I/K})_{i,t-1}$	0.265	0.198	0.248	0.392	0.363	0.364
	(0.011)	(0.012)	(0.011)	(0.022)	(0.023)	(0.022)
$(\mathrm{LIQ/K})_{i,t-1}$	0.0161	0.110	0.119	0.308	0.253	0.0278
	(0.007)	(0.007)	(0.007)	(0.024)	(0.027)	(0.025)
$(\mathrm{S/K})_{i,t-1}$		0.023			0.025	
		(0.001)			(0.006)	
q_{it}			0.011			0.009
			(0.0006)			(0.002)
Actual F	2.04	1.84	2.22	2.50	2.19	2.10
Critical F	1.09	1.07	1.07	1.20	1.17	1.17
Numerator d.f.	834	1251	1251	170	255	255
Denominator d.f.	6592	6174	6174	1368	1282	1282
Number of firms	418	418	418	86	86	86

Note: The dependent variable is $(\mathrm{I/K})_{it}$. Less-capital-intensive firms are those with minimum (K/S) between 0.15 and 0.55 over the sample period. For more-capital-intensive firms, the minimum (K/S) is greater than 0.55. The regressions include company-specific intercepts. Actual F is the F statistic for testing the homogeneity of slope coefficients. For the F test, a 5 percent significance level is chosen. The estimation period is 1974–1992. Standard errors are in parentheses.
Source: Hsiao and Tahmiscioglu (1997, Table 5).

the estimated coefficients are not significant under various specifications. So Hsiao and Tahmiscioglu (1997) classify firms into reasonably homogeneous groups using the capital intensity ratio of 0.55 as a cutoff point. Capital intensity is defined as the minimum value of the ratio of capital stock to sales over the sample period. It is the most statistically significant and most stable variable under different specifications.

Table 13.5 presents the variable intercept estimates for the groups of less- and more-capital intensive firms. The liquidity variable is highly significant in all three variants of the liquidity model. There are also significant differences in the coefficients of the liquidity variable across the two groups. However, Table 13.5 also shows that the null hypothesis of the equality of slope coefficients conditioning on the firm-specific effects are strongly rejected for all specifications for both groups. In other words, using the capital intensity ratio of 0.55 as a cutoff point, there is still substantial heterogeneity within the groups.

Since there appears not to be a set of explanatory variables that adequately explains the variation of β_{i1}, nor can homogeneity be achieved by classifying firms into groups, one is left with either treating $\boldsymbol{\beta}_i$ as fixed and different, or treating $\boldsymbol{\beta}_i$ as random draws from

Table 13.6. *Estimation of mixed fixed- and random-coefficient models for less- and more-capital-intensive firms*

	Estimate					
Variable	Less-capital-intensive firms			More-capital-intensive firms		
$(\text{I/K})_{i,t-1}$	0.230 (0.018)	0.183 (0.017)	0.121 (0.019)	0.321 (0.036)	0.302 (0.037)	0.236 (0.041)
$(\text{LIQ/K})_{i,t-1}$	0.306 (0.021)	0.252 (0.023)	0.239 (0.027)	0.488 (0.065)	0.449 (0.067)	0.416 (0.079)
$(\text{S/K})_{i,t-1}$		0.024 (0.003)			0.038 (0.015)	
q_{it}		0.019 (0.003)			0.022 (0.008)	
Number of firms	418	418	418	86	86	86

Note: The dependent variable is $(\text{I/K})_{it}$. The regressions include fixed firm-specific effects. The estimation period is 1974–1992. Standard errors are in parentheses.
Source: Hsiao and Tahmiscioglu (1997, Table 7).

a common distribution. Within the random-effects framework, individual differences are viewed as random draws from a population with constant mean and variance. Therefore, it is appropriate to pool the data and try to draw some generalization about the population. On the other hand, if individual differences reflect fundamental heterogeneity or if individual response coefficients depend on the values of the included explanatory variables, estimation of the model parameters based on the conventional random effects formulation can be misleading. To avoid this bias, heterogeneity among individuals must be treated as fixed. In other words, one must investigate investment behavior firm by firm, and there is no advantage of pooling. Without pooling, the shortage of degrees of freedom and multicollinearity can render the resulting estimates meaningless and make drawing general conclusions difficult.

Table 13.6 presents the estimates of the mixed fixed- and random-coefficients model of the form (13.4.5) by assuming that conditional on company-specific effects, the remaining slope coefficients are randomly distributed around a certain mean within each of less- and more- capital intensive groups. To evaluate the appropriateness of these specifications, Table 13.7 presents the comparison of the recursive predictive density of the mixed fixed- and random-coefficients models and the fixed-coefficients model assuming that each company has different coefficients for the three variants of the liquidity model by dividing the sample into pre- and post-1989 periods. The numbers reported in Table 13.7 are the logarithms of (13.4.27). The results indicate that the mixed fixed- and random-coefficients model is favored over the fixed-coefficients model for both groups. Similar comparison between the liquidity model, Tobin's q, and sales accelerator models also favor liquidity as an important explanatory variable.

Table 13.6 shows that the estimated liquidity coefficients are highly significant, and there are significant differences between different classes of companies. The mean coefficient of the liquidity variable turns out to be 60%–80% larger for the more-capital-intensive group than for the less-capital-intensive group. The implied long-run relationships between the liquidity variable and the fixed investment variable are also statistically significant. For instance, for model (13.6.1), a 10% increase in liquidity capital ratio leads to a 4% increase in fixed investment capital ratio in the long run for the less-capital-intensive group compared to a 7% increase in the ratio for the more-capital-intensive group. The mixed model also yields substantially larger coefficient estimates of the liquidity variable then

Table 13.7. *Prediction Comparison of Fixed-Coefficients and Mixed Fixed and Random-Coefficients Models for Less- and More-Capital-Intensive Firms (Recursive Predictive Density)*

Sample	Model	Liquidity	Liquidity with q	Liquidity with sales
Less-capital intensive firms	Fixed-slope coefficients	2,244	2,178	2,172
	Random-slope coefficients	2,299	2,272	2,266
More-capital intensive firms	Fixed-slope coefficients	587	544	557
	Random-slope coefficients	589	556	576

Note: The recursive predictive density is the logarithm of (13.4.27). The fixed-coefficients model assumes different coefficients for each firm. The random-coefficients model assumes randomly distributed slope coefficients with constant mean conditional on fixed firm-specific effects. The prediction period is 1990–1992.
Source: Hsiao and Tahmiscioglu (1997, Table 6).

those obtained from the variable-intercept model. If the coefficients are indeed randomly distributed and the explanatory variables are positively autocorrelated, then this is precisely what one would expect from the within estimates (Pesaran and Smith 1995).

In short, there are substantial differences across firms in their investment behavior. When these differences are ignored by constraining the parameters to be identical across firms, the impact of the liquidity variable on firm investment is seriously underestimated. The mixed fixed- and random-coefficients model appears to fit the data well. The mixed model allows pooling and allows some general conclusions to be drawn about a group of firms. The estimation results and prediction tests appear to show that financial constraints are the most important factor affecting actual investment expenditure, at least for a subset of U.S. manufacturing companies.

13.7 COEFFICIENTS THAT VARY OVER TIME

There is a large amount of empirical evidence that parameters of a model change over time. For instance, financial liberalization or changes in monetary policy can cause the relationships between economic variables to alter. If a constant parameter model is used, misspecification may occur. On the other hand, if a model is too flexible in its treatment of parameter change, overfitting or imprecise inferences can occur. In this section, we discuss some commonly used time-varying parameter models that entail a smooth evolvement.

In most models with coefficients evolving over time, it is assumed that there is no individual heterogeneity (e.g., Zellner, Hong, and Min 1991). At a given t, the coefficient vectors $\boldsymbol{\beta}_t$ are identical for all cross-sectional units. For this reason we shall discuss the main issues of time-varying-parameter models assuming that $N = 1$, then indicate how this analysis can be modified when $N > 1$.

13.7.1 Parameters Slowly Evolving over Time

Sun et al. (2020) consider the parameters of a model

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_t + u_t, \quad t = 1, \ldots, T. \tag{13.7.1}$$

where $\boldsymbol{\beta}_t$ are slowly evolving over time in the sense that $\boldsymbol{\beta}_t$ is a smooth function of the ratio (t/T) as in Cai (2007), Chen and Hong (2012), and Robinson (1989),

$$\boldsymbol{\beta}_s \approx \boldsymbol{\beta}_t, \quad s \in (t - Th, t + Th). \tag{13.7.2}$$

Then $\boldsymbol{\beta}_t$ can be estimated by

$$\hat{\boldsymbol{\beta}}_t = (X'K_tX)^{-1}(X'K_t\boldsymbol{y}), \tag{13.7.3}$$

where $\boldsymbol{y} = (y_1, \ldots, y_T)'$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)'$, and K_t is a smooth kernel, $K_t = \text{diag}\{k_{1t}, \ldots, k_{Tt}\}$ with $k_{st} = k(\frac{s-t}{Th})$ being a prespecified symmetric probability density function and h is a bandwidth parameter such that $h \to 0$ and $Th \to \infty$ as $T \to \infty$. To reduce the bias of $\hat{\boldsymbol{\beta}}_t$, Sun et al. (2020) suggest using a jackknife estimator in lieu of (13.7.3) with $K_{-t} = \text{diag}\{k_{1,t}, k_{2,t}, \ldots, k_{(t-1),t}, 0, k_{t+1,t}, \ldots, k_{T,t}\}$

$$\hat{\boldsymbol{\beta}}_t = (X'K_{-t}X)^{-1}(X'K_{-t}\boldsymbol{y}). \tag{13.7.4}$$

13.7.2 Parameters Randomly Vary over Time[21]

13.7.2.1 Models

As shown by Chow (1983, chapter 10), a wide variety of time-varying-parameter models can be put in the general form

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta}_t + \boldsymbol{u}_t, \tag{13.7.5}$$

and

$$\boldsymbol{\beta}_t = H\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad t = 1, \ldots, T, \tag{13.7.6}$$

where $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})'$, $X_t = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt})'$, $\boldsymbol{u}_t = (u_{1t}, \ldots, u_{Nt})'$, $\boldsymbol{x}_{it}$ is a $K \times 1$ vector of exogenous variables, $\boldsymbol{u}_t$ is independently normally distributed over t with mean zero and variance Ω, $\boldsymbol{\eta}_t$ is a K-variant independent normal random variable, with mean zero and covariance matrix Ψ, and $\boldsymbol{\eta}$ and $\boldsymbol{u}$ are independent. For instance, when $H = I_K$, it is the random-walk model of Cooley and Prescott (1976). When $H = I_K$ and $\Psi = \mathbf{0}$, this model is reduced to the standard regression model.

The Rosenberg (1972, 1973) return-to-normality model can also be put in this form. The model corresponds to replacing $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t-1}$ in (13.7.6) by $(\boldsymbol{\beta}_t - \boldsymbol{\beta})$ and $(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta})$ and restricting the absolute value of the characteristic roots of H to less than 1. Although this somewhat changes the formulation, if we define $\boldsymbol{\beta}_t^* = \boldsymbol{\beta}_t - \boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}_t = \boldsymbol{\beta}$, the return-to-normality model can be rewritten as

$$\begin{aligned} \boldsymbol{y}_t &= (X_t, X_t)\begin{bmatrix} \bar{\boldsymbol{\beta}}_t \\ \\ \boldsymbol{\beta}_t^* \end{bmatrix} + \boldsymbol{u}_t, \\ \begin{bmatrix} \bar{\boldsymbol{\beta}}_t \\ \boldsymbol{\beta}_t^* \end{bmatrix} &= \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & H \end{bmatrix}\begin{bmatrix} \bar{\boldsymbol{\beta}}_{t-1} \\ \boldsymbol{\beta}_{t-1}^* \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\eta}_t \end{bmatrix}, \end{aligned} \tag{13.7.7}$$

which becomes a special case of (13.7.5) and (13.7.6).

[21] This section is largely drawn from the work of Chow (1983, chapter 10).

Similarly, we can allow $\boldsymbol{\beta}_t$ to be stationary, with constant mean $\boldsymbol{\beta}$ (Pagan 1980). Suppose

$$\begin{aligned} \boldsymbol{y}_t &= X_t\boldsymbol{\beta} + X_t\boldsymbol{\beta}_t^* + \boldsymbol{u}_t, \\ \boldsymbol{\beta}_t^* &= \boldsymbol{\beta}_t - \boldsymbol{\beta} = A^{-1}(\mathcal{L})\boldsymbol{\epsilon}_t, \end{aligned} \tag{13.7.8}$$

where $A(\mathcal{L})$ is a ratio of polynomials of orders p and q in the lag operator $\mathcal{L}(\mathcal{L}\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_{t-1})$, and $\boldsymbol{\epsilon}$ is independent normal, so that $\boldsymbol{\beta}_t^*$ follows an autoregressive moving-average (ARMA) (p, q) process. Because an ARMA of order p and q can be written as a first-order autoregressive process, this model can again be put in the form of (13.7.5) and (13.7.6). For example,

$$\boldsymbol{\beta}_t^* = B_1\boldsymbol{\beta}_{t-1}^* + B_2\boldsymbol{\beta}_{t-2}^* + \boldsymbol{\epsilon}_t + B_3\boldsymbol{\epsilon}_{t-1} \tag{13.7.9}$$

can be written as

$$\tilde{\boldsymbol{\beta}}_t^* = \begin{bmatrix} \boldsymbol{\beta}_t^* \\ \boldsymbol{\beta}_{t-1}^* \\ \boldsymbol{\epsilon}_t \end{bmatrix} = \begin{bmatrix} B_1 & B_2 & B_3 \\ I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1}^* \\ \boldsymbol{\beta}_{t-2}^* \\ \boldsymbol{\epsilon}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_t \\ \mathbf{0} \\ \boldsymbol{\epsilon}_t \end{bmatrix} = H\tilde{\boldsymbol{\beta}}_{t-1}^* + \boldsymbol{\eta}_t. \tag{13.7.10}$$

Thus, we can write Pagan's model in the form

$$\boldsymbol{y}_t = (X_t, \tilde{X}_t) \begin{bmatrix} \bar{\boldsymbol{\beta}}_t \\ \tilde{\boldsymbol{\beta}}_t^* \end{bmatrix} + \boldsymbol{u}_t, \tag{13.7.11}$$

where $\tilde{X}_t' = (X_t, \mathbf{0}', \mathbf{0}')$. Equation (13.7.11) is then formally equivalent to (13.7.7).

The Kalman filter (Kalman 1960) provides a basis for computing the maximum likelihood estimators and predicting the evolution of the time path of $\boldsymbol{\beta}_t$ for this type of the model. In this section we first consider the problem of estimating $\boldsymbol{\beta}_t$ using information $\mathcal{I}_s$, up to the time s, assuming that σ_u^2, Ψ, and H are known. We denote the conditional expectation of $\boldsymbol{\beta}_t$, given $\mathcal{I}_s$, as $E(\boldsymbol{\beta}_t \mid \mathcal{I}_s) = \boldsymbol{\beta}_{t|s}$. The evaluation of $\boldsymbol{\beta}_{t|s}$ is called filtering when $t = s$; it is called smoothing when $s > t$; it is called prediction when $s < t$. We then study the problem of estimating σ_u^2, Ψ, and H by the method of maximum likelihood. Finally, we consider the problem of testing for constancy of the parameters.

13.7.2.2 Predicting $\boldsymbol{\beta}_t$ by the Kalman Filter

For a panel data analysis of models (13.7.5) and (13.7.6), let $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})'$, $X_t = (\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt})'$, $Y_t = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$, $\tilde{X}_t = (X_1, \ldots, X_t)$. We assume that $(\boldsymbol{u}_t, \boldsymbol{\eta}_t)'$ are independently normally distributed over t with block diagonal covariance matrix

$$\begin{pmatrix} \Omega & \mathbf{0} \\ \mathbf{0}' & \Psi \end{pmatrix}, \tag{13.7.12}$$

and $E[(\boldsymbol{u}_t', \boldsymbol{\eta}_t')' | X_s] = \mathbf{0}$ for any s.

Let $\boldsymbol{\beta}_{t|t} = E(\boldsymbol{\beta}_t | Y_t, \tilde{X}_t)$ and $\boldsymbol{\beta}_{t|t-j} = E(\boldsymbol{\beta}_t | Y_{t-j}, \tilde{X}_{t-j})$, then from (13.7.6)

$$\boldsymbol{\beta}_{t|t-j} = H^j \boldsymbol{\beta}_{t-j|t-j}, \tag{13.7.13}$$

and

$$
\begin{aligned}
\boldsymbol{\beta}_{t|t} &= E(\boldsymbol{\beta}_t | Y_t, \tilde{X}_t) \\
&= E(\boldsymbol{\beta}_t | Y_{t-1}, \tilde{X}_{t-1}) + L_t(\boldsymbol{y}_t - E(\boldsymbol{y}_t | Y_{t-1}, \tilde{X}_t)) \\
&= \boldsymbol{\beta}_{t|t-1} + L_t(\boldsymbol{y}_t - X_t E(\boldsymbol{\beta}_t | Y_{t-1}, \tilde{X}_{t-1})) \\
&= H\boldsymbol{\beta}_{t-1|t-1} + L_t(\boldsymbol{y}_t - X_t \boldsymbol{\beta}_{t|t-1}).
\end{aligned} \tag{13.7.14}
$$

where $\boldsymbol{y}_t - E(\boldsymbol{y}_t | Y_{t-1}, \tilde{X}_t)$ denotes the additional information due to $(\boldsymbol{y}_t, X_t)$ and L_t denotes the adjustment factor from $\boldsymbol{\beta}_{t|t-1}$ because of this additional information. If L_t is known, (13.7.14) can be used to update our estimate $\boldsymbol{\beta}_{t|t-1}$ to form $\boldsymbol{\beta}_{t|t}$.

To derive L_t from the normality assumption on $\boldsymbol{\eta}_t$ and $\boldsymbol{u}_t$, we know that $(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})$ and $(\boldsymbol{y}_t - X_t\boldsymbol{\beta}_{t-1|t-1})$ are jointly normally distributed. The normal-distribution theory (Anderson 1985, chapter 2) states that,

$$
L_t = [E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})(\boldsymbol{y}_t - E(\boldsymbol{y}_t \mid Y_{t-1}, \tilde{X}_t))']\mathrm{Var}(\boldsymbol{y}_t \mid Y_{t-1}, \tilde{X}_t)^{-1}. \tag{13.7.15}
$$

Noting that $E(\boldsymbol{y}_t |, Y_{t-1}, \tilde{X}_t) = X_t\boldsymbol{\beta}_{t|t-1}$ and denoting the covariance matrix $\mathrm{Cov}(\boldsymbol{\beta}_t \mid Y_{t-1}, \tilde{X}_t) = E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})'$ by $\Sigma_{t|t-1}$, we have

$$
\begin{aligned}
&E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})(\boldsymbol{y}_t - X_t\boldsymbol{\beta}_{t|t-1})' \\
&\quad = E\{(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})'X_t' + \boldsymbol{u}_t']\} = \Sigma_{t|t-1}X_t',
\end{aligned} \tag{13.7.16}
$$

and

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{y}_t \mid Y_{t-1}, \tilde{X}_t) &= E[X_t(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}) + \boldsymbol{u}_t][(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})'X_t' + \boldsymbol{u}_t'] \\
&= X_t\Sigma_{t|t-1}X_t' + \Omega.
\end{aligned} \tag{13.7.17}
$$

Hence, (13.7.15) becomes

$$
L_t = \Sigma_{t|t-1}X_t'(X_t\Sigma_{t|t-1}X_t' + \Omega)^{-1}. \tag{13.7.18}
$$

From (13.7.13) we have

$$
\boldsymbol{\beta}_{t|t-1} = H\boldsymbol{\beta}_{t-1|t-1}. \tag{13.7.19}
$$

Thus, we can compute $\Sigma_{t|t-1}$ recursively by

$$
\begin{aligned}
\Sigma_{t|t-1} &= E(\boldsymbol{\beta}_t - H\boldsymbol{\beta}_{t-1|t-1})(\boldsymbol{\beta}_t - H\boldsymbol{\beta}_{t-1|t-1})' \\
&= E[H(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1}) + \boldsymbol{\eta}_t] \\
&\quad \times [H(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1}) + \boldsymbol{\eta}_t]' \\
&= H\Sigma_{t-1|t-1}H' + \Psi.
\end{aligned} \tag{13.7.20}
$$

Next, from (13.7.14) we can write

$$
\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t} = \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1} - L_t[X_t(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}) + \boldsymbol{u}_t]. \tag{13.7.21}
$$

Taking the expectation of the product of (13.7.21) and its transpose, and using (13.7.18), we obtain

$$
\begin{aligned}
\Sigma_{t|t} &= \Sigma_{t|t-1} - L_t(X_t\Sigma_{t|t-1}X_t' + \Omega)L_t' \\
&= \Sigma_{t|t-1} - \Sigma_{t|t-1}X_t'(X_t\Sigma_{t|t-1}X_t' + \Omega)^{-1}X_t\Sigma_{t|t-1}.
\end{aligned} \tag{13.7.22}
$$

Equations (13.7.20) and (13.7.22) can be used to compute $\Sigma_{t|t}(t = 1, 2, \ldots)$ successively given $\Sigma_{0|0}$. Having computed $\Sigma_{t|t-1}$, we can use (13.7.18) to compute L_t. Given L_t, (13.7.19), and (13.7.20), (13.7.14) can be used to compute $\boldsymbol{\beta}_{t|t}$ from $\boldsymbol{\beta}_{t-1|t-1}$ if $\boldsymbol{\beta}_{0|0}$ is known.

Similarly, we can predict $\boldsymbol{\beta}_t$ using future observations $\boldsymbol{y}_{t+1}, \boldsymbol{y}_{t+2}, \ldots, \boldsymbol{y}_{t+n}$. We first consider the regression of $\boldsymbol{\beta}_t$ on $\boldsymbol{y}_{t+1}$, conditional on Y_t. Analogous to (13.7.14) and (13.7.18) are

$$\boldsymbol{\beta}_{t|t+1} = \boldsymbol{\beta}_{t|t} + F_{t|t+1}(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1|t}) \tag{13.7.23}$$

and

$$F_{t|t+1} = [E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t})(\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1|t})'][\mathrm{Cov}(\boldsymbol{y}_{t+1} \mid Y_t, \tilde{X}_t)]^{-1}. \tag{13.7.24}$$

To derive the matrix $F_{t|t+1}$ of regression coefficients, we use (13.7.5) and (13.7.6) to write

$$\begin{aligned} \boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1|t} &= X_{t+1}(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{t+1|t}) + \boldsymbol{u}_{t+1} \\ &= X_{t+1}H(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t}) + X_{t+1}\boldsymbol{\eta}_{t+1} + \boldsymbol{u}_{t+1}. \end{aligned} \tag{13.7.25}$$

Combining (13.7.24), (13.7.25), (13.7.17), and (13.7.18), we have

$$\begin{aligned} F_{t|t+1} &= \Sigma_{t|t}H'X'_{t+1}(X_{t+1}\Sigma_{t+1|t}X'_{t+1} + \Omega)^{-1} \\ &= \Sigma_{t|t}H'\Sigma^{-1}_{t+1|t}L_{t+1}. \end{aligned} \tag{13.7.26}$$

Therefore, from (13.7.26) and (13.7.21), we can rewrite (13.7.23) as

$$\boldsymbol{\beta}_{t|t+1} = \boldsymbol{\beta}_{t|t} + \Sigma_{t|t}H'\Sigma^{-1}_{t+1|t}(\boldsymbol{\beta}_{t+1|t+1} - \boldsymbol{\beta}_{t+1|t}). \tag{13.7.27}$$

Equation (13.7.27) can be generalized to predict $\boldsymbol{\beta}_t$ using future observations $\boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_{t+n}$,

$$\boldsymbol{\beta}_{t|t+n} = \boldsymbol{\beta}_{t|t+n-1} + F^*_t(\boldsymbol{\beta}_{t+1|t+n} - \boldsymbol{\beta}_{t+1|t+n-1}), \tag{13.7.28}$$

where $F^*_t = \Sigma_{t|t}H'\Sigma^{-1}_{t+1|t}$. The proof of this is given by Chow (1983, chapter 10).

When H, Ψ, and Ω are known, (13.7.14) and (13.7.28) trace out the time path of $\boldsymbol{\beta}_t$ and provide the minimum-mean-square-error forecast of the future values of the dependent variable, given the initial values $\boldsymbol{\beta}_{0|0}$ and $\Sigma_{0|0}$. To obtain the initial values of $\boldsymbol{\beta}_{0|0}$ and $\Sigma_{0|0}$, Sant (1977) suggested using the generalized least squares method on the first K observations of $\boldsymbol{y}_t$ and X_t. Noting that

$$\begin{aligned} \boldsymbol{\beta}_t &= H\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \\ &= H^2\boldsymbol{\beta}_{t-2} + \boldsymbol{\eta}_t + H\boldsymbol{\eta}_{t-1} \\ &= H^{t-j}\boldsymbol{\beta}_j + \boldsymbol{\eta}_t + H\boldsymbol{\eta}_{t-1} + \cdots + H^{t-j-1}\boldsymbol{\eta}_{j+1}, \end{aligned} \tag{13.7.29}$$

and assuming that H^{-1} exists, we can also write $\boldsymbol{y}_k$ in the form

$$\begin{aligned} \boldsymbol{y}_k &= X_k\boldsymbol{\beta}_k + \boldsymbol{u}_k \\ &= X_k[H^{-K+k}\boldsymbol{\beta}_K - H^{-K+k}\boldsymbol{\eta}_K - \cdots - H^{-1}\boldsymbol{\eta}_{k+1}] + \boldsymbol{u}_k. \end{aligned}$$

Thus, $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$ can be written as

$$
\begin{aligned}
\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_K \end{bmatrix} &= \begin{bmatrix} X_1 H^{-K+1} \\ X_2 H^{-K+2} \\ \vdots \\ X_K \end{bmatrix} \boldsymbol{\beta}_K + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_K \end{bmatrix} \\
&\quad - \begin{bmatrix} X_1 H^{-1} & X_1 H^{-2} & \ldots & X_1 H^{-K+1} \\ \mathbf{0}' & X_2 H^{-1} & \ldots & X_2 H^{-K+2} \\ & & \ldots & \cdot \\ & & & X_{K-1} H^{-1} \\ & & & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \\ \vdots \\ \boldsymbol{\eta}_K \end{bmatrix}.
\end{aligned}
\tag{13.7.30}
$$

Applying GLS to (13.7.30) gives

$$
\begin{aligned}
\Sigma_{K|K} = \{&[H'^{-K+1} X_1, H'^{-K+2} X_2, \ldots, X_K] \\
&\cdot [I_K + A_K (I_{K-1} \otimes P) A_K']^{-1} [H^{-K+1} X_1, \ldots, X_K]'\}^{-1}
\end{aligned}
\tag{13.7.31}
$$

and

$$
\begin{aligned}
\boldsymbol{\beta}_{K|K} = &\frac{1}{\sigma_u^2} \Sigma_{K|K} [H'^{-K+1} X_1, H'^{-K+2} X_2, \ldots, X_K] \\
&\cdot [I_K + A_K (I_{K-1} \otimes P) A_K']^{-1} \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_K \end{bmatrix},
\end{aligned}
\tag{13.7.32}
$$

where $P = \Omega^{-1/2}\Psi$, and A_K is the coefficient matrix of $(\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K)'$ in (13.7.30). The initial estimators, $\boldsymbol{\beta}_{K|K}$ and $\Sigma_{K|K}$, are functions of Ω, Ψ, and H.

13.7.3 Maximum Likelihood Estimation

When H, Ψ, and Ω are unknown, (13.7.14) opens the way for maximum likelihood estimation without the need for repeated inversions of covariance matrices of large dimensions. To form the likelihood function, we note that

$$
\boldsymbol{y}_t - \boldsymbol{y}_{t|t-1} = X_t(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}) + \boldsymbol{u}_t = \boldsymbol{y}_t - X_t \boldsymbol{\beta}_{t|t-1} \tag{13.7.33}
$$

is normal and serially uncorrelated. Hence, the joint density of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ can be written as the product of the conditional density of $(\boldsymbol{y}_{K+1}, \ldots, \boldsymbol{y}_T \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$ and the marginal density of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$. The log-likelihood function of $(\boldsymbol{y}_{K+1}, \ldots, \boldsymbol{y}_T)$, given $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K)$, is

$$
\begin{aligned}
\log L = &- \frac{T-K}{2} \log 2\pi - \frac{1}{2} \sum_{t=K+1}^{T} \log |X_t \Sigma_{t|t-1} X_t' + \Omega| \\
&- \frac{1}{2} \sum_{t=K+1}^{T} (\boldsymbol{y}_t - X_t \boldsymbol{\beta}_{t|t-1})' (X_t \Sigma_{t|t-1} X_t' + \Omega)^{-1} (\boldsymbol{y}_t - X_t \boldsymbol{\beta}_{t|t-1}).
\end{aligned}
\tag{13.7.34}
$$

The first K observations are used to compute $\Sigma_{K|K}$ and $\boldsymbol{\beta}_{K|K}$ (Equations 13.7.31 and 13.7.32) as functions of Ω, Ψ, and H. Hence, the data $\boldsymbol{\beta}_{t|t-1}$ and $\Sigma_{t|t-1}$ $(t = K+1, \ldots, T)$

required to evaluate log L are functions of Ω, Ψ, and H, as given by (13.7.18)–(13.7.22). To find the maximum of (13.7.34) is tedious, so numerical methods will have to be used. For details with regard to the computation of MLE, see Harvey (1978) and Harvey and Phillips (1982).

13.7.4 Tests for Parameter Constancy

A direct test of the constant parameter model is to test $\Psi = \mathbf{0}$. However, the probability distribution of $\hat{\Psi}$ is very complicated under the null of $\Psi = \mathbf{0}$. An indirect test is to note that if $\Psi = \mathbf{0}$, it is equivalent to the null

$$H_0\colon \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T = \boldsymbol{\beta}. \tag{13.7.35}$$

An F test for the null (13.7.35) can be constructed like that discussed in Section 2.6. Alternatively, we can use a likelihood ratio test statistic

$$2\left\{\left[\log L_t(\hat{\boldsymbol{\beta}}_t) + \log L_2(\hat{\boldsymbol{\beta}}_2) + \cdots + \log L_T(\hat{\boldsymbol{\beta}}_T)\right] - \log L(\hat{\boldsymbol{\beta}})\right\}, \tag{13.7.36}$$

where $\log L_t(\hat{\boldsymbol{\beta}}_t)$ denotes the maximum log-likelihood value of the function,

$$-\frac{1}{2}\log|\Omega| - \frac{1}{2}(\boldsymbol{y}_t - X_t\boldsymbol{\beta})'\Omega^{-1}(\boldsymbol{y}_t - X_t\boldsymbol{\beta}_t), \tag{13.7.37}$$

and log $L(\hat{\boldsymbol{\beta}})$ denotes the maximum log-likelihood value of the function,

$$-\frac{T}{2}\log|\Omega| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t - X_t\boldsymbol{\beta})'\Omega^{-1}(\boldsymbol{y}_t - X_t\boldsymbol{\beta}). \tag{13.7.38}$$

When N is large, (13.7.36) is asymptotically chi-square distributed with $(T-1)K$ degrees of freedom under the null.

If the null hypothesis is rejected, we can use the information that under mild regularity conditions $\text{plim}_{N\to\infty}\hat{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t, t = 1, \ldots, T$, to investigate the nature of variation in the parameters over time or to conduct the Chen and Hong (2012) test for smooth versus abrupt structural change. Or we can apply the Box–Jenkins (1970) method on $\hat{\boldsymbol{\beta}}_t$ to identify a suitable stochastic process with which to model the parameter variation, i.e., the form of H in (13.7.6).

13.8 COEFFICIENTS THAT ARE FUNCTIONS OF OTHER EXOGENOUS VARIABLES

Sometimes, instead of assuming that parameters are random draws from a common distribution, an investigation of possible dependence of $\boldsymbol{\beta}_{it}$ on characteristics of the "individuals" or "time" is of considerable interest (e.g., Amemiya 1978; Hendricks, Koenker, and Poirier 1979; Singh et al. 1976; Swamy and Mehta 1977; Wachter 1970). A general formulation of stochastic-parameter models with systematic components can be expressed within the context of the linear model. Suppose that

$$\boldsymbol{y}_i = X_{i1}\boldsymbol{\beta}_1 + X_{i2}\boldsymbol{\beta}_{2i} + \boldsymbol{u}_i, \quad i = 1, \ldots, N, \tag{13.8.1}$$

and

$$\boldsymbol{\beta}_{2i} = Z_i\boldsymbol{\gamma} + \boldsymbol{\eta}_{2i}, \tag{13.8.2}$$

where X_{i1} and X_{i2} denote the $T \times K_1$ and $T \times K_2$ matrices of the time series observations of the first K_1 and last $K_2 (= K - K_1)$ exogenous variables for the ith individual, $\boldsymbol{\beta}_1$ is a $K_1 \times 1$ vector of fixed constants, $\boldsymbol{\beta}_{2i}$ is a $K_2 \times 1$ vector that varies according to (13.8.2), Z_i and $\boldsymbol{\gamma}$ are a $K_2 \times M$ matrix of known constants and a $M \times 1$ vector of unknown constants, respectively, and $\boldsymbol{u}_i$ and $\boldsymbol{\eta}_{2i}$ are $T \times 1$ and $K_2 \times 1$ vectors of unobservable random variables that are assumed independent of X_i and Z_i. For example, in Wachter (1970), $\boldsymbol{y}_i$ is a vector of time-series observations on the logarithm of the relative wage rate in the ith industry. X_{i1} contains the logarithm of such variables as the relative value-added in the ith industry and the change in the consumer price, X_{i2} consists of a single vector of time series observations on the logarithm of unemployment, and Z_i contains the degree of concentration and the degree of unionization in the ith industry.

For simplicity, we assume that $\boldsymbol{u}_i$ and $\boldsymbol{\eta}_{2i}$ are uncorrelated with each other and have zero means. The variance–covariance matrix of $\boldsymbol{u}_i$ and $\boldsymbol{\eta}_{2i}$ are given by

$$E\boldsymbol{u}_i\boldsymbol{u}_j' = \sigma_{ij} I_T \tag{13.8.3}$$

and

$$E\boldsymbol{\eta}_{2i}\boldsymbol{\eta}_{2j}' = \begin{cases} \Lambda & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases} \tag{13.8.4}$$

Let $\Sigma = (\sigma_{ij})$. We can write the variance–covariance of $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_N')'$ and $\boldsymbol{\eta}_2 = (\boldsymbol{\eta}_{21}', \ldots, \boldsymbol{\eta}_{2N}')'$ as

$$E\boldsymbol{u}\boldsymbol{u}' = \Sigma \otimes I_T \tag{13.8.5}$$

and

$$E\boldsymbol{\eta}_2\boldsymbol{\eta}_2' = \begin{bmatrix} \Lambda & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \Lambda \end{bmatrix} = \tilde{\Lambda}. \tag{13.8.6}$$

Combining (13.8.1) and (13.8.2), we have

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + W\boldsymbol{\gamma} + \tilde{X}_2\boldsymbol{\eta}_2 + \boldsymbol{u}, \tag{13.8.7}$$

where

$$\underset{NT\times 1}{\boldsymbol{y}} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N')',$$

$$\underset{NT\times K_1}{X_1} = (X_{11}', \ldots, X_{N1}')',$$

$$\underset{NT\times M}{W} = (Z_1'X_{12}', Z_2'X_{22}', \ldots, Z_N'X_{N2}')',$$

$$\underset{NT\times NK_2}{\tilde{X}_2} = \begin{bmatrix} X_{12} & & & \mathbf{0} \\ & X_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & X_{N2} \end{bmatrix},$$

and

$$\underset{NK_2\times 1}{\boldsymbol{\eta}_2} = (\boldsymbol{\eta}_{21}', \ldots, \boldsymbol{\eta}_{2N}')'.$$

The BLUE of $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}$ of (13.8.7) is the GLS estimator.

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\gamma}}\end{bmatrix}_{GLS} &= \left\{\begin{bmatrix}X_1'\\ W'\end{bmatrix}[\Sigma\otimes I_T+\tilde{X}_2\tilde{\Lambda}\tilde{X}_2']^{-1}(X_1,W)\right\}^{-1}\\ &\quad\left\{\begin{bmatrix}X_1'\\ W'\end{bmatrix}[\Sigma\otimes I_T+\tilde{X}_2\tilde{\Lambda}\tilde{X}_2']^{-1}\boldsymbol{y}\right\}.\end{aligned} \tag{13.8.8}$$

If Σ is diagonal, the variance–covariance matrix of the stochastic term of (13.8.7) is block-diagonal, with the ith diagonal block equal to

$$\Omega_i = X_{i2}\Lambda X_{i2}' + \sigma_{ii}I_T. \tag{13.8.9}$$

The GLS estimator (13.8.8) can be simplified as

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\gamma}}\end{bmatrix}_{GLS} &= \left[\sum_{i=1}^{N}\begin{bmatrix}X_{i1}'\\ Z_i'X_{i2}'\end{bmatrix}\Omega_i^{-1}(X_{i1},X_{2i}Z_i)\right]^{-1}\\ &\quad\left[\sum_{i=1}^{N}\begin{bmatrix}X_{i1}'\\ Z_i'X_{i2}'\end{bmatrix}\Omega_i^{-1}\boldsymbol{y}_i\right].\end{aligned} \tag{13.8.10}$$

Amemiya (1978) suggested estimating Λ and σ_{ij} as follows. Let

$$\begin{aligned}\begin{bmatrix}\boldsymbol{y}_1\\ \vdots\\ \boldsymbol{y}_N\end{bmatrix} &= \begin{bmatrix}X_{11}\\ \vdots\\ X_{N1}\end{bmatrix}\boldsymbol{\beta}_1 + \begin{bmatrix}X_{12}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}_{21} + \begin{bmatrix}\mathbf{0}\\ X_{22}\\ \vdots\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}_{22}\\ &\quad+\cdots+\begin{bmatrix}\mathbf{0}\\ \vdots\\ X_{N2}\end{bmatrix}\boldsymbol{\beta}_{2N} + \begin{bmatrix}\boldsymbol{u}_1\\ \vdots\\ \boldsymbol{u}_N\end{bmatrix}.\end{aligned} \tag{13.8.11}$$

Apply the least squares method to (13.8.11). Denote the resulting estimates by $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_{2i}$, $i=1,\ldots,N$. Then σ_{ij} can be estimated by

$$\hat{\sigma}_{ij} = \frac{1}{T}(\boldsymbol{y}_i - X_{i1}\hat{\boldsymbol{\beta}}_1 - X_{i2}\hat{\boldsymbol{\beta}}_{2i})'(\boldsymbol{y}_j - X_{j1}\hat{\boldsymbol{\beta}}_1 - X_{j2}\hat{\boldsymbol{\beta}}_{2j}), \tag{13.8.12}$$

and $\boldsymbol{\gamma}$ can be estimated by

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{N}Z_i'Z_i\right)^{-1}\left(\sum_{i=1}^{N}Z_i'\hat{\boldsymbol{\beta}}_{2i}\right). \tag{13.8.13}$$

We then estimate Λ by

$$\hat{\Lambda} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{\beta}}_{2i} - Z_i\hat{\boldsymbol{\gamma}})(\hat{\boldsymbol{\beta}}_{2i} - Z_i\hat{\boldsymbol{\gamma}})'. \tag{13.8.14}$$

Once consistent estimates of σ_{ij} and Λ are obtained (as both N and T approach infinity), we can substitute them into (13.8.8). The resulting two-stage Aitken estimator of $(\boldsymbol{\beta}_1', \boldsymbol{\gamma}')$ is consistent and asymptotically normally distributed under general conditions. A test of the hypothesis that $\boldsymbol{\gamma} = 0$ can be performed in the usual regression framework using $\hat{\boldsymbol{\gamma}}_{GLS}'\,\mathrm{Var}\,(\hat{\boldsymbol{\gamma}}_{GLS})^{-1}\hat{\boldsymbol{\gamma}}_{GLS}$, where

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}_{GLS}) = [W'\tilde{\Omega}^{-1}W - W'\tilde{\Omega}^{-1}X_1(X_1'\tilde{\Omega}^{-1}X_1)^{-1}X_1'\tilde{\Omega}^{-1}W]^{-1}, \tag{13.8.15}$$

and

$$\tilde{\Omega} = \tilde{X}_2 \tilde{\Lambda} \tilde{X}_2' + \Sigma \otimes I_T.$$

13.9 CORRELATED RANDOM-COEFFICIENTS MODELS

13.9.1 Introduction

Standard random-coefficients models assume the variation of coefficients are independent of the variation of regressors (e.g., Section 13.2.2; Hsiao and Pesaran 2008). In recent years, a great deal of attention has been devoted to the correlated random-coefficients model (e.g., Card 1996; Heckman and Vytlacil 1998; Heckman, Urzua, and Vytlacil 2006; Heckman, Schmierer, and Urzua 2010). This type of model is motivated by the measurement of treatment effect of a policy. For instance, in the study of return to schooling, it is plausible that there are unmeasured ability or motivation factors that affect the return to schooling and are also correlated with the level of schooling (e.g., Card 1995; Heckman and Vytlacil 1998; Rosen 1977). As a matter of fact, Li and Tobias (2011) find strong evidence that the amount of schooling attained is determined, in part, by the individual's own return to education. Specifically, a one-percentage-point increase in return to schooling is associated with roughly 0.2 more years of education.

A common formulation for a correlated random-coefficients model is to let

$$\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i. \tag{13.9.1}$$

Substituting (13.9.1) into the regression model (13.1.2) yields

$$y_{it} = \boldsymbol{x}_{it}' \bar{\boldsymbol{\beta}} + \boldsymbol{x}_{it}' \boldsymbol{\alpha}_i + u_{it}, \tag{13.9.2}$$

where

$$E\boldsymbol{\alpha}_i = \boldsymbol{0}, \tag{13.9.3}$$

$$E\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j' = \begin{cases} \Delta, & \text{if} \quad i = j. \\ \boldsymbol{0}, & \text{if} \quad i \neq j. \end{cases} \tag{13.9.4}$$

and

$$E(u_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{\beta}_i) = 0, \tag{13.9.5}$$

However, we now assume

$$E\boldsymbol{x}_{it} \boldsymbol{\alpha}_i' \neq \boldsymbol{0}. \tag{13.9.6}$$

Let $v_{it} = \boldsymbol{x}_{it}' \boldsymbol{\alpha}_i + u_{it}$, then

$$E(v_{it} \mid \boldsymbol{x}_{it}) \neq 0. \tag{13.9.7}$$

13.9.2 Identification with Cross-Sectional Data

If only cross-sectional observations of $(y, \boldsymbol{x})$ are available, it is not possible to identify $\bar{\boldsymbol{\beta}}$. Nor does the existence of instruments z_1 such that

$$\text{cov}\ (z_1, \boldsymbol{x}) \neq \boldsymbol{0}, \tag{13.9.8}$$

$$\text{cov}\ (z_1, u) = \boldsymbol{0} \tag{13.9.9}$$

alone is sufficient to identify $\bar{\boldsymbol{\beta}}$ because the instrumental variable estimator

$$\begin{aligned}
\hat{\bar{\boldsymbol{\beta}}}_{iv} &= \left[\left(\sum_{i=1}^{N} \boldsymbol{x}_i z_{1i}'\right)\left(\sum_{i=1}^{N} z_{1i} z_{1i}'\right)^{-1}\left(\sum_{i=1}^{N} z_{1i} \boldsymbol{x}_i'\right)\right]^{-1} \\
&\quad \left[\left(\sum_{i=1}^{N} \boldsymbol{x}_i z_{1i}'\right)\left(\sum_{i=1}^{N} z_{1i} z_{1i}'\right)^{-1}\left(\sum_{i=1}^{N} z_{1i} y_i\right)\right] \\
&= \bar{\boldsymbol{\beta}} + \left[\left(\sum_{i=1}^{N} \boldsymbol{x}_i z_{1i}'\right)\left(\sum_{i=1}^{N} z_{1i} z_{1i}'\right)^{-1}\left(\sum_{i=1}^{N} z_{1i} \boldsymbol{x}_i'\right)\right]^{-1} \\
&\quad \left[\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i z_{1i}'\right)\left(\frac{1}{N}\sum_{i=1}^{N} z_{1i} z_{1i}'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} z_{1i} \boldsymbol{x}_i' \boldsymbol{\alpha}_i + \frac{1}{N}\sum_{i=1}^{N} z_{1i} u_i\right)\right].
\end{aligned} \tag{13.9.10}$$

Although under (13.9.9), plim $\frac{1}{N}\sum_{i=1}^{n} z_{1i} u_i = \mathbf{0}$,

$$\begin{aligned}
\text{plim}\, \frac{1}{N}\sum_{i=1}^{n} z_{1i} \boldsymbol{x}_i' \boldsymbol{\alpha}_i &= E[z_1 E(\boldsymbol{x}'\boldsymbol{\alpha} \mid z_1)] \\
&= E[z_1 E(\boldsymbol{x}' \mid z_1) E(\boldsymbol{\alpha} \mid \boldsymbol{x}, z_1)],
\end{aligned} \tag{13.9.11}$$

which is not equal to zero given (13.9.8) and the assumption that $E(\boldsymbol{\alpha} \mid \boldsymbol{x}) \neq \mathbf{0}$.

To identify $\bar{\boldsymbol{\beta}}$, the variation of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_i$ need to be independent conditional on z_{1i}. In other words, we need exclusion restrictions. Heckman and Vytlacil (1998) consider estimating $\bar{\boldsymbol{\beta}}$ assuming the existence of instruments $z_i = (z_{1i}, z_{2i})$ such that

$$\boldsymbol{x}_i = \Pi z_{1i} + \boldsymbol{v}_i, \tag{13.9.12}$$

$$\boldsymbol{\beta}_i = \Phi z_{2i} + \boldsymbol{\eta}_i, \tag{13.9.13}$$

where z_{1i} and z_{2i} are $m_1 \times 1$ and $m_2 \times 1$ vectors of instruments that satisfy

$$E(u_i \mid z_i) = 0, \tag{13.9.14}$$

$$E(\boldsymbol{\eta}_i \mid z_i) = \mathbf{0}, \tag{13.9.15}$$

$$E(\boldsymbol{v}_i \mid z_i) = \mathbf{0}, \tag{13.9.16}$$

and z_2 contains elements that are not in z_1.
Then, under (13.9.13), if Φ is known, an estimator for $\bar{\boldsymbol{\beta}}$ can be obtained from the relation,

$$E(\boldsymbol{\beta}_i) = \bar{\boldsymbol{\beta}} = \Phi E(z_2). \tag{13.9.17}$$

Substituting (13.9.12), (13.9.13) into (13.2.1) yields

$$\begin{aligned}
y_i &= \boldsymbol{x}_i' \boldsymbol{\beta}_i + u_i \\
&= (z_{1i}' \Pi' + \boldsymbol{v}_i')(\Phi z_{2i} + \boldsymbol{\eta}_i) + u_i \\
&= (z_{1i}' \boldsymbol{\pi}_1 z_{2i}') \boldsymbol{\phi}_1 \;+\; (z_{1i}' \boldsymbol{\pi}_2 z_{2i}') \boldsymbol{\phi}_2 \\
&\quad + \cdots + (z_{1i}' \boldsymbol{\pi}_K z_{2i}') \boldsymbol{\phi}_K \;+\; E(\boldsymbol{v}_i' \boldsymbol{\eta}_i \mid z_i) + \epsilon_i^*,
\end{aligned} \tag{13.9.18}$$

where $\boldsymbol{\pi}_k'$ and $\boldsymbol{\phi}_k'$ denote the kth row of Π and Φ, respectively,

$$\epsilon_i^* = \boldsymbol{v}_i'\Phi z_{2i} \;+\; \boldsymbol{\eta}_i'\Pi z_{1i} + [\boldsymbol{v}_i'\boldsymbol{\eta}_i - E(\boldsymbol{v}_i'\boldsymbol{\eta}_i \mid z_i)] + u_i. \tag{13.9.19}$$

Under (13.9.12)–(13.9.16), $E(\epsilon_i^* \mid z_i) = 0$.

Therefore, a consistent estimator of Φ exists if

$$E(\boldsymbol{v}_i'\boldsymbol{\eta}_i \mid z_{1i}, z_{2i}) = \Sigma_{v\eta} \tag{13.9.20}$$

is not a function of z_{1i} and z_{2i}, and

$$\text{rank}\left[\frac{1}{N}\sum_{i=1}^{n}\left(z_{2i}z_{1i}' \otimes \hat{\Pi}\right)\left(\hat{\Pi}' \otimes z_{1i}z_{2i}'\right)\right] = Km_2. \tag{13.9.21}$$

In other words, the necessary condition for the identification of $E(\boldsymbol{\beta}_i) = \bar{\boldsymbol{\beta}}$ for the correlated random-coefficients model (13.9.1)–(13.9.7) when only cross-sectional data are available is there exist m_1 instruments for $\boldsymbol{x}_i$ and nonzero m_2 instruments for $\boldsymbol{\beta}_i$ that satisfy (13.9.12)–(13.9.16) with $m_1^2 > Km_2, m_2 > 0$, and either (13.9.20) holds or $E(\boldsymbol{v}_i'\boldsymbol{\eta}_i \mid z_i)$ is known.

The requirements that there exist nonzero z_1 and z_2 with $m_1^2 \geq Km_2$, and (13.9.20) holds are stronger than the usual requirement for the existence of an instrumental variable estimator. In fact, the necessary condition requires the existence of both z_1 and z_2 (i.e., $m_1 > 0,\ m_2 > 0$). Neither is (13.9.20) an innocuous assumption. The conditional covariance independent of z holds only if $\boldsymbol{v}, \boldsymbol{\eta}$, and z are joint normal.

13.9.3 Estimation of the Mean Effects with Panel Data

When only cross-sectional data are available, the identification conditions of average effects for a correlated random-coefficients model are very stringent and may not be satisfied for many data sets. The instrumental variable approach requires the estimation of a large number of parameters $[(m_1 + m_2)K]$. Multicollinearity and shortages of degrees of freedom could lead to very unreliable estimates. On the other hand, panel data, by blending inter-individual differences with intra-individual dynamics, can offer several alternatives to get around the difficulties of the correlations between the coefficients and the regressors without the prior conjecture of the existence of certain instruments that satisfy the exclusion restrictions. For ease of exposition, we suppose there are T time series observations of $(y_{it}, \boldsymbol{x}_{it})$ for each individual i. Let $(\boldsymbol{y}_i', \boldsymbol{x}_i')$ be the stacked T time series observations of y_{it} and $\boldsymbol{x}_{it}'$ for each i.

13.9.3.1 Group Mean Estimator

We note that conditional on $\boldsymbol{x}_i, \boldsymbol{\beta}_i$ is a fixed constant. Under (13.9.5), the least squares estimator of the equation

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_i + u_{it}, \quad t = 1, \ldots, T, \tag{13.9.22}$$

yields an unbiased estimator of $\boldsymbol{\beta}_i, \hat{\boldsymbol{\beta}}_i$, for each i with covariance matrix $\sigma_i^2(X_i'X_i)^{-1}$ if u_{it} is independently distributed over t, where X_i denotes the $T \times K$ stacked $(\boldsymbol{x}_{it}')$. If u_{it} is independently distributed over i, taking the simple average of $\hat{\boldsymbol{\beta}}_i$ as in Hsiao, Pesaran, and Tahmiscioglu (1999),

$$\hat{\bar{\boldsymbol{\beta}}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i. \tag{13.9.23}$$

yields a consistent estimator of $\bar{\boldsymbol{\beta}}$ as $N \to \infty$. If $T > K$, the estimator (13.9.23) is consistent and asymptotically normally distributed as $N \to \infty$, and $\sqrt{N}(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\text{Asy Cov } (\hat{\bar{\boldsymbol{\beta}}}) = \left[\Delta + \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2(X_i'X_i)^{-1}\right], \tag{13.9.24}$$

if u_{it} is independently distributed over i and t with variance σ_i^2.

13.9.3.2 Conventional Fixed-Effects Estimator

The estimator (13.9.23) is simple to implement. However, if $T < K$, we cannot estimate $\boldsymbol{\beta}_i$ using the ith individual's time series observations $(y_i, \boldsymbol{x}_i')$. Nevertheless, the conventional fixed-effects estimator can still allow us to obtain consistent estimator of $\bar{\boldsymbol{\beta}}$ in a number of situations.

Let $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\bar{\boldsymbol{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_{it}$. The conventional fixed-effects estimator first takes the deviation of each observation from its time series mean, then regresses $(y_{it} - \bar{y}_i)$ on $(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)$ (e.g., Chapter 2). Model (13.9.2) leads to

$$(y_{it} - \bar{y}_i) = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\bar{\boldsymbol{\beta}} + (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\alpha}_i + (u_{it} - \bar{u}_i), \quad i = 1, \ldots, N, \; t = 1, \ldots, T, \tag{13.9.25}$$

where $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$.

The FE estimator (13.9.25) will converge to $\bar{\boldsymbol{\beta}}$ provided

$$\text{plim } \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\alpha}_i = \mathbf{0} \tag{13.9.26}$$

and

$$\text{plim } \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)(u_{it} - \bar{u}_i) = \mathbf{0} \tag{13.9.27}$$

Under (13.9.5), (13.9.27) holds. Under the assumption that $\boldsymbol{x}_{it}$ and $\boldsymbol{\alpha}_i$ are linearly related with finite variance, $\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$ does not involve $\boldsymbol{\alpha}_i$; hence, (13.9.26) holds by a law of large numbers. Thus, the conventional fixed-effects estimator is $\sqrt{N}$ consistent and asymptotically normally distributed as $N \to \infty$. The asymptotic covariance matrix of the conventional fixed-effects estimator (2.2.10) can be approximated using the Newey–West heteroscedasticity-autocorrelation consistent formula (Vogelsang 2012).

When $(\boldsymbol{x}_{it}, \boldsymbol{\alpha}_i)$ jointly have elliptical distribution[22] (e.g., Fang and Zhang 1990; Gupta and Varga 1993), $\boldsymbol{x}_{it}$ and $\boldsymbol{\alpha}_i$ are linearly related. Another case that the fixed-effects

[22] Many commonly assumed distributions such as uniform, normal, Student's t, double exponential, etc., belong to the family of elliptical distributions.

estimator can be consistent is that $(\boldsymbol{x}_{it}, \boldsymbol{\alpha}_i)$ are jointly symmetrically distributed, then $\frac{1}{NT}\sum_{i=1}^{N}(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)'\boldsymbol{\alpha}_i$ will converge to zero even though $\boldsymbol{x}_{it}$ have a mean different from zero.

13.9.3.3 Panel Pooled Least Squares Estimator

The conventional fixed-effects (FE) estimator (2.2.10) can yield a consistent estimator of $\bar{\boldsymbol{\beta}}$, under certain conditions. However, if $\boldsymbol{\alpha}$ and $\boldsymbol{x}_{it}$ are not linearly related, it is inconsistent. Moreover, if $\boldsymbol{x}_{it}$ contains time-invariant variables, then the mean effects of time-invariant variables cannot be identified by the FE estimator. Furthermore, the FE estimator only makes use of within (-group) variation. Since in general, the between-group variation is much larger than within-group variation, the FE estimator could also mean a loss of efficiency. To get around these limitations on the FE estimator as well as to allow the case that $\boldsymbol{\alpha}_i$ and $\boldsymbol{x}_{it}$ are not linearly related, Hsiao, Li, Liang, and Xie (2019) suggest a modified specification to obtain the estimate of the mean effects.

To illustrate the basic idea, they first assume that $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$ is a linear function of $\boldsymbol{x}_i$. They then show that similar procedures can be applied if $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$ is a function of higher order of $\boldsymbol{x}_i$.

Suppose $(\boldsymbol{x}_i', \boldsymbol{\alpha}_i')$ are independently identically distributed across i with

$$E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i) = \boldsymbol{a} + B\boldsymbol{x}_i, \tag{13.9.28}$$

where $\boldsymbol{a}$ and B are $K \times 1$ and $K \times TK$ constant vector and matrix, respectively. From (13.9.3) and (13.9.28), we have

$$E_x[E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)] = \boldsymbol{a} + BE(\boldsymbol{x}_i) = \boldsymbol{0}. \tag{13.9.29}$$

It follows that

$$E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i) = B(\boldsymbol{x}_i - E\boldsymbol{x}_i) \tag{13.9.30}$$

Substituting

$$\boldsymbol{\alpha}_i = E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i) + \boldsymbol{\omega}_i \tag{13.9.31}$$

and (13.9.30) into (13.9.2) yields

$$y_{it} = \boldsymbol{x}_{it}'\bar{\boldsymbol{\beta}} + \boldsymbol{x}_{it}'B(\boldsymbol{x}_i - E\boldsymbol{x}_i) + v_{it}^*, \tag{13.9.32}$$

where

$$v_{it}^* = \boldsymbol{x}_{it}'\boldsymbol{\omega}_i + u_{it}. \tag{13.9.33}$$

By construction, $E(v_{it}^* \mid \boldsymbol{x}_i) = 0$. Therefore, the least squares regression of

$$\boldsymbol{y}_i = X_i\bar{\boldsymbol{\beta}} + X_i \otimes (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \operatorname{vec}(B') + \boldsymbol{v}_i^* \tag{13.9.34}$$

yields $\sqrt{N}$ consistent and asymptotically normally distributed estimator of $\bar{\boldsymbol{\beta}}$ when $N \to \infty$, where $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$, and $\boldsymbol{v}_i^* = (v_{i1}^*, \ldots, v_{iT}^*)'$ as long as

$$\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} X_i'X_i & X_i'X_i \otimes (\boldsymbol{x}_i' - \bar{\boldsymbol{x}}') \\ (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \otimes X_i'X_i & (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \otimes X_i'X_i \otimes (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \end{bmatrix} \tag{13.9.35}$$

is a full rank matrix.

However, since

$$E\boldsymbol{v}_i^*\boldsymbol{v}_i^{*\prime} = X_i\Delta^*X_i' + \sigma_i^2 I_T, \tag{13.9.36}$$

where $\Delta^* = E(\boldsymbol{\omega}_i \boldsymbol{\omega}_i')$, a more efficient estimator of $\bar{\boldsymbol{\beta}}$ will be a generalized least squares estimator (GLS) if Δ^* and σ_i^2 are known. If Δ^* and σ_i^2 are unknown, we can apply the feasible GLS (FGLS) through a two-step procedure.

Similar reasoning can be applied if $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$ is a higher-order polynomial of $\boldsymbol{x}_i$, say

$$E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i) = \boldsymbol{a} + B\boldsymbol{x}_i + C\boldsymbol{x}_i \otimes \boldsymbol{x}_i. \tag{13.9.37}$$

then from $E_x[E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)] = 0$, it follows that then the least squares regression of

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}'\bar{\boldsymbol{\beta}} + \boldsymbol{x}_{it}' \otimes (\boldsymbol{x}_i - E\boldsymbol{x}_i)' \text{ vec } (B') \\ &\quad + \boldsymbol{x}_{it}' \otimes [(\boldsymbol{x}_i \otimes \boldsymbol{x}_i) - E(\boldsymbol{x}_i \otimes \boldsymbol{x}_i)]' \text{ vec } (C') \\ &\quad + v_{it}, \end{aligned} \tag{13.9.38}$$

is consistent when $N \to \infty$, where $v_{it} = \boldsymbol{x}_{it}'\boldsymbol{\omega}_i + u_{it}$ and $\boldsymbol{\omega}_i = \boldsymbol{\alpha}_i - E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$.

LS (or FGLS) regression of (13.9.38) not only requires the estimation of a large number of parameters, but it could also raise the issue of multicollinearity. However, if $\boldsymbol{x}_{it}$ is stationary, a more parsimonious approximation would be to follow Mundlak (1978a) to replace (13.9.37) by

$$E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i) = \boldsymbol{a} + \tilde{B}\bar{\boldsymbol{x}}_i + \tilde{C}(\bar{\boldsymbol{x}}_i \otimes \bar{\boldsymbol{x}}_i), \tag{13.9.39}$$

where $\bar{\boldsymbol{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_{it}$ and regressing

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}'\bar{\boldsymbol{\beta}} + \boldsymbol{x}_{it}' \otimes (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})' \text{ vec } (\tilde{B}') \\ &\quad + \boldsymbol{x}_{it}' \otimes [(\bar{\boldsymbol{x}}_i \otimes \bar{\boldsymbol{x}}_i) - (\bar{\boldsymbol{x}} \otimes \bar{\boldsymbol{x}})]' \text{ vec } (\tilde{C}') \\ &\quad + v_{it}, \end{aligned} \tag{13.9.40}$$

where $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$ and $(\bar{\boldsymbol{x}} \otimes \bar{\boldsymbol{x}}) = \frac{1}{N}\sum_{i=1}^{N}[\bar{\boldsymbol{x}}_i \otimes \bar{\boldsymbol{x}}_i]$.

Remark 13.9.1 When $\boldsymbol{x}_{it}$ contains an intercept term, (13.9.35) is not a full rank matrix. Let $\boldsymbol{x}_{it}' = (1, \tilde{\boldsymbol{x}}_{it})$, where $\tilde{\boldsymbol{x}}_{it}$ denotes the $(1 \times (K-1))$ time-varying explanatory variables. Let $\boldsymbol{\alpha}_i' = (\alpha_{1i}, \tilde{\boldsymbol{\alpha}}_i')$ and $\bar{\boldsymbol{\beta}}' = (\bar{\boldsymbol{\beta}}_1, \bar{\tilde{\boldsymbol{\beta}}})$ be the corresponding partitions. Rewrite (13.9.28) in the form

$$E\begin{pmatrix} \alpha_{1i} & \mid & \\ & \mid & \boldsymbol{x}_{it} \\ \tilde{\boldsymbol{\alpha}}_i & \mid & \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{b}}_1' \\ \tilde{B} \end{pmatrix} (\tilde{\boldsymbol{x}}_{it} - E\tilde{\boldsymbol{x}}_{it}). \tag{13.9.41}$$

Then

$$\begin{aligned} E(\boldsymbol{y}_i \mid X_i) &= X_i\bar{\boldsymbol{\beta}} + \boldsymbol{e}(\tilde{\boldsymbol{x}}_i - E\bar{\tilde{\boldsymbol{x}}}_i)'\tilde{\boldsymbol{b}}_1^* \\ &\qquad + \tilde{X}_i \otimes [\tilde{X}_i - E(\tilde{X}_i)]' \text{ vec } (\hat{B}'), \end{aligned} \tag{13.9.42}$$

where $\tilde{\boldsymbol{b}}_1^* = T\tilde{\boldsymbol{b}}_1$, $\tilde{X}_i$ is the $T \times (K-1)$ stacked $\tilde{\boldsymbol{x}}_{it}'$, and $\boldsymbol{e}$ is a $(T \times 1)$ vector of 1's.

Therefore, a consistent estimator of $\bar{\boldsymbol{\beta}}' = (\bar{\boldsymbol{\beta}}_1, \bar{\tilde{\boldsymbol{\beta}}})$ can be obtained by the least squares regression of

$$y_{it} = \bar{\beta}_1 + \tilde{\boldsymbol{x}}_{it}'\bar{\tilde{\boldsymbol{\beta}}} + (\bar{\tilde{\boldsymbol{x}}}_i - \bar{\tilde{\boldsymbol{x}}})'\tilde{\boldsymbol{b}}_1^{\,*} + (\tilde{\boldsymbol{x}}_{it}' \otimes (\tilde{\boldsymbol{x}}_{it} - \bar{\tilde{\boldsymbol{x}}})') \text{ vec } (\tilde{B}') + v_{it}, \quad i = 1, \ldots, N, \; t = 1, \ldots, T. \tag{13.9.43}$$

13.9.3.4 Semiparametric Estimates

The application of a panel least squares estimator requires precise knowledge of $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$. If the pattern of correlation is unknown, Hsiao, Li, Liang, and Xie (2019) suggest a semiparametric estimator if there exists a q dimensional random variables $z_i (q \leq K)$ such that conditional on $z_i, \boldsymbol{\beta}_i$, and $\boldsymbol{x}_{it}$ are conditionally independent, $(\boldsymbol{\beta}_i \perp \boldsymbol{x}_{it} \mid z_i)$, and $E[\boldsymbol{x}_{it}\boldsymbol{x}'_{it} \mid z_i]$ is a full rank matrix. Then $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i, z_i) = E(\boldsymbol{\alpha}_i \mid z_i) \equiv \boldsymbol{g}(z_i)$. The random variable z_i can contain components of $\boldsymbol{x}_i$; it could be a function of $\boldsymbol{x}_i$, say, the propensity score (Rosenbaum and Rubin 1983); or instruments for $\boldsymbol{x}_{it}$ and $\boldsymbol{\beta}_i$ (e.g., Heckman and Vytlacil 2001, 2005, 2007; Heckman, Schmierer, and Urzua 2010); or simply the time-series mean of the ith individual's $\boldsymbol{x}_{it}, \bar{\boldsymbol{x}}_i$ (Mundlak 1978a).

$$
\begin{aligned}
y_{it} &= \boldsymbol{x}'_{it}\bar{\boldsymbol{\beta}} + \boldsymbol{x}'_{it}\boldsymbol{g}(z_i) + \epsilon_{it} \\
&= \boldsymbol{x}'_{it}\boldsymbol{\theta}(z_i) + \epsilon_{it},
\end{aligned}
\tag{13.9.44}
$$

where $\boldsymbol{\theta}(z_i) = \bar{\boldsymbol{\beta}} + \boldsymbol{g}(z_i)$ and $E\boldsymbol{g}(z_i) = \mathbf{0}$. Given $\boldsymbol{\theta}(z_i)$ and $E(\boldsymbol{\theta}(z_i)) = \bar{\boldsymbol{\beta}}$, a consistent estimator of $\bar{\boldsymbol{\beta}}$ is

$$
\hat{\bar{\boldsymbol{\beta}}}_{semi} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\theta}(z_i). \tag{13.9.45}
$$

Hsiao, Li, Liang, and Xie (2019) suggest two types of semiparametric estimators for $\boldsymbol{\theta}(z)$: local constant and local polynomial estimation methods. The local constant estimator of $\boldsymbol{\theta}(z)$ is given by

$$
\hat{\boldsymbol{\theta}}_{LC}(z) = \left(\sum_{j=1}^{N}\sum_{s=1}^{T} x_{js}x'_{js}K_{h,z_j z}\right)^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T} x_{js}y_{js}K_{h,z_j z}, \tag{13.9.46}
$$

where $K_{h,z_j z} = \Pi_{l=1}^{q} k\left(\frac{z_{jl}-z_l}{h_l}\right)$ is the product kernel, $k(\cdot)$ is the univariate kernel function, and z_{jl} and z_l are the lth component of z_j and z, respectively.

The local polynomial estimation minimizes the kernel weighted sum of squared errors

$$
\sum_{j=1}^{N}\sum_{s=1}^{T}\left[y_{js} - \sum_{0 \leq |k| \leq p} x'_{js}b_k(z)(z_j - z)^k\right]^2 K_{h,z_j z}, \tag{13.9.47}
$$

with respect to each $b_k(z)$ which gives an estimate of $\hat{b}_k(z)$. Thus, $\hat{\boldsymbol{\theta}}_{LP} = \hat{\boldsymbol{b}}_0(z)$ is the pth order local polynomial estimator of $\boldsymbol{\theta}(z)$. The local polynomial estimator of β is then given by

$$
\hat{\boldsymbol{\beta}}_{LP} = \frac{1}{N_i}\sum_{i=1}^{N}\hat{\boldsymbol{\theta}}_{LP}(z_i) \tag{13.9.48}
$$

Simple Monte Carlo studies conducted by Hsiao et al. (2019) show that if the exact order of $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$ is known, the panel pooled least squares estimator performs well. If $E(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i)$ is unknown, the group mean estimator (13.9.23) and semiparametric estimator (13.9.45) are robust to a variety of joint distribution of $(\boldsymbol{\alpha}_i, \boldsymbol{x}_{it})$, but not the conventional fixed effects estimator.

APPENDIX 13A COMBINATION OF TWO NORMAL DISTRIBUTIONS

Suppose that conditional on $X, \boldsymbol{\beta}, \boldsymbol{y} \sim N(X\boldsymbol{\beta}, \Omega)$ and $\boldsymbol{\beta} \sim N(A\bar{\boldsymbol{\beta}}, C)$. Then the posterior of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}$ given $\boldsymbol{y}$ is

$$P(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \boldsymbol{y}) \propto \exp\left[-\frac{1}{2}\{(\boldsymbol{y} - X\boldsymbol{\beta})'\Omega^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) + (\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}})'C^{-1}(\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}})\}\right], \tag{13A.1}$$

where "$\propto$" denotes "proportionality." Using the identity (e.g., Rao 1971, p. 33)

$$(D + BFB')^{-1} = D^{-1} - D^{-1}B(B'D^{-1}B + F^{-1})^{-1}B'D^{-1}. \tag{13A.2}$$

and

$$(D + F)^{-1} = D^{-1} - D^{-1}(D^{-1} + F^{-1})^{-1}D^{-1}, \tag{13A.3}$$

we can complete the squares of

$$\begin{aligned}(\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}})'C^{-1}(\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}}) + (\boldsymbol{y} - X\boldsymbol{\beta})'\Omega^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) \\ = \boldsymbol{\beta}'C^{-1}\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}'A'C^{-1}A\bar{\boldsymbol{\beta}} - 2\boldsymbol{\beta}'C^{-1}A\bar{\boldsymbol{\beta}} \\ + \boldsymbol{y}'\Omega^{-1}\boldsymbol{y} + \boldsymbol{\beta}'X'\Omega^{-1}X\boldsymbol{\beta} - 2\boldsymbol{\beta}'X'\Omega^{-1}\boldsymbol{y}.\end{aligned} \tag{13A.4}$$

Let

$$\begin{aligned}Q_1 = [\boldsymbol{\beta} - (X'\Omega^{-1}X + C^{-1})^{-1}(X\Omega^{-1}\boldsymbol{y} + C^{-1}A\bar{\boldsymbol{\beta}})]'(C^{-1} + X'\Omega^{-1}X) \\ \cdot [\boldsymbol{\beta} - (X'\Omega^{-1}X + C^{-1})^{-1}(X'\Omega^{-1}\boldsymbol{y} + C^{-1}A\bar{\boldsymbol{\beta}})],\end{aligned} \tag{13A.5}$$

then

$$\begin{aligned}\boldsymbol{\beta}'C^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}'X'\Omega^{-1}X\boldsymbol{\beta} - 2\boldsymbol{\beta}'C^{-1}A\bar{\boldsymbol{\beta}} - 2\boldsymbol{\beta}'X'\Omega^{-1}\boldsymbol{y} \\ = Q_1 - (X'\Omega^{-1}\boldsymbol{y} + C^{-1}A\bar{\boldsymbol{\beta}})'(X'\Omega^{-1}X + C^{-1})^{-1}(X'\Omega^{-1}\boldsymbol{y} + C^{-1}A\bar{\boldsymbol{\beta}}).\end{aligned} \tag{13A.6}$$

Substituting (13A.6) into (13A.4) yields

$$\begin{aligned}Q_1 &+ \boldsymbol{y}'[\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X + C^{-1})^{-1}X'\Omega^{-1}]\boldsymbol{y} \\ &+ \bar{\boldsymbol{\beta}}'A'[C^{-1} - C^{-1}(X'\Omega^{-1}X + C^{-1})^{-1}C^{-1}]A\bar{\boldsymbol{\beta}} \\ &- 2\bar{\boldsymbol{\beta}}'A'C^{-1}(X'\Omega^{-1}X + C^{-1})^{-1}X'\Omega^{-1}\boldsymbol{y} \\ &= Q_1 + \boldsymbol{y}'(XCX' + \Omega)^{-1}\boldsymbol{y} + \bar{\boldsymbol{\beta}}'A'X'(XCX' + \Omega)^{-1}XA\bar{\boldsymbol{\beta}} \\ &- 2\bar{\boldsymbol{\beta}}'A'X'(XCX' + \Omega)^{-1}\boldsymbol{y} \\ &= Q_1 + Q_2 + Q_3\end{aligned} \tag{13A.7}$$

where

$$\begin{aligned}Q_2 = \{\bar{\boldsymbol{\beta}} - [A'X'(XCX' + \Omega)^{-1}XA]^{-1}[A'X'(XCX' + \Omega)^{-1}\boldsymbol{y}]\}' \\ \cdot [A'X'(XCX' + \Omega)^{-1}XA] \\ \cdot \{\bar{\boldsymbol{\beta}} - [A'X'(XCX' + \Omega)^{-1}XA]^{-1}[A'X'(XCX' + \Omega)^{-1}\boldsymbol{y}]\},\end{aligned} \tag{13A.8}$$

$$Q_3 = \boldsymbol{y}'\{(XCX' + \Omega)^{-1} - (XCX' + \Omega)^{-1} XA[A'X(XCX' + \Omega)^{-1} XA]^{-1} A'X'(XCX' + \Omega)^{-1}\}\, \boldsymbol{y} \tag{13A.9}$$

Since Q_3 is a constant independent of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}$, we can write $P(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \boldsymbol{y})$ in the form of $P(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \boldsymbol{y}) P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{y})$, which becomes

$$P\{\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \boldsymbol{y}\} \propto \exp\left\{-\frac{1}{2} Q_1\right\} \exp\left\{-\frac{1}{2} Q_2\right\} \tag{13A.10}$$

where $\exp\left\{-\frac{1}{2}Q_1\right\}$ is proportional to $P(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \boldsymbol{y})$ and $\exp\left\{-\frac{1}{2}Q_2\right\}$ is proportional to $P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{y})$. That is, $P(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \boldsymbol{y})$ is $N\{(X'\Omega^{-1}X + C^{-1})^{-1}(X'\Omega^{-1}\boldsymbol{y} + C^{-1}A\bar{\boldsymbol{\beta}}), (C^{-1} + X'\Omega^{-1}X)^{-1}\}$ and $P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{y})$ is $N\{[A'X'(XCX' + \Omega)^{-1}XA]^{-1}[A'X'(XCX' + \Omega)^{-1}\boldsymbol{y}], [A'X'(XCX' + \Omega)^{-1}XA]^{-1}\}$.

Alternatively, we may complete the square of the left side of (13A.4) with the aim of writing $P(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \boldsymbol{y})$ in the form of $P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{\beta}, \boldsymbol{y}) P(\boldsymbol{\beta} \mid \boldsymbol{y})$,

$$\begin{aligned} & Q_4 + \boldsymbol{\beta}'[X'\Omega^{-1}X + C^{-1} - C^{-1}A(A'CA)^{-1}A'C^{-1}]\boldsymbol{\beta} \\ & \qquad - 2\boldsymbol{\beta}'X'\Omega^{-1}\boldsymbol{y} + \boldsymbol{y}'\Omega^{-1}\boldsymbol{y} \\ & \quad = Q_4 + Q_5 + Q_3, \end{aligned} \tag{13A.11}$$

where

$$Q_4 = [\bar{\boldsymbol{\beta}} - (A'C^{-1}A)^{-1}A'C^{-1}\boldsymbol{\beta}]'(A'C^{-1}A)[\bar{\boldsymbol{\beta}} - (A'C^{-1}A)^{-1}A'C^{-1}\boldsymbol{\beta}], \tag{13A.12}$$

$$Q_5 = [\boldsymbol{\beta} - D^{-1}X'\Omega^{-1}\boldsymbol{y}]'D[\boldsymbol{\beta} - D^{-1}X'\Omega^{-1}\boldsymbol{y}]. \tag{13A.13}$$

and

$$D = X'\Omega^{-1}X + C^{-1} - C^{-1}A(A'C^{-1}A)^{-1}A'C^{-1}. \tag{13A.14}$$

Therefore, $P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{\beta}, \boldsymbol{y}) \sim N\{(A'C^{-1}A)^{-1}C^{-1}\boldsymbol{\beta}, (A'C^{-1}A)^{-1}\}$ and $P(\boldsymbol{\beta} \mid \boldsymbol{y}) \sim N\{D^{-1}X'\Omega^{-1}\,\boldsymbol{y}, D^{-1}\}$.