CHAPTER 11

Incomplete Panel Data

Thus far our discussions have been concentrated on situations in which the sample of N cross-sectional units over T time periods is sufficient to identify a behavioral model. In this chapter we turn to issues of incomplete panel data. We first discuss issues when some individuals are dropped out of the experiment or survey. We note that when individuals are followed over time, there is a high probability that this may occur. Because the situations where individuals are missing for a variety of behavioral reasons have been discussed in Chapter 8, Section 8.3, in this chapter we consider only the situations where (1) individuals are missing randomly or are being rotated; (2) a series of independent cross-sections are observed over time; and (3) only a single set of cross-sectional data is available in conjunction with the aggregate time series observations. We then consider the problems of estimating dynamic models when the length of time series is shorter than the maximum order of the lagged variables included in the equation.

11.1 ROTATING OR RANDOMLY MISSING DATA

In many situations we do not have complete time series observations on cross-sectional units. Instead, individuals are selected according to a "rotating" scheme that can be briefly stated as follows: Let all individuals in the population be numbered consecutively. Suppose the sample in period 1 consists of individuals $1, 2, \ldots, N$. In period 2, individuals $1, \ldots, m_1$ $(0 \leq m_1 \leq N)$ are replaced by individuals $N+1, \ldots, N+m_1$. In period 3, individuals $m_1+1, \ldots, m_1+m_2$ $(0 \leq m_2 \leq N)$ are replaced by individuals $N+m_1+1, \ldots, N+m_1+m_2$, and so on. This procedure of dropping the first m_{t-1} individuals from the sample selected in the previous period and augmenting the sample by drawing m_{t-1} individuals from the population so that the sample size remains the same continues through all periods. Hence, for T periods, although the total number of observations remains at NT, we have observed $N + \sum_{t=1}^{T-1} m_t$ individuals.

"Rotation" of a sample of micro units over time is quite common. It can be caused by deliberate policy of the data-collecting agency (e.g., the Bureau

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

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

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

$$L = \prod_{i=1}^{N+\sum_{t=1}^{T-1} m_t} f(y_{it_i}, \ldots, y_{iT_i}), \tag{11.1.1}$$

where t_i and T_i denote the first and the last periods during which the ith individual was observed. Apart from the minor modifications of t_i for 1 and T_i for T, (11.1.1) is basically of the same form as the likelihood functions for the complete panel data.

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

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + v_{it}, \tag{11.1.2}$$

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

$$v_{it} = \alpha_i + u_{it}. \tag{11.1.3}$$

The error terms α_i and u_{it} are independent of each other and are independently distributed, with zero means and constant variances σ_α^2 and σ_u^2, respectively. For ease of exposition, we assume that α_i and u_{it} are uncorrelated with $\mathbf{x}_{it}$.[1]

[1] If α_i are correlated with $\mathbf{x}_{it}$, we can eliminate the linear dependence between α_i and $\mathbf{x}_{it}$ by assuming $\alpha_i = \Sigma_t \mathbf{a}'_t \mathbf{x}_{it} + \epsilon_i$. For details, see Chapter 3 or Mundlak (1978a).

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

We also assume that in each period a fixed number of individuals are dropped out of the sample and the same number of individuals from the population are added back to the sample (namely, $m_t = m$ for all t). Thus, the total number of individuals observed is

$$H = (T-1)m + N. \tag{11.1.4}$$

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

$$\mathbf{y}_i = X_i \boldsymbol{\beta} + \mathbf{v}_i, \tag{11.1.5}$$

where

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

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

The variance–covariance matrix of v_i is

$$V_i = \sigma_u^2 + \sigma_\alpha^2 \quad \text{if } q_i = 1 \tag{11.1.6a}$$

and is

$$V_i = E\mathbf{v}_i\mathbf{v}'_i = \sigma_u^2 I_{q_i} + \sigma_\alpha^2 J_i \quad \text{if } q_i > 1, \tag{11.1.6b}$$

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

$$V_i^{-1} = (\sigma_u^2 + \sigma_\alpha^2)^{-1}, \tag{11.1.7a}$$

and for $q_i > 1$,

$$V_i^{-1} = \frac{1}{\sigma_u^2}\left[I_{q_i} - \frac{\sigma_\alpha^2}{\sigma_u^2 + q_i\sigma_\alpha^2} J_i\right]. \tag{11.1.7b}$$

Because $\mathbf{y}_i$ and $\mathbf{y}_j$ are uncorrelated, the variance–covariance matrix of the stacked equations $(y'_1, \ldots, y'_{N+(T-1)m})'$ is block-diagonal. Therefore, the GLS estimator of β is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\sum_{i=1}^{N+(T-1)m} X'_i V_i^{-1} X_i\right]^{-1} \left[\sum_{i=1}^{N+(T-1)m} X'_i V_i^{-1} \mathbf{y}_i\right]. \tag{11.1.8}$$

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

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

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

$$\hat{\sigma}_u^2 = \frac{1}{H^*} \sum_{i \in \Theta} \sum_{t=t_i}^{Ti} [(y_{it} - \bar{y}_i) - \hat{\boldsymbol{\beta}}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)]^2, \tag{11.1.9}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{N + (T-1)m} \sum_{i=1}^{N+(T-1)m} \left[(\bar{y}_i - \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_i)^2 - \frac{1}{q_i} \hat{\sigma}_u^2 \right]. \tag{11.1.10}$$

Similarly, we can apply the MLE by maximizing the logarithm of the likelihood function (11.1.1):

$$\begin{aligned} \log L &= -\frac{NT}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{N+(T-1)m} \log \mid V_i \mid \\ &\quad - \frac{1}{2} \sum_{i=1}^{N+(T-1)m} (\mathbf{y}_i - X_i \boldsymbol{\beta})' V_i^{-1} (\mathbf{y}_i - X_i \boldsymbol{\beta}) \\ &= -\frac{NT}{2} \log 2\pi - \frac{1}{2} \left[\sum_{i=1}^{N+(T-1)m} (q_i - 1) \right] \log \sigma_u^2 \\ &\quad - \frac{1}{2} \sum_{i=1}^{N+(T-1)m} \log(\sigma_u^2 + q_i \sigma_\alpha^2) \\ &\quad - \frac{1}{2} \sum_{i=1}^{N+(T-1)m} (\mathbf{y}_i - X_i \boldsymbol{\beta})' V_i^{-1} (\mathbf{y}_i - X_i \boldsymbol{\beta}). \end{aligned} \tag{11.1.11}$$

Conditioning on σ_u^2 and σ_α^2, the MLE is the GLS (11.1.8). Conditioning on $\boldsymbol{\beta}$, the MLEs of σ_u^2 and σ_α^2 are the simultaneous solutions of the following equations:

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma_u^2} &= -\frac{1}{2\sigma_u^2} \left[\sum_{i=1}^{N+(T-1)m} (q_i - 1) \right] \\ &\quad - \frac{1}{2} \left[\sum_{i=1}^{N+(T-1)m} \frac{1}{(\sigma_u^2 + q_i \sigma_\alpha^2)} \right] \\ &\quad + \frac{1}{2\sigma_u^4} \sum_{i=1}^{N+(T-1)m} (y_i - X_i \boldsymbol{\beta})' Q_i (y_i - X_i \boldsymbol{\beta}) \\ &\quad + \frac{1}{2} \sum_{i=1}^{N+(T-1)m} \frac{q_i}{(\sigma_u^2 + q_i \sigma_\alpha^2)^2} (\bar{y}_i - \bar{\mathbf{x}}_i' \boldsymbol{\beta})^2 = 0 \end{aligned} \tag{11.1.12}$$

and

$$\frac{\partial \log L}{\partial \sigma_\alpha^2} = -\frac{1}{2}\sum_{i=1}^{N+(T-1)m} \times \left[\frac{q_i}{\sigma_u^2 + q_i\sigma_\alpha^2} - \frac{q_i^2}{(\sigma_u^2 + q_i\sigma_\alpha^2)^2}(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta})^2\right] = 0 \tag{11.1.13}$$

where $Q_i = I_{q_i} - (1/q_i)\mathbf{e}_{q_i}\mathbf{e}'_{q_i}$, and $\mathbf{e}_{q_i}$ is a $q_i \times 1$ vector of ones. Unfortunately, because q_i are different for different i, (11.1.12) and (11.1.13) cannot be put in the simple form of (3.3.25) and (3.3.26). Numerical methods will have to be used to obtain a solution. However, computation of the MLEs of $\boldsymbol{\beta}$, σ_u^2 and σ_α^2 can be simplified by iteratively switching between (11.1.8) and (11.1.12)–(11.1.13).

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

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=t_i}^{Ti}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\right]^{-1} \cdot \left[\sum_{i=1}^{N}\sum_{t=t_i}^{Ti}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i)\right]. \tag{11.1.14}$$

If the model is dynamic, similar modification of the GMM (e.g., (4.3.47)) (e.g., Collado (1997); Moffitt 1993) can be applied to obtain consistent estimators of the coefficients. The likelihood approach for dynamic models has the issue of initial conditions.[2] Different assumptions about initial conditions will suggest different ways of incorporating new observations with those already in the sample. If α_i are treated as random, it would appear a reasonable approximation in this case is to modify the methods based on the assumption that initial observations are correlated with individual effects and have stationary variances (Chapter 4, Case IVc or IVc′). However, the assumption imposed on the model will have to be even more restrictive. If α_i are treated as fixed, a similar modification can be applied to the transform MLE (e.g., (4.5.6)).

When data are randomly missing, a common procedure is to focus on the subset of individuals for which complete time series observations are available. However, the subset of incompletely observed individuals also contains some information about unknown parameters. A more efficient and computationally somewhat more complicated way is to treat randomly missing samples in the same way as rotating samples. For instance, the likelihood function (11.1.1), with the modification that $t_i = 1$ for all i, can also be viewed the likelihood function for this situation: In time period 1 there are $N + \sum_{t=1}^{T-1} m_t$ individuals; in period 2, m_1 of them randomly drop out, and so on, such that at the end of T periods there are only N individuals remaining in the sample. Thus, the

[2] For details, see Chapters 4 and 6.

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

procedure for obtaining the GLS or MLE for unknown parameters with all the observations utilized is similar to the situation of rotating samples.

To test if attrition is indeed random, we note that either the complete sample unbalanced panel estimators discussed above or the estimators based on the balanced panel subsample estimators converge to the true value under the null. Under the alternative that attrition is behaviorally related, neither estimators are consistent. However, if the individual-specific effects α_i and the error u_{it} are independent of the regressors $\mathbf{x}_{it}$, and are independently normally distributed, a test of random attrition versus behaviorally related attrition is a student t-test of the significance of sample selection effect (e.g., (8.2.7)). If α_i are correlated with $\mathbf{x}_{it}$, one can construct a Hausman (1978) type test statistic for the significance of the difference between the Kyriazidou (1997) fixed effects sample selection estimator (e.g., (8.5.4)) and the complete sample unbalanced panel data with estimator (11.1.14). Further, if all initial samples are observed for at least two periods before attrition occurs, then the within estimator based on initial complete samples within estimator and the within estimator based on all observed samples (unbalanced panel) converge to the true value under the null and converge to different values under the alternative. A straightforward Hausman (1978) test statistic,

$$(\hat{\boldsymbol{\beta}}_{cv} - \tilde{\boldsymbol{\beta}}_{cvs})' \left[\text{Cov}(\tilde{\boldsymbol{\beta}}_{cvs}) - \text{Cov}(\hat{\boldsymbol{\beta}}_{cv}) \right]^{-1} (\hat{\boldsymbol{\beta}}_{cv} - \tilde{\boldsymbol{\beta}}_{cvs}) \qquad (11.1.15)$$

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

11.2 PSEUDO-PANELS (OR REPEATED CROSS-SECTIONAL DATA)

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

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}, \qquad (11.2.1)$$

if α_i and $\mathbf{x}_{it}$ are correlated through the fixed effects estimator discussed in Chapter 3.[3] However, several authors have argued that with some additional assumptions $\boldsymbol{\beta}$ may be identifiable from a single cross-section or a series of

[3] If α_i and $\mathbf{x}_{it}$ are uncorrelated, there is no problem of consistently estimating (11.2.1) with repeated cross-sectional data because $E(\alpha_i + u_{it} \mid \mathbf{x}_{it}) = 0$.

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

independent cross-sections (e.g., Blundell, Browning, and Meghir 1994; Deaton 1985; Heckman and Robb 1985; Moffitt 1993).

Deaton (1985) suggests using a cohort approach to obtain consistent estimators of $\boldsymbol{\beta}$ of (11.2.1) if repeated cross-sectional data are available. In this approach individuals sharing common observed characteristics, such as age, sex, education, or socioeconomic background are grouped into *cohorts*. For instance, suppose that one can divide the sample into C cohorts in terms of an $L \times 1$ vector of individual characteristics, $\mathbf{z}_c$, $c = 1, \ldots, C$. Let $\mathbf{z}_{it}$ be the corresponding L-dimensional vector of individual-specific variables for the ith individual of the tth cross-sectional data. Then $(y_{it}, \mathbf{x}_{it})$ belong to the cth cohort if $\mathbf{z}_{it} = \mathbf{z}_c$. Let $\psi_{ct} = \{i \mid \mathbf{z}_{it} = \mathbf{z}_c$ for the tth cross-sectional data$\}$ be the set of individuals that belong to the cohort c at time t, $c = 1, \ldots, C, t = 1, \ldots, T$. Let N_{ct} be the number of individuals in ψ_{ct}. Deaton (1985) assumes individuals belonging to the same cohort have the same specific effects,

$$\alpha_i = \sum_{c=1}^{C} \alpha_c d_{itc}, \tag{11.2.2}$$

where $d_{itc} = 1$ if the ith individual of the tth cross-sectional data belongs to cohort c and 0 otherwise. Let $\bar{y}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} y_{it}$ and $\bar{\mathbf{x}}_{ct} = \frac{1}{N_{ct}} \sum_{i \epsilon \psi_{ct}} \mathbf{x}_{it}$, then the data $(\bar{y}_{ct}, \bar{\mathbf{x}}'_{ct})$ becomes a pseudo-panel with repeated observations on C cohorts over T time periods. Aggregation of observations to cohort level for the model (11.2.1) leads to

$$\bar{y}_{ct} = \bar{\mathbf{x}}'_{ct}\boldsymbol{\beta} + \alpha_c + \bar{u}_{ct}, \qquad \begin{array}{l} c = 1, \ldots, C, \\ t = 1, \ldots, T, \end{array} \tag{11.2.3}$$

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

If $\mathbf{x}_{it}$ are uncorrelated with u_{it}, the within estimator (3.2.8) can be applied to the pseudo panel

$$\hat{\boldsymbol{\beta}}_w = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)' \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{y}_{ct} - \bar{y}_c) \right), \tag{11.2.4}$$

where $\bar{\boldsymbol{x}}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{\boldsymbol{x}}_{ct}$, and $\bar{y}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{y}_{ct}$. When $T \rightarrow \infty$ or if T is fixed but $N \rightarrow \infty$, $C \rightarrow \infty$, and $\frac{C}{N} \longrightarrow 0$, (11.2.4) is consistent.

Although the cohort approach offers a useful framework to make use of independent cross-sectional information, there are problems with some of its features. First, the assertion of intra-cohort homogeneity (11.2.2) appears very strong, in particular, in view of the cohort classification is often arbitrary. Second, the practice of establishing the large sample properties of econometric estimators and test statistics by assuming that the number of cohorts, C, tends to infinity is not satisfactory. There is often a physical limit beyond which one cannot increase the number of cohorts. The oft-cited example

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

of date of birth cohorts is a case in point. Third, grouping or aggregating individuals may result in the loss of information. Moreover, in general, the number of individuals at different cohorts or different time are different, $N_{ct} \neq N_{c's}$. Even u_{it} is homoscedastic and independently distributed, Var $(\bar{u}_{ct}) = \frac{\sigma_u^2}{N_{ct}} \neq$ Var $(\bar{u}_{c's}) = \frac{\sigma_u^2}{N_{c's}}$. Therefore, the t-statistic based on the conventional within estimator formula is not asymptotically standard normally distributed unless $N_{ct} = N_{c's}$ for all c, c', t, s, and var (u_{it}) is a constant across i. Hence, the resulting inference can be misleading (Inoue 2008).

Suppose (11.2.2) indeed holds and if u_{it} is independently, identically distributed the problem of heterocesdasticity of $\bar{u}_{ct}$ can be corrected by applying the weighted within estimator,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ww} = &\left\{\sum_{c=1}^{C}\sum_{t=1}^{T}\left[N_{ct}(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)'\right]\right\}^{-1} \\ &\cdot\left\{\sum_{c=1}^{C}\sum_{t=1}^{T}\left[N_{ct}(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{y}_{ct} - \bar{y}_c)\right]\right\}.\end{aligned} \tag{11.2.5}$$

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

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{ww}) = \sigma^2\left\{\sum_{c=1}^{C}\sum_{t=1}^{T}\left[N_{ct}(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{x}_{ct} - \bar{\mathbf{x}}_c)'\right]\right\}^{-1}. \tag{11.2.6}$$

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

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}, \tag{11.2.7}$$

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

$$E^*(y_{it} \mid \mathbf{z}_{it}) = \mathbf{z}_{it}'\boldsymbol{\delta}_t, \tag{11.2.8}$$

where $E^*(y \mid \mathbf{z})$ denotes the minimum mean-square-error linear predictor of y by $\mathbf{z}$. Let $\hat{y}_{i,t-1} = \mathbf{z}_{i,t-1}\hat{\boldsymbol{\delta}}_{t-1}$, then (11.2.7) becomes

$$y_{it} = \gamma\hat{y}_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{11.2.9}$$

where

$$v_{it} = \alpha_i + u_{it} + \gamma(y_{i,t-1} - \hat{y}_{i,t-1}). \tag{11.2.10}$$

Girma (2000); Moffitt (1993); and McKenzie (2004) assume that $\mathbf{z}_{it} = \mathbf{z}_c$, are a set of cohort dummies for all i belonging to cohort c. This is equivalent to simply using the dummy variable d_{itc}, as instruments for y_{it} where $d_{itc} = 1$ if y_{it} belongs to cohort c and 0 otherwise, for $c = 1, \ldots, C$. Taking the average of y_{it} or $\hat{y}_{it}$ for i belonging to cohort c leads to the following pseudo panel dynamic model

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \alpha_c + \bar{\mathbf{x}}'_{ct}\boldsymbol{\beta} + v_{ct}, \quad c = 1, \ldots, C, \quad t = 1, \ldots, T, \tag{11.2.11}$$

where all variables denote period-by-period averages within each cohort. The covariance estimator of (11.2.11) would be consistent estimators of γ and $\boldsymbol{\beta}$ provided

$$\text{Cov}\,(v_{ct}, \bar{y}_{c,t-1}) = 0, \tag{11.2.12}$$

and

$$\text{Cov}\,(v_{ct}, \bar{\mathbf{x}}_{ct}) = \mathbf{0}. \tag{11.2.13}$$

However, even under the assumption (11.2.2),

$$E[(\alpha_i + u_{it})\mathbf{z}_{i,t-1} \mid i \in \psi_{c,t-1}] = \mathbf{0}, \tag{11.2.14}$$

in general,

$$E[(y_{i,t-1} - \hat{y}_{i,t-1})\hat{y}_{i,t-1}] \neq 0. \tag{11.2.15}$$

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

$$\text{Cov}\,[(\alpha_i + u_{it})\mathbf{x}_{it}] = \mathbf{0}, \tag{11.2.16}$$

(11.2.13) is unlikely to hold because $\mathbf{x}_{i,t-1}$ drives $y_{i,t-1}$ and $\mathbf{x}_{it}$ is likely to be serially correlated. To overcome the problem of correlations between the regressors and errors in (11.2.11), one will have to also find instruments for $\mathbf{x}_{it}$ as well. Unfortunately, the availability of such instruments in addition to $\mathbf{z}_i$ in many applications may be questionable (e.g., Verbeek and Vella 2005). It remains to be seen whether in empirical applications of cohort approach suitable instruments can be found that have time-varying relationships with $\mathbf{x}_{it}$ and $y_{i,t-1}$, while in the meantime they should not have any time-varying relationship with the error term (11.2.10) (e.g., Verbeek 2007; Verbeek and Vella 2005).

11.3 POOLING OF SINGLE CROSS-SECTIONAL AND SINGLE TIME SERIES DATA

11.3.1 Introduction

In this section we consider the problem of pooling when we have a single cross-sectional and a single time series data set. Empirical studies based solely on time series data often result in very inaccurate parameter estimates because of the high collinearity among the explanatory variables. For instance, income

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

and price time series can be highly correlated. On the other hand, a cross-sectional data set may contain good information on household income, but not on price, because the same price is likely to be faced by all households. Thus, each data set contains useful information on some of the variables, but not on all the variables to allow accurate estimates of all the parameters of interest. A classic example of this is provided in a study (Stone 1954) of aggregate-demand systems in which there was no cross-sectional variation in commodity prices and inadequate time-series variation in real incomes.

To overcome the problem of lack of information on interesting parameters from time series or cross-sectional data alone, one frequently estimates some parameters from cross-sectional data, then introduces these estimates into time series regression to estimate other parameters of the model. For instance, Tobin (1950) calculated income elasticity from cross-sectional data, then multiplied it by the time series income variable and subtracted the product from the annual time series of quantity demand to form a new dependent variable. This new dependent-variable series was then regressed against the time series of the price variable to obtain an estimate of the price elasticity of demand.

The purpose of pooling here, as in the cases analyzed earlier, is to get more efficient estimates for the parameters that are of interest. In a time series, the number of observations is usually limited, and variables are highly correlated. Moreover, an aggregate data set, or a single individual time series data set does not contain information on micro-sociodemographic variables that affect economic behavior. Neither are cross-sectional data more structurally complete. Observations on individuals at one point in time are likely to be affected by prior observations. These raise two fundamental problems: One is that the source of estimation bias in cross-sectional estimates may be different from that in time series estimates. In fact, many people have questioned the suitability and comparability of estimates from different kinds of data (micro or aggregate, cross section or time series) (e.g., Kuh 1959; Kuh and Meyer 1957). The second is that if pooling is desirable, what is the optimal way to do it? It turns out that both problems can be approached simultaneously from the framework of an analysis of the likelihood functions (Maddala 1971b) or a Bayesian approach (Hsiao et al. 1995).

The likelihood function provides a useful way to extract the information contained in the sample provided that the model is correctly specified. Yet a model is a simplification of complex real-world phenomena. To be most useful, a model must strike a reasonable balance between realism and manageability. It should be realistic in incorporating the main elements of the phenomena being represented and at the same time be manageable in eliminating extraneous influences. Thus, when specifying a regression equation, it is common to assume that the numerous factors that affect the outcome of the dependent variable, but are individually unimportant or unobservable, can be appropriately summarized by a random disturbance term. However, the covariations of these omitted variables and the included explanatory variables in a cross-sectional regression may be different from those in a time series regression. For example,

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

if high income is associated with high consumption levels and is also correlated with age, the regression of consumption on income cross-sectionally will yield an income coefficient that measures the joint effects of age and income on consumption, unless age is introduced as another explanatory variable. But the age composition of the population could either be constant or be subject only to gradual, slow change in aggregate time series. Hence, the time series estimate of the income elasticity, ignoring the age variable, could be smaller than the cross-sectional estimates because of the negligible age-income correlation.

Another reason that cross-sectional and time series estimates in demand analysis may differ is that cross-sectional estimates tend to measure long-run behavior and time series estimates tend to measure short-run adjustment (Kuh 1959; Kuh and Meyer 1957). The assumption is that the majority of the observed families have enjoyed their present positions for some time, and the disequilibrium among households tends to be synchronized in response to common market forces and business cycles. Hence, many disequilibrium effects wash out (or appear in the regression intercept), so that the higher cross-sectional slope estimates may be interpreted as long-run coefficients. However, this will not be true for time series observations. Specifically, changes over time usually represent temporary shifts. Recipients or losers from this change probably will not adjust immediately to their new levels. A incompletely adjusted response will typically have a lower coefficient than the fully adjusted response.

These observations on differential cross-sectional and time series behavior suggest that the impacts of omitted variables can be strikingly different in time series and cross sections. Unless the assumption that the random term (representing the omitted-variables effect) is uncorrelated with the included explanatory variables holds, the time series and cross-sectional estimates of the common coefficients can diverge. In fact, if the time series and cross-sectional estimates differ, this is an indication that either or both models are misspecified. In Chapter 3 we discussed specification tests without using extraneous information. We now discuss a likelihood approach when extraneous information in the form of cross-sectional data for the time series model, or time series data for the cross-sectional model, is available.

11.3.2 The Likelihood Approach to Pooling Cross-Sectional and Time Series Data

Assume that we have a single cross section consisting of N units and a time series extending over T time periods. Suppose that the cross-sectional model is

$$\mathbf{y}_c = Z_1\boldsymbol{\delta}_1 + Z_2\boldsymbol{\delta}_2 + \mathbf{u}_c, \tag{11.3.1}$$

where $\mathbf{y}_c$ is an $N \times 1$ vector of observations on the dependent variable, Z_1 and Z_2 are $N \times K$ and $N \times L$ matrices of independent variables, and $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

$K \times 1$ and $L \times 1$ vectors of parameters, respectively. The $N \times 1$ error term $\mathbf{u}_c$ is independently distributed, with variance–covariance matrix $\sigma_u^2 I_N$.

The time series model is

$$\mathbf{y}_T = X_1 \boldsymbol{\beta}_1 + X_2 \boldsymbol{\beta}_2 + \mathbf{v}_T. \tag{11.3.2}$$

where $\mathbf{y}_T$ is a $T \times 1$ vector of observations on the dependent variable, X_1 and X_2 are $T \times K$ and $T \times M$ matrices of observations on the independent variables, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $K \times 1$ and $M \times 1$ vectors of parameters, and $\mathbf{v}_T$ is a $T \times 1$ vector of disturbances.[4] For simplicity, we assume that $\mathbf{v}_T$ is uncorrelated with $\mathbf{u}_c$ and is serially uncorrelated, with the variance–covariance matrix $E\mathbf{v}_T\mathbf{v}_T' = \sigma_v^2 I_T$.

The null hypothesis here is that $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1$. So with regard to the question whether or not to pool, we can use a likelihood-ratio test. Let L_1^* and L_2^* denote the maxima of the log joint likelihood functions for (11.3.1) and (11.3.2) with and without the restriction that $\delta_1 = \beta_1$. Then, under the null hypothesis, $2(L_2^* - L_1^*)$ is asymptotically χ^2 distributed, with K degrees of freedom. The only question is: What is the appropriate level of significance? If the costs of mistakenly accepting the pooling hypothesis and rejecting the pooling hypothesis are the same, Maddala (1971b) suggested using something like a 25 to 30 percent level of significance, rather than the conventional 5 percent, in our preliminary test of significance.

The specifications of the maximum-likelihood estimates and their variance–covariances merely summarize the likelihood function in terms of the location of its maximum and its curvature around the maximum. It is possible that the information that the likelihood function contains is not fully expressed by these. When the compatibility of cross-sectional and time series estimates is investigated, it is useful to plot the likelihood function extensively. For this purpose, Maddala (1971b) suggested that one should also tabulate and plot the relative maximum likelihoods of each data set,

$$R_M(\delta_1) = \frac{\underset{\boldsymbol{\theta}}{\text{Max}} \; L(\boldsymbol{\delta}_1, \boldsymbol{\theta})}{\underset{\boldsymbol{\delta}_1, \boldsymbol{\theta}}{\text{Max}} \; L(\boldsymbol{\delta}_1, \boldsymbol{\theta})}, \tag{11.3.3}$$

where $\boldsymbol{\theta}$ represents the set of nuisance parameters, $\max_\theta L(\boldsymbol{\delta}_1, \boldsymbol{\theta})$ denotes the maximum of L with respect to $\boldsymbol{\theta}$, given $\boldsymbol{\delta}_1$ and $\max_{\boldsymbol{\delta}_1, \boldsymbol{\theta}} L(\boldsymbol{\delta}_1, \boldsymbol{\theta})$ denotes the maximum of L with respect to both $\boldsymbol{\delta}_1$ and $\boldsymbol{\theta}$. The plot of (11.3.3) summarizes almost all the information contained in the data on $\boldsymbol{\delta}_1$. Hence, the shapes and locations of the relative maximum likelihoods will reveal more information

[4] If the cross-sectional data consist of all individuals in the population, then in the year in which cross-sectional observations are collected, the sum across individual observations of a variable should be equal to the corresponding aggregate time-series variable. Because in most cases cross-sectional samples consist of a small portion of the population, we shall ignore this relation and assume that the variables are unrelated.

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

about the compatibility of the different bodies of data than a single test statistic can.

If the hypothesis $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1$ is acceptable, then, as Chetty (1968), Durbin (1953), and Maddala (1971b) have suggested, we can stack (11.3.1) and (11.3.2) together as

$$\begin{bmatrix} \mathbf{y}_c \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} Z_1 \\ X_1 \end{bmatrix} \boldsymbol{\delta}_1 + \begin{bmatrix} Z_2 \\ \mathbf{0} \end{bmatrix} \boldsymbol{\delta}_2 + \begin{bmatrix} \mathbf{0} \\ X_2 \end{bmatrix} \boldsymbol{\beta}_2 + \begin{bmatrix} \mathbf{u}_c \\ \mathbf{v}_T \end{bmatrix}. \qquad (11.3.4)$$

It is clear that an efficient method of estimating of $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$, and $\boldsymbol{\beta}_2$ is to apply the maximum-likelihood method to (11.3.4). An asymptotically equivalent procedure is to first apply least-squares separately to (11.3.1) and (11.3.2) to obtain consistent estimates of σ_u^2 and σ_v^2, and then substitute the estimated σ_u^2 and σ_v^2 into the equation

$$\begin{bmatrix} \frac{1}{\sigma_u} y_c \\ \frac{1}{\sigma_v} y_T \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_u} Z_1 \\ \frac{1}{\sigma_v} X_1 \end{bmatrix} \boldsymbol{\delta}_1 + \begin{bmatrix} \frac{1}{\sigma_u} Z_2 \\ \mathbf{0} \end{bmatrix} \boldsymbol{\delta}_2 + \begin{bmatrix} \mathbf{0} \\ \frac{1}{\sigma_v} X_2 \end{bmatrix} \boldsymbol{\beta}_2 + \begin{bmatrix} \frac{1}{\sigma_u} \mathbf{u}_c \\ \frac{1}{\sigma_v} \mathbf{v}_T \end{bmatrix} \qquad (11.3.5)$$

and apply the least-squares method to (11.3.5).

The conventional procedure of substituting the cross-sectional estimates of $\boldsymbol{\beta}_1$, $\hat{\boldsymbol{\delta}}_{1c}$, into the time series model

$$\mathbf{y}_T - X_1\hat{\boldsymbol{\delta}}_{1c} = X_2\boldsymbol{\beta}_2 + \mathbf{v}_T + X_1(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\delta}}_{1c}), \qquad (11.3.6)$$

and then regressing $(\mathbf{y}_T - X_1\hat{\boldsymbol{\delta}}_{1c})$ on X_2, yields only conditional estimates of the parameters $\boldsymbol{\beta}_2$ – conditional on the estimates obtained from the cross-sectional data.[5] However, there is also some information about $\boldsymbol{\beta}_1$ in the time series sample, and this should be utilized. Moreover, one should be careful in the use of two-step procedures. Proper evaluation of the asymptotic variance–covariance matrix of $\boldsymbol{\beta}_2$ should take account of the uncertainty (variance) in substituting $\hat{\boldsymbol{\delta}}_{1c}$ for $\boldsymbol{\beta}_1$. (For details, see Chetty 1968; Hsiao et al. 1995; Jeong 1978; and Maddala 1971b.)

[5] In the Bayesian framework this is analogous to making inferences based on the conditional distribution of $\boldsymbol{\beta}_2$, $f(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1 = \boldsymbol{\delta}_{1c})$, whereas it is the marginal distribution of $\boldsymbol{\beta}_2$ that should be used whenever $\boldsymbol{\beta}_1$ is not known with certainty. For details see Chetty (1968).

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

11.3.3 An Example

To illustrate application of the likelihood approach to pooling, Maddala (1971b) analyzed a simple econometric model relating to the demand for food in the United States. The model and the data were taken from Tobin (1950).

The cross-sectional demand equation is

$$y_{1i} = \delta_0 + \delta_1 z_{1i} + \delta_2 z_{2i} + u_i, \qquad i = 1, \ldots, N, \tag{11.3.7}$$

where y_{1i} is the logarithm of the average food consumption of the group of families at a point in time, and z_{1i} and z_{2i} are the logarithms of the average income of the ith family and the average family size, respectively. The time series demand function is

$$y_{2t} = \beta_0 + \beta_1(x_{1t} - \beta_2 x_{2t}) + \beta_3(x_{2t} - x_{2,t-1}) + v_t, \qquad t = 1, \ldots, T. \tag{11.3.8}$$

where y_{2t}, x_{1t}, and x_{2t} are the logarithms of the food price index, per capita food supply for domestic consumption, and per capita disposable income, respectively. The income elasticity of demand, δ_1, was assumed common to both regressions, namely, $\delta_1 = \beta_2$. The error terms u_i and v_t were independent of each other and were assumed independently normally distributed, with 0 means and constant variances σ_u^2 and σ_v^2, respectively.

The results of the cross-sectional estimates are

$$\begin{aligned} \hat{y}_{1i} = 0.569 + {} & 0.5611 z_{1i} + 0.2540 z_{2i} \\ & (0.0297) + (0.0367) \end{aligned}, \tag{11.3.9}$$

where standard errors are in parentheses. The results of the time series regression are

$$\begin{aligned} \hat{y}_{2t} = 7.231 + {} & 1.144 x_{2t} - 0.1519(x_{2t} - x_{2,t-1}) - 3.644 x_{1t} \\ & (0.0612) \quad (0.0906) \qquad\qquad\qquad (0.4010). \end{aligned} \tag{11.3.10}$$

The implied income elasticity, δ_1, is 0.314.

When the cross-sectional estimate of δ_1, 0.56, is introduced into the time series regression, the estimated β_1 is reduced to -1.863, with a standard error of 0.1358. When δ_1 and β_1 are estimated simultaneously by the maximum-likelihood method, the estimated δ_1 and β_1 are 0.5355 and -1.64, with a covariance $\begin{bmatrix} 0.00206 & 0.00827 \\ & 0.04245 \end{bmatrix}$.

Although there is substantial improvement in the accuracy of the estimated coefficient using the combined data, the likelihood-ratio statistic turns out to be 17.2, which is significant at the 0.001 level with 1 degree of freedom. It strongly suggests that in this case we should not pool the time series and cross-sectional data.

Figure 11.1 reproduces Maddala's plot of the relative maximum likelihood $R_M(\delta_1)$ for the parameter δ_1 (the income elasticity of demand) in the Tobin

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

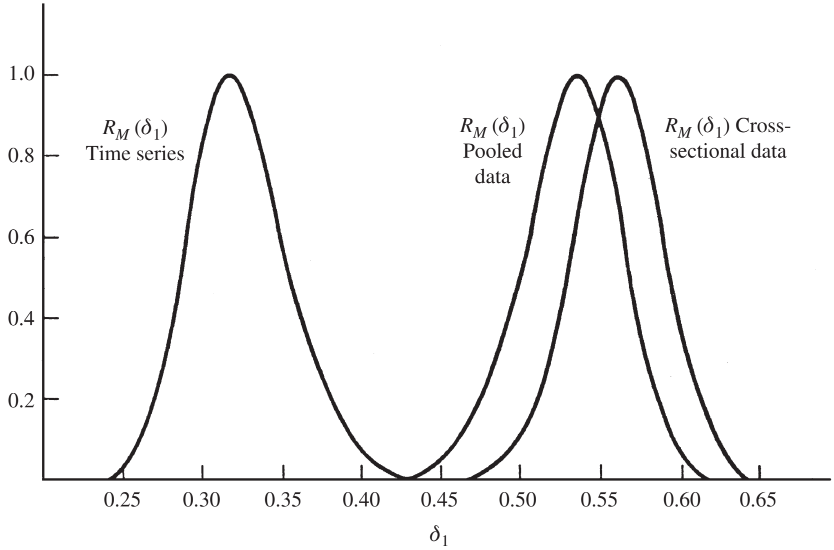

Figure 11.1. Relative maximum likelihood for the parameter δ_1. *Source:* Maddala (1971b, fig. 1).

model from cross-sectional data alone, from time series data alone, and from the pooled sample. The figure reveals that the information on δ_1 provided by the time series data is almost as precise as that provided by the cross-sectional data (otherwise the likelihood function would be relatively flat). Furthermore, there is very little overlap between the likelihood functions from time series and cross-sectional data. Again, this unambiguously suggests that the data should not be pooled.[6]

Given that the time series data arise by aggregating some microeconomic process, there cannot possibly be a conflict between the time series and cross-sectional inferences if individual differences conditional on explanatory variables are viewed as chance outcomes. Thus, whenever the empirical results differ systematically between the two, as in the foregoing example, this is an indication that either or both models may be misspecified. The existence of supporting extraneous information in the form of cross-sectional or time series data provides an additional check to the appropriateness of a model specification that cannot be provided by a single cross-sectional or time series data set, because there may be no internal evidence of this omitted-variable bias. However, until a great deal is learned about the cross-sectional time series relation,

[6] It should be noted that the foregoing result is based on the assumption that both u_i and v_t are independently normally distributed. In practice, careful diagnostic checks should be performed before exploring the pooling issue, using the likelihood-ratio test or relative maximum likelihoods. In fact, Izan (1980) redid the analysis by allowing v_t to follow a first-order autoregressive process. The likelihood-ratio test after allowing for autocorrelation resulted in accepting the pooling hypothesis.

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

there appears no substitute for the completeness of information. Sequential observations on a number of individuals or panel data are essential for a full understanding of the systematic interrelations at different periods of time.

11.4 ESTIMATING DISTRIBUTED LAGS IN SHORT PANELS[7]

11.4.1 Introduction

Because of technical, institutional, and psychological rigidities, often behavior is not adapted immediately to changes in the variables that condition it. In most cases this adaptation is progressive. The progressive nature of adaptations in behavior can be expressed in various ways. Depending on the rationale behind it, we can set up an autoregressive model with the current value of y being a function of lagged dependent variables and exogenous variables, or we can set up a distributed-lag model, with the current value of y being a function of current and previous values of exogenous variables. Although usually a linear distributed-lag model can be expressed in an autoregressive form, and, similarly, as a rule any stable linear autoregressive model can be transformed into a distributed-lag model,[8] the empirical determination of time lags is very important in applied economics. The roles of many economic measures can be correctly understood only if we know when they will begin to take effect and when their effects will be fully worked out. Therefore, we would like to distinguish these two types of dynamic models when a precise specification (or reasoning) is possible. In Chapter 4 we discussed the issues of estimating autoregressive models with panel data. In this section we discuss estimation of distributed-lag models (Pakes and Griliches 1984).

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

$$y_t = \mu + \sum_{\tau=0}^{\infty} \beta_\tau x_{t-\tau} + u_t, \quad t = 1, \ldots, T, \tag{11.4.1}$$

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

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

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

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

priori, that all of them are functions of only a very small number of parameters (Koyck lag, Almon lag, etc.) (Dhrymes 1971; Malinvaud 1970).

On the other hand, when there are N time series, we can use cross-sectional information to identify and estimate (at least some of the) lag coefficients without having to specify a priori that the sequence of lag coefficients progresses in a particular way. For instance, consider the problem of using panel data to estimate the model (11.4.1), which for a given t we rewrite as

$$y_{it} = \mu + \sum_{\tau=0}^{t-1} \beta_\tau x_{i,t-\tau} + b_{it} + u_{it}, \qquad i = 1, \ldots, N, \tag{11.4.2}$$

where

$$b_{it} = \sum_{\tau=0}^{\infty} \beta_{t+\tau} x_{i,-\tau} \tag{11.4.3}$$

is the contribution of the unobserved presample x values to the current values of y, to which we shall refer as the truncation remainder. Under certain assumptions about the relationships between the unobserved b_{it} and the observed x_{it}, it is possible to obtain consistent estimates of β_τ, $\tau = 0, \ldots, t-1$, by regressing (11.4.2) cross-sectionally. Furthermore, the problem of collinearity among $x_t, x_{t-1}, \ldots,$ in a single time series can be reduced or avoided by use of the cross-sectional differences in individual characteristics.

11.4.2 Common Assumptions

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

Thus, to combine N time series to estimate a distributed-lag model, we have to impose restrictions on the distribution of lag coefficients across cross-sectional units and/or on the way the unobserved presample terms affect current

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

behavior. Pakes and Griliches (1984) considered a distributed-lag model of the form

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{\infty} \beta_{i\tau} x_{i,t-\tau} + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{11.4.4}$$

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

Assumption 11.4.1: $E(\beta_{i\tau}) = \beta_\tau$.

Assumption 11.4.2: Let $\bar{\beta}_{i\tau} = \beta_{i\tau} - \beta_\tau, \xi_{it} = \sum_{\tau=0}^{\infty} \bar{\beta}_{i\tau} x_{i,t-\tau}$, and $\boldsymbol{\xi}_i' = (\xi_{i1}, \ldots, \xi_{iT})$; then $E^*[\boldsymbol{\xi}_i \mid \mathbf{x}_i] = \mathbf{0}$.

Assumption 11.4.3: $E^*(\alpha_i^* \mid \mathbf{x}_i) = \mu + \mathbf{a}'\mathbf{x}_i$

Here $E^*(Z_1 \mid Z_2)$ refers to the minimum mean-square-error linear predictor (or the projection) of Z_1 onto Z_2; $\mathbf{x}_i$ denotes the vector of all observed $\mathbf{x}_{it}$. We assume that there are $\ell + 1$ observations on x before the first observation on y, and the $1 \times (\ell + 1 + T)$ vector $\mathbf{x}_i' = [x_{i,-\ell}, \ldots, x_{iT}]$ is an independent draw from a common distribution with $E(\mathbf{x}_i \mathbf{x}_i') = \sum_{xx}$ positive definite.[9]

A sufficient condition for Assumption 11.4.2 to hold is that differences in lag coefficients across individuals are uncorrelated with the $\mathbf{x}_i$ [i.e., $\beta_{i\tau}$ is a random variable defined in the sense of Swamy (1970), or see Chapter 6]. However, Assumption 11.4.3 does allow for individual-specific constant terms (the α_i^*) to be correlated with $\mathbf{x}_i$. The combination of Assumptions 11.4.1–11.4.3 is sufficient to allow us to identify the expected value of the lag-coefficient sequence $\{\beta_\tau\}$ if both N and T tend to infinity.

If T is fixed, substituting Assumptions 11.4.1 and 11.4.2 into equation (11.4.4), we rewrite the distributed-lag model as

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{t+\ell} \beta_\tau x_{i,t-\tau} + b_{it} + \tilde{u}_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{11.4.5}$$

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

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

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

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

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

In this subsection we consider the identification of a distributed-lag model using a kind of restriction different from that in the usual approach of constraining lag coefficients. Our interest is focused on estimating at least some of the population parameters $\beta_\tau = E(\beta_{i\tau})$ for $\tau = 0, 1, \dots,$ without restricting β_τ to be a function of a small number of parameters. We consider a lag coefficient identified if it can be calculated from the matrix of coefficients obtained from the projection of $\mathbf{y}_i$ onto $\mathbf{x}_i$, a $T \times (T + \ell + 1)$ matrix labeled Π, where $E^*(\mathbf{y}_i \mid \mathbf{x}_i) = \boldsymbol{\mu}^* + \Pi \mathbf{x}_i$, $\boldsymbol{\mu}^* = (\mu_1^*, \dots, \mu_T^*)'$ and $\mathbf{y}_i' = (y_{i_1}, \dots, y_{iT})$ is a $1 \times T$ vector.

Equation (11.4.5) makes it clear that each row of Π will contain a combination of the lag coefficients of interest and the coefficients from the projections of the two unobserved components, α_i^* and b_{it}, on $\mathbf{x}_i$. Therefore, the problem is to separate out the lag coefficients from the coefficients defining these two projections.

Using equation (11.4.5), the projection of $\mathbf{y}_i$ onto $\mathbf{x}_i$ and α_i^* is given by[10]

$$E^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i^*) = [B + W]\mathbf{x}_i + [\mathbf{e} + \mathbf{c}]\alpha_i^* \tag{11.4.6}$$

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

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

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

$$E^*[\mathbf{b}_i \mid \mathbf{x}_i, \alpha_i^*] = W\mathbf{x}_i + \mathbf{c}\alpha_i^*. \tag{11.4.7}$$

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

Equation (11.4.6) and the fact that $E^*\{E^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i^*) \mid \mathbf{x}_i\} = E^*[\mathbf{y}_i \mid \mathbf{x}_i] = (\mathbf{e} + \mathbf{c})\mu + \Pi \mathbf{x}_i$ imply that

$$\Pi = B + [W + (\mathbf{e} + \mathbf{c})\mathbf{a}']. \tag{11.4.8}$$

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

Clearly, if the $T \times (T + \ell + 1)$ matrix W is unrestricted, we cannot separate out the lag coefficients, B, and the impact of the truncation-remainder term from the Π matrix. But given that $\mathbf{ca}'$ is a matrix of rank 1, we may be able to identify some elements of B if there are restrictions on W. Thus, to identify some of the lag coefficients from Π, we shall have to restrict W. W will be restricted if it is reasonable to assume that the stochastic process generating $\{x_{it}\}_{t=-\infty}^{T}$ restricts the coefficients on $\mathbf{x}_i$ in the projection of the presample $x_{i,-j}$ values onto the in-sample $\mathbf{x}_i$ and α_i^*. The particular case analyzed by Pakes and Griliches (1984) is given by the following assumption.[11]

Assumption 11.4.4: For $q \geq 1$, $E^*[x_{i,-\ell-q} \mid \mathbf{x}_i,\ \alpha_i^*] = c_q\alpha_i^* + \Sigma_{j=1}^{p}\rho_j^{(q)} x_{i,-\ell+j-1}$. That is, in the projection of the unseen presample x values onto $\mathbf{x}_i$ and α_i^*, only $[x_{i,-\ell}, x_{i,-\ell+1}, \ldots, x_{i,-\ell+p-1}]$ have nonzero coefficients.

If $c_q = 0$, a sufficient condition for Assumption 11.4.4 to hold is that x is generated by a pth-order autoregressive process.[12]

Because each element of $\mathbf{b}_i$ is just a different linear combination of the same presample x values, the addition of Assumption 11.4.4 implies that

$$E^*[b_{it} \mid \mathbf{x}_i, \alpha_i^*] = c_t\alpha_i^* + \sum_{j=1}^{p} w_{t,j-\ell-1}x_{i,j-\ell-1}, \quad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{11.4.9}$$

where $w_{t,j-\ell-1} = \Sigma_{q=1}^{\infty}\beta_{t+\ell+q}\rho_j^{(q)}$, $j = 1, \ldots, p$, and $c_t = \Sigma_{q=1}^{\infty}\beta_{t+l+q}c_q$. This determines the vector $\mathbf{c}$ and the matrix W in (11.4.7). In particular, it implies that W can be partitioned into a $T \times (T + \ell - p + 1)$ matrix of zeros and $T \times p$ matrix of free coefficients,

$$W = \left[\begin{array}{c} \tilde{W} \\ T \times p \end{array} \vdots \begin{array}{c} \mathbf{0} \\ T \times (T + \ell - p + 1). \end{array}\right]. \tag{11.4.10}$$

Substituting (11.4.10) into (11.4.8) and taking partial derivatives of Π with respect to the leading $(T + \ell - p + 1)$ lag coefficients, we can show that the resulting Jacobian matrix satisfies the rank condition for identification of these coefficients (e.g., Hsiao 1983, Theorem 5.1.2). A simple way to check that

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

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

the leading $(T+\ell-p+1)$ lag coefficients are indeed identified is to show that consistent estimators for them exist. We note that by construction, cross-sectional regression of $\mathbf{y}_i$ on $\mathbf{x}_i$ yields consistent estimates of Π. For the special case in which $c_q = 0$, the projections of each period's value of y_{it} on all in-sample values of $\mathbf{x}_i' = (x_{i,-\ell}, x_{i,-\ell+1}, \ldots, x_{iT})$ are[13]

$$\begin{aligned}
E^*(y_{i1} \mid \mathbf{x}_i) &= \mu + \sum_{j=1}^{p} \phi_{1,j-\ell-1} x_{i,j-\ell-1},\\
E^*(y_{i2} \mid \mathbf{x}_i) &= \mu + \beta_0 x_{i2} + \sum_{j=1}^{p} \phi_{2,j-\ell-1} x_{i,j-\ell-1},\\
E^*(y_{i3} \mid \mathbf{x}_i) &= \mu + \beta_0 x_{i3} + \beta_1 x_{i2} + \sum_{j=1}^{p} \phi_{3,j-\ell-1} x_{i,j-\ell-1}\\
&\vdots\\
E^*(y_{iT} \mid \mathbf{x}_i) &= \mu + \beta_0 x_{iT} + \cdots + \beta_{T+\ell-p} x_{i,p-\ell} + \sum_{j=1}^{p} \phi_{T,j-\ell-1} x_{i,j-\ell-1},
\end{aligned} \tag{11.4.11}$$

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

When $c_q \neq 0$, the finding of consistent estimators (hence identification) for the leading $(T+\ell-p+1)$ lag coefficients is slightly more complicated. Substituting (11.4.9) into (11.4.5), we have

$$\begin{aligned}
E^*\left(y_{it} \mid \mathbf{x}_i, \alpha_i^*\right) = (1+c_t)\alpha_i^* &+ \sum_{\tau=0}^{t+\ell-p} \beta_\tau x_{i,t-\tau}\\
&+ \sum_{j=1}^{p} \phi_{t,j-\ell-1} x_{i,j-\ell-1}, \qquad t = 1, \ldots, T,
\end{aligned} \tag{11.4.12}$$

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

[13] The coefficient of (11.4.11) is another way of writing Π (11.4.8).

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

we obtain

$$
\begin{aligned}
E^*(y_{i1} \mid \mathbf{x}_i) &= \mu(1+c_1) + (1+c_1)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p} [(1+c_1)a_{j-\ell-1} + \phi_{1,j-\ell-1}] x_{i,j-\ell-1}, \\
E^*(y_{i2} \mid \mathbf{x}_i) &= \mu(1+c_2) + \beta_0 x_{i2} + (1+c_2)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p} [(1+c_2)a_{j-\ell-1} + \phi_{2,j-\ell-1}] x_{i,j-\ell-1}, \qquad (11.4.13) \\
&\vdots \\
E^*(y_{iT} \mid \mathbf{x}_i) &= \mu(1+c_T) + \sum_{\tau=0}^{T+\ell-p} \beta_\tau x_{i,T-\tau} + (1+c_T)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p} [(1+c_T)a_{j-\ell-1} + \phi_{T,j-\ell-1}] x_{i,j-\ell-1}.
\end{aligned}
$$

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

$$
y_{it} = \tilde{c}_t y_{i1} + \sum_{\tau=0}^{t+\ell-p} \beta_\tau x_{i,t-\tau} + \sum_{j=1}^{p} \tilde{\phi}_{t,j-\ell-1} x_{i,j-\ell-1} + \nu_{it}, \qquad (11.4.14)
$$

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

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

and

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

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

For given t, the only variable on the right-hand side of (11.4.14) that is correlated with ν_{it} is y_{i1}. If we know the values of $\{\tilde{c}_t\}_{t=2}^{T}$, the system (11.4.14) will allow us to estimate the leading $(T + \ell - p + 1)$ lag coefficients consistently by first forming $\tilde{y}_{it} = y_{it} - \tilde{c}_t y_{i1}$ (for $t = 2, \ldots, T$) and then regressing this sequence on in-sample x_{it} values cross-sectionally. In the case in which all c_t values are identical, we know that the sequence $\{\tilde{c}_t\}_{t=2}^{T}$ is just a sequence of 1's. In the case in which α_i^* have a free coefficient in each period of the sample, we have unknown $(1 + c_t)$. However, we can consistently estimate $\tilde{c}_t$, β_τ, and $\tilde{\phi}_{t,j}$

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

by the instrumental-variable method, provided there is at least one x_{is} that is excluded from the determinants of $y_{it} - \tilde{c}_t y_{i1}$ and that is correlated with y_{i1}. If $T \geq 3, x_{i3}, \ldots, x_{iT}$ are excluded from the equation determining $(y_{i2} - \tilde{c}_2 y_{i1})$, and provided that not all of a_3 to a_T are 0, at least one of them will have the required correlation with y_{i1}.

We have shown that under Assumptions 11.4.1–11.4.4, the use of panel data allows us to identify the leading $T + \ell - p + 1$ lag coefficients without imposing any restrictions on the sequence $\{\beta_\tau\}_{\tau=0}^{\infty}$. Of course, if $T + \ell$ is small relative to p, we will not be able to build up much information on the tail of the lag distribution. This simply reflects the fact that short panels, by their very nature, do not contain unconstrained information on that tail. However, the early coefficients are often of significant interest in themselves. Moreover, they may provide a basis for restricting the lag structure (to be a function of a small number of parameters) in further work.

11.4.4 Identification Using Prior Structure on the Lag Coefficients

In many situations we may know that all β_τ are positive. We may also know that the first few coefficients β_0, β_1, and β_2 are the largest and that β_τ decreases with τ at least after a certain value of τ. In this subsection we show how the conventional approach of constraining the lag coefficients to be a function of a finite number of parameters can be used and generalized for identification of a distributed-lag model in the panel data context. Therefore, we drop Assumption 11.4.4. Instead, we assume that we have prior knowledge of the structure of lag coefficients. The particular example we use here is the one assumed by Pakes and Griliches (1984), where the sequence of lag coefficients, after the first few free lags, has an autoregressive structure. This restriction is formalized as follows.

Assumption 11.4.5:

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

where the roots of the characteristic equation $1 - \sum_{j=1}^{J} \delta_j L^j = 0$, say, $\lambda_1^{-1}, \ldots, \lambda_J^{-1}$, lie outside the unit circle.[14] For simplicity, we assume that $k_1 = \ell + 1$, and that $\lambda_1, \ldots, \lambda_J$ are real and distinct.

Assumption 11.4.5 implies that β_τ declines geometrically after the first k_1 lags. Solving the Jth-order difference equation

$$\beta_\tau - \delta_1 \beta_{\tau-1} - \cdots - \delta_J \beta_{\tau-J} = 0, \tag{11.4.15}$$

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

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

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

$$\beta_\tau = \sum_{j=1}^{J} A_j \lambda_j^\tau, \tag{11.4.16}$$

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

Substituting (11.4.16) into (11.4.5), we write the truncation-remainder term b_{it} as

$$\begin{aligned} b_{it} &= \sum_{\tau=\ell+1}^{\infty} \left(\sum_{j=1}^{J} A_j \lambda_j^{t+\tau} \right) x_{i,-\tau} \\ &= \sum_{j=1}^{J} \lambda_j^t \left(A_j \sum_{\tau=\ell+1}^{\infty} \lambda_j^\tau x_{i,-\tau} \right) \\ &= \sum_{j=1}^{J} \lambda_j^t b_{ij}, \end{aligned} \tag{11.4.17}$$

where $b_{ij} = A_j \sum_{\tau=\ell+1}^{\infty} \lambda_j^\tau x_{i,-\tau}$. That is, we can represent the truncation remainder b_{it} in terms of J unobserved initial conditions $(b_{i1}, \dots, b_{iJ})$. Thus, under Assumptions 11.4.1–11.4.3 and 11.4.5, the distributed-lag model becomes a system of T regressions with $J + 1$ freely correlated unobserved factors $(\alpha_i^*, b_{i1}, \dots, b_{iJ})$ with J of them decaying geometrically over time.

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

When $J = 1$, it is the familiar case of a modified Koyck (or geometric) lag model. The truncation remainder becomes an unobserved factor that follows an exact first-order autoregression (i.e., $b_{it} = \delta b_{i,t-1}$). Substituting this result into (11.4.5), we have

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{\ell+1} \beta_\tau x_{i,t-\tau} + \beta_{\ell+1} \sum_{\tau=\ell+2}^{t+\ell} \delta^{\tau-(\ell+1)} x_{i,t-\tau} + \delta^{t-1} b_i + \tilde{u}_{it}, \tag{11.4.18}$$

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

Recall from the discussion in Section 11.4.3 that to identify the lag parameters we require a set of restrictions on the projection matrix $E^*(\mathbf{b}_i \mid \mathbf{x}_i) = [W + \mathbf{ca}']\mathbf{x}_i$ [equation (11.4.7)]. The Koyck lag model implies that $b_{it} = \delta b_{i,t-1}$, which implies that $E^*(b_{it} \mid \mathbf{x}_i) = \delta E^*(b_{i,t-1} \mid \mathbf{x}_i)$; that is, $w_{tr} = \delta w_{t-1,r}$ for $r = 1, \dots, T + \ell + 1$ and $t = 2, \dots, T$. It follows that the Π matrix has

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

the form

$$\Pi = B^* + \boldsymbol{\delta}^* \mathbf{w}^{*'} + \mathbf{e}\mathbf{a}', \tag{11.4.19}$$

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

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

Taking partial derivatives of (11.4.19) with respect to unknown parameters, it can be shown that the resulting Jacobian matrix satisfies the rank condition for identification of the lag coefficients, provided $T \geq 3$ (e.g., Hsiao 1983, Theorem 5.1.2). In fact, an easy way to see that the lag coefficients are identified is to note that (11.4.18) implies that

$$\begin{aligned} (y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) &= \beta_0 x_{it} + [\beta_1 - \beta_0(1+\delta)]x_{i,t-1} \\ &\quad + \sum_{\tau=2}^{\ell} [\beta_\tau - (1+\delta)\beta_{\tau-1} + \delta\beta_{\tau-2}]x_{i,t-\tau} + \nu_{it}, \\ i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{11.4.20}$$

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

In the more general case, with $J > 1$, $\boldsymbol{\delta}^*\mathbf{w}^{*'}$ in (11.4.19) will be replaced by $\sum_{j=1}^{J} \boldsymbol{\lambda}_j^* \mathbf{w}_j^{*'}$, where $\boldsymbol{\lambda}_j^{*'} = [1, \lambda_j, \ldots, \lambda_j^{T-1}]$, and $\mathbf{w}_j^*$ is the vector of coefficients from the projection of b_{ij} on $\mathbf{x}_i$. Using a similar procedure, we can show that the Π matrix will identify the lag coefficients if $T \geq J + 2$.

Of course, if in addition to Assumption 11.4.5 we also have information on the structure of x process, there will be more restrictions on the Π matrices than in the models in this subsection. Identification conditions can consequently be relaxed.

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

11.4.5 Estimation and Testing

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

$$\underset{T\times 1}{\mathbf{y}_i} = \boldsymbol{\mu}^* + [I_T \otimes \mathbf{x}_i']\boldsymbol{\pi} + \boldsymbol{\nu}_i, \quad i = 1, \ldots, N, \tag{11.4.21}$$

where $\boldsymbol{\nu}_i = \mathbf{y}_i - E^*[\mathbf{y}_i \mid \mathbf{x}_i]$, and $\boldsymbol{\pi}' = [\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_T']$, where $\boldsymbol{\pi}_j'$ is the jth row of the matrix Π. By construction, $E(\boldsymbol{\nu}_i \otimes \mathbf{x}_i) = \mathbf{0}$. Under the assumption that the N vectors $(\mathbf{y}_i', \mathbf{x}_i')$ are independent draws from a common distribution, with finite fourth-order moments and with $E\mathbf{x}_i\mathbf{x}_i' = \Sigma_{xx}$ positive definite, the least-squares estimator of $\boldsymbol{\pi}$, $\hat{\boldsymbol{\pi}}$, is consistent, and $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix Ω, which is given by (3.8.11).

The models of Sections 11.4.3 and 11.4.4 imply that $\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of the model's parameters of dimensions $m \leq (T + \ell + 1)$. We can impose these restrictions by a minimum-distance estimator that chooses $\hat{\boldsymbol{\theta}}$ to minimize

$$[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})], \tag{11.4.22}$$

where $\hat{\Omega}$ is a consistent estimator of (3.8.11). Under fairly general conditions, the estimator $\hat{\boldsymbol{\theta}}$ is consistent, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed, with asymptotic variance–covariance matrix

$$(F'\Omega^{-1}F)^{-1}, \tag{11.4.23}$$

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

$$N[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})]'\Omega^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})] \tag{11.4.24}$$

is asymptotically χ^2 distributed with $T(T + \ell + 1) - m$ degrees of freedom.

Equation (11.4.24) provides us with a test of the $T(T + \ell + 1) - m$ constraints $\mathbf{f}(\boldsymbol{\theta})$ placed on $\boldsymbol{\pi}$. To test nested restrictions, consider the null hypothesis $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ is a k-dimensional vector $(k \leq m)$ of the parameters of the restricted model. Let $\mathbf{h}(\boldsymbol{\omega}) = \mathbf{f}[\mathbf{g}(\boldsymbol{\omega})]$; that is, $\mathbf{h}$ embodies the restrictions of the constrained model. Then, under the null hypothesis,

$$N[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\omega}})]'\Omega^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\omega}})] \tag{11.4.25}$$

is asymptotically χ^2 distributed with $T(T + \ell + 1) - k$ degrees of freedom, where $\hat{\boldsymbol{\omega}}$ minimizes (11.4.25). Hence, to test the null hypothesis, we can use

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press

the statistic[15]

$$N[\hat{\pi} - \mathbf{h}(\hat{\boldsymbol{\omega}})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\omega}})] \\ - N[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})], \qquad (11.4.26)$$

which is asymptotically χ^2 distributed, with $m - k$ degrees of freedom.

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

$$E^*[y_{it} \mid G_{it}, \alpha_i^*] = \alpha_i^* + \gamma G_{it}, \qquad (11.4.27)$$

and

$$G_{it} = \sum_{\tau=1}^{\infty} \beta_{i\tau}\, x_{it-\tau}. \qquad (11.4.28)$$

Substituting (11.4.28) into (11.4.27), we have a model that consists in regressing the operating profits of firms on a distributed lag of their past investment expenditures.

Using a sample of 258 manufacturing firms' annual profit data for the years 1964–72 and investment data for the years 1961–71, and assuming that p in Assumption 11.4.4 equals three,[16] they found that the estimated lag coefficients rose over the first three periods and remained fairly constant over the next four or five. This pattern implies that the contribution of past investment to the capital stock first "appreciates" in the early years as investments are completed, shaken down, or adjusted to. This is distinctly different from the pattern implied by the commonly used straight-line or declining-balance depreciation formula to construct the "stock of capital." Both formulas imply that the lag coefficients decline monotonically in τ, with the decline being the greatest in earlier periods for the second case.

[15] See Neyman (1949) or Hsiao (1984).

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

https://doi.org/10.1017/CBO9781139839327.014 Published online by Cambridge University Press