CHAPTER 9

Cross-Sectionally Dependent Panel Data

Most panel inference procedures discussed so far assume that apart from the possible presence of individual invariant but period varying time-specific effects, the effects of omitted variables are independently distributed across cross-sectional units. Often economic theory predicts that agents take actions that lead to interdependence among themselves. For example, the prediction that risk-averse agents will make insurance contracts allowing them to smooth idiosyncratic shocks implies dependence in consumption across individuals. Cross-sectional units could also be affected by common omitted factors. The presence of cross-sectional dependence can substantially complicate statistical inference for a panel data model. However, properly exploiting the dependence across cross-sectional units in panel data not only can improve the efficiency of parameter estimates, but it may also simplify statistical inference than the situation where only cross-sectional data are available. In Section 9.1 we discuss issues of ignoring cross-sectional dependence. Sections 9.2 and 9.3 discuss spatial and factor approaches for modeling cross-sectional dependence. Section 9.4 discusses cross-sectional mean augmented approach for controlling the impact of cross-sectional dependency. Section 9.5 discusses procedures for testing cross-sectional dependence. Section 9.6 demonstrates that when panel data are cross-sectionally dependent, sometimes it may considerably simplify statistical analysis compared to the case of when only cross-sectional data are available by considering the measurement of treat effects.

9.1 ISSUES OF CROSS-SECTIONAL DEPENDENCE

Standard two-way effects models (e.g. (3.6.8)) imply observations across individuals are equal correlated. However, the impacts of common omitted factors could be different for different individuals. Unfortunately, contrary to the observations along the time dimension in which the time label, t or s, gives a natural ordering and structure, there is no natural ordering of observations along the cross-sectional dimension. The cross-sectional labeling, i or j, is arbitrary.

Let $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$ be the $N \times 1$ vector of cross-sectionally stacked error vector at time t and the $N \times N$ constant matrix, $\sum = (\sigma_{ij})$ be its

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

covariance matrix. When N is fixed and T is large, one can estimate the covariance between i and j, σ_{ij}, by $\frac{1}{T}\sum_{t=1}^{T}\hat{v}_{it}\hat{v}_{jt}$ directly, using individual time series regression residuals, $\hat{v}_{it}$ if the conditional variables, $\mathbf{x}_{it}$, are uncorrelated with v_{it}. One can then apply the feasible generalized least-squares method (FGLS) or Zellner's (1962) seemingly unrelated regression method (SUR) to estimate the slope coefficients. The FGLS or SUR estimator is consistent and asymptotically normally distributed.

When T is finite, unrestricted $\sum$ cannot be consistently estimated. However, if each row of $\sum$ only has a maximum of h_N elements that are nonzero (i.e., cross-sectional dependence is in a sense "local")[1] and $\frac{h_N}{N} \to 0$ as $N \to \infty$, panel estimators that ignore cross-sectional dependence could still be consistent and asymptotically normally distributed, although they will not be efficient. The test statistics based on the formula ignoring cross-correlations could also lead to severe size distortion (e.g., Breitung and Das 2008). On the other hand, if $\frac{h_N}{N} \to c \neq 0$ as $N \to \infty$, estimators that ignore the presence of cross-sectional dependence could be inconsistent no matter how large N is (e.g., Hsiao and Tahmiscioglu (2008), Phillips and Sul (2007)) if T is finite.[2] To see this, consider the simple regression model,

$$\begin{aligned} y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \quad & i = 1, \ldots, N \\ & t = 1, \ldots, T, \end{aligned} \tag{9.1.1}$$

where $E(v_{it} \mid \mathbf{x}_{it}) = 0$, $E(v_{it}v_{js}) = 0$ if $t \neq s$ and $E(v_{it}v_{jt}) = \sigma_{ij}$. Let $\sum = (\sigma_{ij})$ be the $N \times N$ covariance matrix of the cross-sectionally stacked error, $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$. Then the covariance matrix of pooled least-squares estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\text{LS}}$, is equal to

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{\text{LS}}) = \left[\sum_{t=1}^{T} X_t'X_t\right]^{-1}\left[\sum_{t=1}^{T} X_t'\Sigma X_t\right]\left[\sum_{t=1}^{T} X_t'X_t\right]^{-1}, \tag{9.1.2}$$

where $X_t = (\mathbf{x}_{it}')$ denotes the $N \times K$ cross-sectionally stacked explanatory variables $\mathbf{x}_{it}$ for time period t. Since $\sum$ is a symmetrical positive definite matrix, $\sum$ can be decomposed as

$$\Sigma = \vee\Lambda\vee', \tag{9.1.3}$$

where $\wedge$ is a diagonal matrix with the diagonal elements being the eigenvalues of Σ and $\vee$ is an orthonormal matrix. If one or more eigenvalues of Σ is of order N, $\mathbf{X}_t'\Sigma\mathbf{X}_t$ could be of order N^2 under the conventional assumption that $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}$ converges to a constant vector. Hence the

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{\text{LS}}) = O(\frac{1}{T}).$$

[1] This is equivalent to saying the eigenvalues of $\sum$ are bounded as $N \longrightarrow \infty$. Pesaran and Tosetti (2010) define this case as the "weak dependence."

[2] This is equivalent to saying the eigenvalues of $\sum$ are $O(N)$, which Pesaran and Tosetti (2010) called the "strong dependence."

In other words, the least-squares estimator of $\boldsymbol{\beta}$ converges to a random variable rather than a constant when T is finite and N is large.

9.2 SPATIAL APPROACH

9.2.1 Introduction

In regional science, correlation across cross-sectional units is assumed to follow a certain spatial ordering, that is, dependence among cross-sectional units is related to location and distance, in a geographic or more general economic or social network space (e.g., Anselin 1988; Anselin and Griffith 1988; Anselin, Le Gallo, Jayet 2008). The neighbor relation is expressed by a so-called spatial weights matrix, $W = (w_{ij})$, an $N \times N$ positive matrix in which the rows and columns correspond to the cross-sectional units, is specified to express the prior strength of the interaction between location i (in the row of the matrix) and location j (column), w_{ij}. By convention, the diagonal elements, $w_{ii} = 0$. The weights are often standardized so that the sum of each row, $\sum_{j=1}^{N} w_{ij} = 1$ through row-normalization; for instance, let the ith row of W, $\mathbf{w}_i' = (d_{i1}, \dots, d_{iN}) / \sum_{j=1}^{N} d_{ij}$, where $d_{ij} \geq 0$ represents a function of the spatial distance of the ith and jth units in some (characteristic) space. A side effect of this standardization is that whereas the original weights may be symmetrical, the row-standardized form no longer is.

The spatial weights matrix, W, is often included into a model specification to the dependent variable, or to the error term or to both through a so-called *spatial lag operator*. For instance, a *spatial lag* model for the $NT \times 1$ variable $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_N')'$, $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$, may take the form

$$\mathbf{y} = \rho(W \otimes I_T)\mathbf{y} + X\boldsymbol{\beta} + \mathbf{u} \tag{9.2.1}$$

where X and $\mathbf{u}$ denote the $NT \times K$ explanatory variables and $NT \times 1$ vector of error terms, respectively (called the mixed regressive, spatial autoregression model by Anselin (1988) and Ord (1975)). A *spatial error* model may take the form

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{v}, \tag{9.2.2}$$

where $\mathbf{v}$ may be specified as in a *spatial autoregressive* form,

$$\mathbf{v} = \theta(W \otimes I_T)\mathbf{v} + \mathbf{u}, \tag{9.2.3}$$

or a *spatial moving average* form,

$$\mathbf{v} = \delta(W \otimes I_T)\mathbf{u} + \mathbf{u}, \tag{9.2.4}$$

and $\mathbf{u}_i' = (u_{i1}, \dots, u_{iT})$ is assumed to be independently distributed across i with $E\mathbf{u}_i\mathbf{u}_i' = \sigma_u^2 I_T$.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

The joint determination of $\mathbf{y}$ for model (9.2.1) or $\mathbf{v}$ for (9.2.2) when $\mathbf{v}$ is given by (9.2.3) is through $[(I_N - \rho W)^{-1} \otimes I_T]$ or $[(I_N - \theta W)^{-1} \otimes I_T]$. Since

$$(I_N - \rho W)^{-1} = I_N + \rho W + \rho^2 W^2 + \cdots, \tag{9.2.5}$$

or

$$(I_N - \theta W)^{-1} = I_N + \theta W + \theta^2 W^2 + \cdots, \tag{9.2.6}$$

to ensure a "distance" decaying effect among the cross-sectional units, ρ and θ are assumed to have absolute values less than 1.[3]

The *spatial autoregressive* form (9.2.3) implies that the covariance matrix of the N cross-sectional units at time t, $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$ takes the form

$$E\mathbf{v}_t\mathbf{v}_t' = \sigma_u^2[I_N - \theta W]^{-1}[I_N - \theta W']^{-1} = V. \tag{9.2.7}$$

The *spatial moving average* form (9.2.4) implies that the covariance matrix of $\mathbf{v}_t$ takes the form

$$\begin{aligned} E\mathbf{v}_t\mathbf{v}_t' &= \sigma_u^2[I_N + \delta W][I_N + \delta W]' \\ &= \sigma_u^2[I_N + \delta(W + W') + \delta^2 WW'] = \tilde{V}. \end{aligned} \tag{9.2.8}$$

When W is *sparse*, that is, many elements of W are prespecified to be 0, for instance, W could be a block diagonal matrix in which only observations in the same region are considered *neighbors*, and observations across regions are uncorrelated. W can also be a sparse matrix by some neighboring specification, for example, if a district is a spatial unit, some specifications assume that a neighbor for this district is another one which has a common boundary. The spatial moving average form allows the cross-correlations to be "local" (9.2.8). On the other hand, the *spatial autoregressive* form suggests a much wider range of spatial covariance than specified by the nonzero elements of the weights matrix W, implying a "global" covariance structure (9.2.7).

Generalizing the spatial approach, Conley (1999) suggests using the notion of "economic distance" to model proximity between two economic agents. The joint distribution of random variables at a set of points is assumed to be invariant to a shift in location and is a function only of the "economic distances" between them. For instance, the population of individuals is assumed to reside in a low dimensional Euclidean space, say R^2, with each individual i located at a point s_i. The sample then consists of realization of agents' random variables at a collection of locations $\{s_i\}$ inside a sample region. If two agents' locations s_i and s_j are close, then y_{it} and y_{js} may be highly correlated. As the distance

[3] The combination of the row sum of W equal to 1 and γ or θ having absolute value less than 1 implies that the spatial models assume cross-sectional dependence being "weak."

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

between s_i and s_j grows large, y_{it} and y_{js} approach independence. Under this assumption, the dependence among cross-sectional data can be estimated using methods analogous to time series procedures either parametrically or nonparametrically (e.g., Hall, Fisher, and Hoffman 1992; Priestley 1982; Newey and West 1987).

While the approach of defining cross-sectional dependence in terms of "economic distance" measure allows for more complicated dependence than models with time-specific (or group-specific) effects alone (e.g. Chapter 3, Section 3.6), it still requires that the econometricians have information regarding this "economic distance." In certain urban, environmental, development, growth, and other areas of economics, this information may be available. For instance, in the investigation of peoples' willingness to pay for local public goods, the relevant economic distance may be the time and monetary cost of traveling between points to use these local public goods. Alternatively, if the amenity is air quality, then local weather conditions might constitute the major unobservable common to cross-sectional units in the neighborhood. Other examples include studies of risk sharing in rural developing economies where the primary shocks to individuals in such agrarian economies may be weather related. If so, measures of weather correlation on farms of two individuals could be the proxy for the economic distance between them. In many other situations, prior information such as this may be difficult to come by. However, the combination of the row sum of W equal to 1 and δ or θ having absolute value less than 1 implies that the population consists of N cross-sectional units. In other words, the spatial approach is an analysis of the population based on T time series realized observations. If N is considered sample size, then the spatial autoregressive model implies that the cross-sectional dependence is "weak." In other words, each cross-sectional unit is correlated only with a fixed number of other cross-sectional units. Under an assumpton of weak cross-sectional dependence, the covariance estimator (3.2.8) for models with only individual-specific effects or (3.6.13) for models with both individual- and time-specific effects of $\boldsymbol{\beta}$ remains consistent if T is fixed and $N \to \infty$ or if N is fixed and T tends to infinity or both. However, there could be severe size distortion in hypothesis testing if cross-sectional dependence is ignored. Vogelsang (2012) showed that the covariance matrix estimate proposed by Driscoll and Kraay (1998) based on the Newey–West (1987) heteroscedastic autocorrelation (HAC) covariance matrix estimator of cross-sectional averages,

$$T\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)^{-1}\hat{\bar{\Omega}}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)^{-1}, \tag{9.2.9}$$

is robust to heteroscedasticity, autocorrelation and spatial dependence, where $\tilde{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ if (3.2.8) is used or $\tilde{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\mathbf{x}})$ if (3.6.13) is used, $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}$, $\bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}$, $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{x}}_t$, and

$$\hat{\hat{\Omega}} = \frac{1}{T}\left\{\sum_{t=1}^{T}\hat{\hat{\mathbf{v}}}_t^*\hat{\hat{\mathbf{v}}}_t^{*\prime} + \sum_{j=1}^{T-1}k\left(\frac{j}{m}\right)\left[\sum_{t=j+1}^{T}\hat{\hat{\mathbf{v}}}_t^*\hat{\hat{\mathbf{v}}}_{t-j}^{*\prime} + \sum_{t=j+1}^{T}\hat{\hat{\mathbf{v}}}_{t-j}^*\hat{\hat{\mathbf{v}}}_t^{*\prime}\right]\right\}$$

$$\hat{\mathbf{v}}_{it}^* = \tilde{\mathbf{x}}_{it}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\hat{\boldsymbol{\beta}}_{cv}), \tag{9.2.10}$$

$$\hat{\hat{\mathbf{v}}}_t^* = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{it}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}\hat{\boldsymbol{\beta}}_{cv}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{v}}_{it}^*,$$

$\tilde{y}_{it} = (y_{it} - \bar{y}_i)$ if (3.2.8) is used or $\tilde{y}_{it} = (y_i - \bar{y}_i - \bar{y}_t + \bar{y})$ if (3.6.13) is used, and $k(\frac{j}{m}) = 1 - |\frac{j}{m}|$ if $| \frac{j}{m} | < 1$ and $k(\frac{j}{m}) = 0$ if $| \frac{j}{m} | > 1$, m is an a priori chosen positive constant less than or equal to T. The choice of m depends on how strongly an investigator thinks about the serial correlation of the error u_{it}.

9.2.2 Spatial Error Model

The log-likelihood function for the spatial error model (9.2.2) takes the form

$$-\frac{1}{2}\log \mid \Omega \mid -\frac{1}{2}\mathbf{v}'\Omega^{-1}\mathbf{v}, \tag{9.2.11}$$

where

$$\Omega = V \otimes I_T \tag{9.2.12}$$

if $\mathbf{v}$ is a spatial autoregressive form (9.2.3), and

$$\Omega = \tilde{V} \otimes I_T \tag{9.2.13}$$

is $\mathbf{v}$ is a spatial moving average form (9.2.4). Conditional on θ or δ, the MLE of $\boldsymbol{\beta}$ is just the GLS estimator

$$\hat{\boldsymbol{\beta}} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}\mathbf{y}). \tag{9.2.14}$$

When Ω takes the form of (9.2.12), the log-likelihood function (9.2.11) takes the form

$$\begin{aligned} &T \log \mid I_N - \theta W \mid -\frac{NT}{2}\log\sigma_u^2 \\ &-\frac{1}{2\sigma_u^2}(\mathbf{y} - X\boldsymbol{\beta})'[(I_N - \theta W)'(I_N - \theta W) \otimes I_T](\mathbf{y} - X\boldsymbol{\beta}). \end{aligned} \tag{9.2.15}$$

Ord (1975) notes that

$$\mid I_N - \theta W \mid = \prod_{j=1}^{N}(1 - \theta\omega_j), \tag{9.2.16}$$

where ω_j are the eigenvalues of W, which are real even W after row normalization is no longer symmetric. Substituting (9.2.16) into (9.2.15), the

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

log-likelihood values can be evaluated at each possible $(\theta, \boldsymbol{\beta}')$ with an iterative optimization routine. However, when N is large, the computation of the eigenvalues becomes numerically unstable.

9.2.3 Spatial Lag Model

For the spatial lag model (9.2.1), the right-hand side, $(W \otimes I_T)\mathbf{y}$, and $\mathbf{u}$ are correlated. When $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 I_{NT})$, the log-likelihood function of (9.2.1) is

$$\begin{aligned} &T \log \mid I_N - \rho W \mid - \frac{NT}{2} \log \sigma_u^2 \\ &- \frac{1}{2\sigma_u^2}[\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta}]'[\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta}], \mid \rho \mid < 1. \end{aligned} \tag{9.2.17}$$

When T is fixed, the MLE is $\sqrt{N}$ consistent and asymptotically normally distributed under the assumption that w_{ij} are at most of order h_N^{-1}, and the ratio $h_N/N \to 0$ as N goes to infinity (Lee (2004)). However, when N is large, just like the MLE for (9.2.11), the MLE for (9.2.1) is burdensome and numerically unstable (e.g., Kelejian and Prucha (2001), Lee (2004)). The $\mid I_N - \rho W \mid$ is similar in form to (9.2.16). A similar iterative optimization routine as that for (9.2.15) can be evaluated at each possible $(\rho, \boldsymbol{\beta}')$. When N is large, the computation of the eigenvalues becomes numerically unstable.

The parameters $(\rho, \boldsymbol{\beta}')$ can also be estimated by the instrumental variables or generalized method of moments estimator (or two-stage least squares estimator) (Kelejian and Prucha 2001),

$$\begin{pmatrix} \hat{\rho} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = [Z'H(H'H)^{-1}H'Z]^{-1}[Z'H(H'H)^{-1}H'\mathbf{y}], \tag{9.2.18}$$

where $Z = [(W \otimes I_T)\mathbf{y}, X]$ and $H = [(W \otimes I_T)X, X]$. Lee (2003) shows that an optimal instrumental variables estimator is to let $H = [(W \otimes I_T)E\mathbf{y}, X]$, where $E\mathbf{y} = [I_{NT} - \rho(W \otimes I_T)]^{-1}X\boldsymbol{\beta}$. The construction of optimal instrumental variables requires some initial consistent estimators of ρ and $\boldsymbol{\beta}$.

When $w_{ij} = O(N^{-(\frac{1}{2}+\delta)})$, where $\delta > 0$, $E((W \otimes I_T)\mathbf{y}\mathbf{u}') = o(N^{-\frac{1}{2}})$, one can ignore the correlations between $(W \otimes I_T)\mathbf{y}$ and $\mathbf{u}$. Applying the least-squares method to (9.2.1) yields a consistent and asymptotically normally distributed estimator of $(\rho, \boldsymbol{\beta}')$ (Lee 2002). However, if W is "sparse," this condition may not be satisfied. For instance, in Case (1991), "neighbors" refers to households in the same district. Each neighbor is given equal weight. Suppose there are r districts and m members in each district, $N = mr$. Then $w_{ij} = \frac{1}{m}$ if i and j are in the same district and $w_{ij} = 0$ if i and j belong to different districts. If $r \longrightarrow \infty$ as $N \longrightarrow \infty$ and N is relatively much larger than r in the sample, one might regard the condition $w_{ij} = O(N^{-(\frac{1}{2}+\delta)})$ being satisfied. On the other hand, if r is relatively much larger than m or $\lim_{N\to\infty} \frac{r}{m} = c \neq 0$, then $w_{ij} = O(N^{-\frac{1}{2}(N+\delta)})$ cannot hold.

9.2.4 Spatial Error Models with Individual-Specific Effects

One can also combine the spatial approach with the error components or fixed effects specification (e.g., Kapoor, Kelejian, and Prucha 2007; Lee and Yu (2010a,b)). For instance, one may generalize the spatial error model by adding the individual-specific effects,

$$\mathbf{y} = X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v}, \tag{9.2.19}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$. Suppose $\boldsymbol{\alpha}$ are treated as fixed constants and $\mathbf{v}$ follows a spatial error autoregressive form (9.2.3), the log-likelihood function is of the form (9.2.11) where Ω is given by (9.2.12), and $\mathbf{v} = (\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha})$. Taking partial derivatives of the log-likelihood function with respect to $\boldsymbol{\alpha}$ and setting it equal to $\mathbf{0}$ yields the MLE estimates of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$ and θ. Substituting the MLE estimates of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$ and θ into the log-likelihood function, we obtain the concentrated log-likelihood function

$$\begin{aligned} -\frac{NT}{2} \log \sigma_u^2 + T \log \mid I_N - \theta W \mid \\ - \frac{1}{2\sigma_u^2} \tilde{\mathbf{v}}'[(I_N - \theta W)'(I_N - \theta W) \otimes I_T]\tilde{\mathbf{v}}, \end{aligned} \tag{9.2.20}$$

where the element $\tilde{\mathbf{v}}, \tilde{v}_{it} = (y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta}$, $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, and $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$. In other words, the MLE of $\boldsymbol{\beta}$ is equivalent to first taking the covariance transformation of each y_{it} and $\mathbf{x}_{it}$ to get rid of the individual-specific effects, α_i, then maximizing (9.2.20) to obtain the MLE of the spatial error model with fixed individual-specific effects.

The MLE of $\boldsymbol{\beta}$ and θ are consistent when either N or T or both tend to infinity. However, the MLE of $\boldsymbol{\alpha}$ and σ_u^2 is consistent only if $T \longrightarrow \infty$. To obtain consistent estimate of $(\beta, \theta, \sigma_u^2)$ with finite T and large N, Lee and Yu (2010a,b) suggest maximizing[4]

$$\begin{aligned} -\frac{N(T-1)}{2} \log \sigma_u^2 + (T-1) \log \mid I_N - \theta W \mid \\ -\frac{1}{2\sigma_u^2} \tilde{\mathbf{v}}'[(I_N - \theta W)'(I_N - \theta W) \otimes I_T]\tilde{\mathbf{v}}. \end{aligned} \tag{9.2.21}$$

When α_i are treated as random and are independent of $\mathbf{u}$, The $NT \times NT$ covariance matrix of $\mathbf{v}$ takes the form

$$\Omega = \sigma_\alpha^2 (I_N \otimes J_T) + \sigma_u^2((B'B)^{-1} \otimes I_T), \tag{9.2.22}$$

if α_i and u_{it} are independent of X and are i.i.d. with mean 0 and variance σ_α^2 and σ_u^2, respectively, where J_T is a $T \times T$ matrix with all elements equal to 1,

[4] As a matter of fact, (9.2.21) is derived by the transformation matrix Q^* where $Q^* = [F, \frac{1}{\sqrt{T}} I_T]$, where F is the $T \times (T-1)$ eigenvector matrix of $Q = I_T - \frac{1}{T}\mathbf{e}_T\mathbf{e}_T'$ that correspond to the eigenvalues of 1.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

$B = (I_N - \theta W)$. Using the results in Wansbeek and Kapteyn (1978), one can show that (e.g., Baltagi et al. 2007)

$$\Omega^{-1} = \sigma_u^{-2}\left\{\frac{1}{T}J_T \otimes \left[T\phi I_N + (B'B)^{-1}\right]^{-1} + E_T \otimes B'B\right\}, \tag{9.2.23}$$

where $E_T = I_T - \frac{1}{T}J$ and $\phi = \frac{\sigma_\alpha^2}{\sigma_u^2}$,

$$\mid \Omega \mid = \sigma_u^{2NT} \mid T\phi I_N + (B'B)^{-1} \mid \cdot \mid (B'B)^{-1} \mid^{T-1} . \tag{9.2.24}$$

The MLE of $\boldsymbol{\beta}, \theta, \sigma_u^2$, and σ_α^2 can then be derived by substituting (9.2.23) and (9.2.24) into the log-likelihood function (e.g., Anselin 1988, p. 154).

The FGLS estimator (9.2.14) of the random-effects spatial error model $\boldsymbol{\beta}$ is to substitute initial consistent estimates of ϕ and θ into (9.2.23). Kapoor et al. (2007) propose a method of moments estimation with moment conditions in terms of $(\theta, \sigma_u^2, \tilde{\sigma}^2 = \sigma_u^2 + T\sigma_\alpha^2)$.

9.2.5 Spatial Lag Model with Individual-Specific Effects

For the spatial lag model with individual-specific effects,

$$\mathbf{y} = \rho(W \otimes I_T)\mathbf{y} + X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{u}. \tag{9.2.25}$$

If $\boldsymbol{\alpha}$ is treated as fixed constants, the log-likelihood function of (9.2.25) is of similar form as (9.2.20)

$$\begin{aligned} &T \log \mid I_N - \rho W \mid - \frac{NT}{2} \log \sigma_u^2 \\ &- \frac{1}{2\sigma_u^2}\{[\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}]' \\ &\qquad [\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}]\}. \end{aligned} \tag{9.2.26}$$

The MLE of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be computed similarly as that of (9.2.20).

When α_i are treated as randomly distributed across i with constant variance σ_α^2 and independent of X, then

$$\begin{aligned} &E\left\{(\mathbf{u} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha})(\mathbf{u} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha})'\right\} \\ &= I_N \otimes V^*, \end{aligned} \tag{9.2.27}$$

where $V^* = \sigma_u^2 I_T + \sigma_\alpha^2 \mathbf{e}_T \mathbf{e}_T'$. The MLE or quasi-MLE for the spatial lag model (9.2.1) can be obtained by maximizing

$$\begin{aligned} &T \log \mid I_N - \rho W \mid - \frac{N(T-1)}{2} \log \sigma_u^2 - \frac{N}{2} \log\left(\sigma_u^2 + T\sigma_\alpha^2\right) \\ &\qquad - \frac{1}{2}(\mathbf{y}^* - X\boldsymbol{\beta})'(I_N \otimes V^{*-1})(\mathbf{y}^* - X\boldsymbol{\beta}), \end{aligned} \tag{9.2.28}$$

where $\mathbf{y}^* = (I_{NT} - \rho(W \otimes I_T)\mathbf{y}$. Conditional on ρ, σ_u^2, and σ_α^2, the MLE of $\boldsymbol{\beta}$ is the GLS estimator

$$\hat{\boldsymbol{\beta}} = (X'[I_N \otimes V^{*-1}]X)^{-1}(X'(I_N \otimes V^{*-1})(I_{NT} - \rho(W \otimes I_T))\mathbf{y}, \quad (9.2.29)$$

where V^{*-1} is given by (3.3.7). Kapoor et al. (2007) have provided moment conditions to obtain initial consistent estimates σ_u^2, σ_α^2, and ρ.

One can also combine the random individual-specific effects specification of $\boldsymbol{\alpha}$ with a spatial specification for the error $\mathbf{v}$. For instance, we can let

$$\mathbf{y} = \rho(W_1 \otimes I_T)\mathbf{y} + X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v}, \quad (9.2.30)$$

with

$$\mathbf{v} = \theta(W_2 \otimes I_T)\mathbf{v} + \mathbf{u}, \quad (9.2.31)$$

where W_1 and W_2 are $N \times N$ spatial weights matrices and $\boldsymbol{\alpha}$ is an $N \times 1$ vector of individual effects. Let $S(\rho) = I_N - \rho W_1$ and $R(\theta) = I_N - \theta W_2$. Under the assumption that u_{it} is independently normally distributed, the log-likelihood function of (9.2.30) takes the form

$$\begin{aligned}\log L = & -\frac{NT}{2}\log\sigma_u^2 + T\log \mid S(\rho)\mid + T\log \mid R(\theta)\mid \\ & -\frac{1}{2}\tilde{\mathbf{v}}^{*'}\tilde{\mathbf{v}}^*,\end{aligned} \quad (9.2.32)$$

where

$$\tilde{\mathbf{v}}^* = [R(\theta) \otimes I_T][(S(\rho) \otimes I_T)\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}]. \quad (9.2.33)$$

The MLE (or quasi-MLE if u is not normally distributed) can be computed similarly as that of (9.2.20). For details, see Lee and Yu (2010a,b).

9.2.6 Spatial Dynamic Panel Data Models

Consider a dynamic panel data model of the form

$$\mathbf{y} = \mathbf{y}_{-1}\gamma + X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v} \quad (9.2.34)$$

where $\mathbf{y}_{-1}$ denotes the $NT \times 1$ vector of y_{it} lagged by one period, $\mathbf{y}_{-1} = (y_{10}, \dots, y_{1,T-1}, \dots, y_{N,T-1})'$, X denotes the $NT \times K$ matrix of exogenous variables, $X = (\mathbf{x}_{it}')$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)'$ denotes the $N \times 1$ fixed individual-specific effects. If the error term follows a spatial autoregressive form of (9.2.3), even $\mid \gamma \mid < 1$, there could be spatial cointegration if $\gamma + \theta = 1$ while $\gamma \neq 1$ (Yu and Lee (2010)). Yu et al. (2012) show that the MLE of $(\gamma, \theta, \boldsymbol{\beta}, \boldsymbol{\alpha})$ are $\sqrt{NT}$ consistent with T tends to infinity. However, if $\gamma + \theta = 1$, then the asymptotic covariance matrix of the MLE is singular when the estimator is multiplied by the scale factor $\sqrt{NT}$ because the sum of the spatial and dynamic effects converge at a higher rate (e.g., Yu and Lee (2010)).

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Yu et al. (2012) also consider the estimation of a dynamic spatial lag model with the spatial-time effect,

$$\begin{aligned}\mathbf{y} =&(\rho W \otimes I_T)\mathbf{y} + \mathbf{y}_{-1}\gamma + (\rho^* W \otimes I_T)\mathbf{y}_{-1} + X\boldsymbol{\beta} \\ &+ (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v}\end{aligned} \tag{9.2.35}$$

Model (9.2.35) is stable if $\gamma + \rho + \rho^* < 1$ and spatially cointegrated if $\gamma + \rho + \rho^* = 1$ but $\gamma \neq 1$. They develop the asymptotics of (quasi)–MLE when both N and T are large and propose a bias correction formula.

9.3 FACTOR APPROACH

Another approach to model cross-sectional dependence is to assume that the error follows a linear factor model,

$$v_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + u_{it} = \mathbf{b}_i' \, \mathbf{f}_t + u_{it}, \tag{9.3.1}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of random factors with mean 0, $\mathbf{b}_i = (b_{i1}, \ldots, b_{ir})'$ is a $r \times 1$ nonrandom factor loading coefficient (to avoid with u_{it} nonseparability), u_{it} represents the effects of idiosyncratic shocks, which is independent of $\mathbf{f}_t$ and is independently distributed across i with constant variance over t.

Factor models have been suggested as an effective way of synthesizing information contained in large data sets (e.g., Bai 2003, 2009; Bai and Ng 2002). The conventional time-specific effects model (e.g., Chapter 3) is a special case of (9.3.1) when $r = 1$ and $b_i = b_\ell$ for all i and ℓ. An advantage of factor model over the spatial approach is that there is no need to prespecify the strength of correlations between units i and j.

Let $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$, then

$$\mathbf{v}_t = B \, \mathbf{f}_t + \mathbf{u}_t, \tag{9.3.2}$$

where $B = (b_{ij})$ is the $N \times r$ factor loading matrix, and $\mathbf{u}_t = (\mathbf{u}_{1t}, \ldots, \mathbf{u}_{Nt})'$. Then

$$E\mathbf{v}_t\mathbf{v}_t' = B(E\, \mathbf{f}_t \, \mathbf{f}_t')B' + D \tag{9.3.3}$$

where D is an $N \times N$ diagonal covariance matrix of $\mathbf{u}_t$. The covariance between v_{it} and $v_{\ell t}$ is given by

$$E v_{it} v_{\ell t} = \mathbf{b}_i'(E\, \mathbf{f}_t \, \mathbf{f}_t')\mathbf{b}_\ell \tag{9.3.4}$$

However, $B\, \mathbf{f}_t = BAA^{-1}\, \mathbf{f}_t$ for any $r \times r$ nonsingular matrix A. That is, without r^2 normalizations, (or prior restrictions) $\mathbf{b}_i$ and $\mathbf{f}_t$ are not uniquely determined. A common normalization is to assume $E\, \mathbf{f}_t \, \mathbf{f}_t' = I_r$. Nevertheless, even with this assumption, it only yields $\frac{r(r+1)}{2}$ restrictions on B. B is only identifiable up to an orthonormal transformation, that is, $BCC'B' = BB'$ for any $r \times r$

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

orthonormal matrix (e.g., Anderson (1985)). Additional $\frac{r(r-1)}{2}$ restrictions are needed, say $B'B$ diagonal.[5]

For given v_{it} and r, B and F can be estimated by minimizing

$$\begin{aligned}\tilde{V}(r) &= (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(v_{it} - \mathbf{b}_i'\,\mathbf{f}_t)^2 \\ &= (NT)^{-1}tr[(V - FB')(V' - BF')],\end{aligned} \tag{9.3.5}$$

where $V = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ is a $T \times N$ matrix with $\mathbf{v}_i = (v_{i1}, \ldots, v_{iT})'$, $F = (\mathbf{f}_1, \ldots, \mathbf{f}_r)$ is a $T \times r$ matrix with $\mathbf{f}_j = (f_{j1}, \ldots, f_{jT})'$. Taking partial derivatives of (9.3.5) with respect to B, setting them equal to 0, and using the normalization $\frac{1}{T}F'F = I_r$, we obtain

$$B = \frac{1}{T}V'F. \tag{9.3.6}$$

Substituting (9.3.6) into (9.3.5), minimizing (9.3.5) is equivalent to maximizing $tr[F'(VV')F]$ subject to $B'B$ being diagonal. Therefore, the $T \times r$ common factor $F = (\mathbf{f}_1, \ldots, \mathbf{f}_r)$ is estimated as $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix $\sum_{i=1}^{N}\mathbf{v}_i\mathbf{v}_i'$, denoted by $\hat{F}$ Anderson (1985)). Given $\hat{F}$, the factor loading matrix B can be estimated as $\hat{B} = \frac{1}{T}V'\hat{F}$.

To identify r, Bai and Ng (2002) note that if $\lim_{N\to\infty}\frac{1}{N}B'B$ converges to a nonsingular $r \times r$ constant matrix A, the largest r eigenvalues of (9.3.3) are of order N because the r positive eigenvalues $B'B$ are equal to those of BB'. In other words, the r common factors, $\mathbf{f}_t$, practically drive all $N \times 1$ errors, $\mathbf{v}_t$. Therefore, when r is unknown, under the assumption that $\lim_{N\to\infty}\frac{1}{N}B'B = A$, Bai and Ng (2002) suggest using the criterion

$$\min_{k} \text{PC}(k) = \hat{V}(k) + kg(N,T), \tag{9.3.7}$$

or

$$\min_{k} \text{IC}(k) = ln\hat{V}(k) + kg(N,T),, \tag{9.3.8}$$

where $k <\ \min(N, T)$,

$$\hat{V}(k) = \min_{B^k,\, F^k}\quad (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{v}_{it} - \hat{\mathbf{b}}_i^{k\prime}\,\hat{\mathbf{f}}_t^k)^2, \tag{9.3.9}$$

[5] Even in this case uniqueness is only up to a sign change. For instance, $-\mathbf{f}_t$ and $-B$ also satisfy the restrictions. However, the covariance between v_{it} and v_{jt} remains the same, $E(v_{it}v_{jt}) = \mathbf{b}_i'\mathbf{b}_j = \mathbf{b}_i^{*\prime}\mathbf{b}_j^* = \mathbf{b}_i'CC'\mathbf{b}_j$ for any $r \times r$ orthonormal matrix.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

where $\hat{v}_{it}$ is the estimated v_{it}, $\hat{\mathbf{f}}_t^k$ denotes the k-dimensional estimated factor at time t, $\hat{\mathbf{b}}_i^k$ denotes the estimated loading factor for the ith individual, and $g(N, T)$ is a penalty function satisfying (1) $g(N, T) \longrightarrow 0$ and (2) $\min\{N, T\}g(N, T) \longrightarrow \infty$ as $N, T \longrightarrow \infty$ to select r. The reason for using these criteria is because $\hat{V}(k)$ decreases with k and $\hat{V}(k) - \hat{V}(r)$ converges to a nonzero positive number for $k < r$ and $\hat{V}(k) - \hat{V}(r) \longrightarrow 0$ at certain rate, say $C(N, T)$. Choosing a penalty function $kg(N, T)$ increases with k. When $g(N, T)$ diminishes to 0 at a rate slower than $C(N, T)$, the penalty will eventually become dominant and prevent choosing a $k > r$. Bai and Ng (2002) show that $C(N, T) = \min\{N, T\}$ when u_{it} satisfy the stationarity assumption (allowing for weak serial and cross-sectional dependence). Therefore, they propose the specific forms of $g(N, T)$ as $\hat{\sigma}_u^2 \cdot \frac{N+T}{NT} ln\left(\frac{NT}{N+T}\right)$ or $\hat{\sigma}_u^2 \cdot \frac{N+T}{NT} ln\left(\min\left(N, T\right)\right)$, etc.[6] They show that when both N and $T \longrightarrow \infty$, the criterion (9.3.8) or (9.3.7) selects $\hat{k} \longrightarrow r$ with probability 1. Moreover, $\hat{\mathbf{f}}_t \longrightarrow \mathbf{f}_t$ if $(\sqrt{T}/N) \longrightarrow \infty$.

To estimate a regression model of the form

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{9.3.10}$$

where v_{it} follows (9.3.1), Bai (2009) and Pesaran (2006) suggest the least-squares regression of the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{b}_i'\,\mathbf{f}_t + u_{it}, \tag{9.3.11}$$

subject to $\frac{1}{T}\sum_{t=1}^T \mathbf{f}_t\,\mathbf{f}_t' = I_r$ and $B'B$ being diagonal. Noting that conditional on $\mathbf{f}_t$, the least squares estimator of $\boldsymbol{\beta}$ is equal to

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^N X_i'MX_i)^{-1}(\sum_{i=1}^N X_i'M\mathbf{y}_i), \tag{9.3.12}$$

where $\mathbf{y}_i$ and X_i denote the stacked T time series observations of y_{it} and $\mathbf{x}_{it}'$, $M = I - F(F'F)^{-1}F'$ where F is the $T \times r$ matrix of $F = (\mathbf{f}_t')$. Conditional on $\boldsymbol{\beta}$, the residual v_{it} is a pure factor structure (9.3.1). The least squares estimator of F is equal to the first r eigenvectors (multiplied by $\sqrt{T}$ due to the restriction $F^{r'}F^r/T = I$) associated with the largest r eigenvalues of the matrix (Anderson (1985)),

$$\frac{1}{N}\sum_{i=1}^N(\mathbf{y}_i - X_i\boldsymbol{\beta})(\mathbf{y}_i - X_i\boldsymbol{\beta})', \tag{9.3.13}$$

$$\left[\frac{1}{NT}\sum_{i=1}^N(\mathbf{y}_i - X_i\boldsymbol{\beta})(\mathbf{y}_i - X_i\boldsymbol{\beta})'\right]\hat{F} = \hat{F}\Lambda, \tag{9.3.14}$$

[6] When u_{it} exhibit considerable serial correlation and the sample size is not sufficiently large, the Bai and Ng (2002) criterion may overfit (e.g.,Greenaway-McGrevy, Han, and Sul 2012).

where Λ is a diagonal matrix that consists of the r largest eigenvalues of (9.3.13) multiplied by $\frac{1}{T}$. Conditional on $(\hat{\boldsymbol{\beta}}, \hat{F})$, (9.3.6) leads to

$$\hat{B}' = T^{-1}\left[\hat{F}'(\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}), \ldots, \hat{F}'(\mathbf{y}_N - X_N\hat{\boldsymbol{\beta}})\right] = \frac{1}{T}\hat{V}'\hat{F}. \quad (9.3.15)$$

Iterating between (9.3.12), (9.3.14), and (9.3.15) leads to the least-squares estimator of (9.3.11) as if $\mathbf{f}$ were observable.

When both N and T are large, the least-squares estimator (9.3.12) is consistent and asymptotically normally distributed with covariance matrix

$$\sigma_u^2\left(\sum_{i=1}^{N} X_i' M X_i\right)^{-1} \qquad (9.3.16)$$

if u_{it} is independently, identically distributed with mean 0 and constant variance σ_u^2. However, if u_{it} is heteroscedastic and cross-sectionally or serially correlated, when $\frac{T}{N} \to c \neq 0$ as $N, T \to \infty$, $\sqrt{NT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically biased of the form

$$\left(\frac{T}{N}\right)^{\frac{1}{2}} C + \left(\frac{N}{T}\right)^{\frac{1}{2}} D^*, \qquad (9.3.17)$$

where C denotes the bias induced by heteroscadasticity and cross-sectional correlation and D^* denotes the bias induced by serial correlation and heteroscedasticity of u_{it}. Bai (2009) has provided the formulas for constructing the bias-corrected estimator.

To estimate a model with both additive and interactive effects,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \mathbf{b}_i'\,\mathbf{f}_t + u_{it}, \qquad (9.3.18)$$

in addition to the normalization conditions, $F'F = I_r$ and $B'B$ diagonal, we also need to impose the restriction $\sum_{i=1}^{N}\alpha_i = 0$, $\sum_{i=1}^{N}\mathbf{b}_i = \mathbf{0}$, $\sum_{t=1}^{T}\mathbf{f}_t = \mathbf{0}$, to obtain a unique solution of $(\beta, \alpha_i, \mathbf{b}_i, \mathbf{f}_t)$ (Bai 2009, p. 1253). Just like the standard fixed-effects estimator (Chapter 3) we can first take individual observations from its time series mean, $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$, to get rid of α_i from (9.3.18), and then iteratively estimate $\boldsymbol{\beta}$ and F by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{M}\tilde{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\tilde{X}_i\tilde{M}\tilde{\mathbf{y}}_i\right), \qquad (9.3.19)$$

where $\tilde{\mathbf{y}}_i, \tilde{X}_i$ denote the stacked T time series observations of $\tilde{y}_{it}$ and $\tilde{\mathbf{x}}_{it}'$, $\tilde{M} = I - \hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}'$, and $\hat{F}$ is the $T \times r$ matrix consisting of the first r eigenvectors (multiplied by $\sqrt{T}$) associated with the r largest eigenvalues of the matrix

$$\frac{1}{NT}\sum_{i=1}^{N}(\tilde{\mathbf{y}}_i - \tilde{X}_i\hat{\boldsymbol{\beta}})(\tilde{\mathbf{y}}_i - \tilde{X}_i\hat{\boldsymbol{\beta}})'. \qquad (9.3.20)$$

After convergent solutions of $\hat{\boldsymbol{\beta}}$ and $\hat{F}$ are obtained, one can obtain $\hat{\alpha}_i$ and $\hat{B}'$ by

$$\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}, \tag{9.3.21}$$

$$\hat{B}' = T^{-1}[\hat{F}'(\tilde{\mathbf{y}}_1 - \tilde{X}_1\hat{\boldsymbol{\beta}}), \dots, \hat{F}'(\tilde{\mathbf{y}}_N - \tilde{X}_N\hat{\boldsymbol{\beta}})]. \tag{9.3.22}$$

Ahn, Lee, and Schmidt (2001, 2013) have proposed a nonlinear GMM method to estimate a linear panel data model with interactive effects (9.3.1). For ease of exposition, suppose $r = 1$. Let $\theta_t = \frac{f_t}{f_{t-1}}$, then

$$(y_{it} - \theta_t y_{i,t-1}) = \mathbf{x}_{it}'\boldsymbol{\beta} - \mathbf{x}_{i,t-1}'\boldsymbol{\beta}\theta_t + (u_{it} - \theta_t u_{i,t-1}),\ t = 2, \dots, T. \tag{9.3.23}$$

It follows that

$$E[\mathbf{x}_i(u_{it} - \theta_t u_{i,t-1})] = \mathbf{0} \tag{9.3.24}$$

Let $W_i = I_{T-1} \otimes \mathbf{x}_i$,

$$\underset{(T-1)\times(T-1)}{\ominus =} \begin{bmatrix} \theta_2 & 0 & \dots & \dots & 0 \\ 0 & \theta_3 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \theta_T \end{bmatrix},$$

$$\tilde{\mathbf{u}}_i = (u_{i2}, \dots, u_{iT})',\ \tilde{\mathbf{u}}_{i,-1} = (u_{i1}, \dots, u_{i,T-1})'.$$

Then a GMM estimator of $\boldsymbol{\beta}$ and $\ominus$ can be obtained from the moment conditions,

$$E[W_i(\tilde{\mathbf{u}}_i - \ominus\tilde{\mathbf{u}}_{i,-1})] = \mathbf{0}. \tag{9.3.25}$$

The nonlinear GMM estimator is consistent and asymptotically normally distributed when $N \to \infty$ under fixed T even u_{it} is serially correlated and heteroscedastic. However, the computation can be very cumbersome when $r > 1$. For instance, if $r = 2$, in addition to letting $\theta_t = \frac{f_{it}}{f_{i,t-1}}$, we need to introduce additional parameters $\delta_t = f_{2t} - f_{2,t-1}\theta_t$ and to take the quasi-difference of $(y_{it} - \theta_t y_{i,t-1})$ equation one more time to eliminate the factor error.

Remark 9.3.1: The unique determination of $\mathbf{b}_i$ and $E\,\mathbf{f}_t\mathbf{f}_t'$ is derived under the assumption that v_{it} are observable. The derivation of the least-squares regression of (9.3.11) is based on the assumption that (9.3.11) is identifiable from $(y_{it}, \mathbf{x}_{it}')$. The identification conditions for (9.3.11) remain to be explored. Neither does it appear feasible to simultaneously estimate $\boldsymbol{\beta}$, B and $\mathbf{f}_t$ when N is large. On the other hand, the two-step procedure of (9.3.12)–(9.3.14) depends on the possibility of getting initially consistent estimator of $\boldsymbol{\beta}$.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

9.4 GROUP MEAN AUGMENTED (COMMON CORRELATED EFFECTS) APPROACH TO CONTROL THE IMPACT OF CROSS-SECTIONAL DEPENDENCE

The Frisch-Waugh FGLS approach of iteratively estimating (9.3.12) and (9.3.14) (or (9.3.15) and (9.3.16)) may work for the factor approach only if both N and T are large. However, if N is large, the implementation of FGLS is cumbersome. Nevertheless, when $N \longrightarrow \infty$, $\bar{u}_t = \frac{1}{N}\sum_{i=1}^{N} u_{it} \longrightarrow 0$, model (9.2.2) and (9.3.2) (or (9.3.11)) imply that

$$\bar{\mathbf{b}}'\,\mathbf{f}_t \simeq \bar{y}_t - \bar{\mathbf{x}}_t'\boldsymbol{\beta}, \tag{9.4.1}$$

where $\bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i$, $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$ and $\bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}$. If $\mathbf{b}_i'\,\mathbf{f}_t = c_i\bar{\mathbf{b}}'\,\mathbf{f}_t$, for all t or if $\mathbf{f}_t$ can be approximated by linear combinations of $\bar{y}_t$ and $\bar{\mathbf{x}}_t$ ((9.4.1)), instead of estimating $\hat{\mathbf{f}}_t$, Pesaran (2006) suggests a simple approach to filter out the cross-sectional dependence by augmenting (9.3.18) by $\bar{y}_t$ and $\bar{\mathbf{x}}_t$,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \bar{y}_t c_i + \bar{\mathbf{x}}_t'\mathbf{d}_i + e_{it}. \tag{9.4.2}$$

The pooled estimator,

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i=1}^{N} w_i X_i' M^* X_i\right)^{-1}\left(\sum_{i=1}^{N} w_i X_i' M^* \mathbf{y}_i\right) \tag{9.4.3}$$

is consistent and asymptotically normally distributed when $N \to \infty$ and T either fixed or $\to \infty$, where $w_i = \frac{\sigma_i^2}{\sum_{j=1}^{N}\sigma_j^2}$, $\sigma_j^2 =$ Var (u_{jt}), $M^* = (I - H(H'H)^{-1}H')$, $H = (\mathbf{e}, \bar{\mathbf{y}}, \bar{X})$ and $\bar{\mathbf{y}}$, $\bar{X}$ are $T \times 1$ and $T \times K$ stacked $\bar{y}_t$ and $\bar{\mathbf{x}}_t'$, respectively. Pesaran (2006) called (9.4.3) the common correlated effects pooled estimator (CCEP). The limited Monte Carlo studies conducted by Westerlund and Urbain (2012) appear to show that the Pesaran (2006) CCEP estimator of $\boldsymbol{\beta}$ (9.4.3) is less biased than the Bai (2009) iterated least-squares estimator (9.3.12).

Kapetanios, Pesaran, and Yamagata (2011) further show that the cross-sectional average-based method is robust to a wide variety of data-generating processes. For instance, for the error process generated by a multifactor error structure (9.3.1), whether the unobservable common factors $\mathbf{f}_t$ follow $I(0)$ or unit root processes, the asymptotic properties of (9.4.3) remain similar.

Remark 9.4.1: The advantage of Pesaran's (2006) cross-sectional mean-augmented approach to take account the cross-sectional dependence is its simplicity. However, there are restrictions on its application. The method works when $\mathbf{b}_i'\,\mathbf{f}_t = c_i\bar{\mathbf{b}}'\mathbf{f}_t$ for all t or if $\mathbf{f}_t$ can be considered as linear combinations of $\bar{y}_t$ and $\bar{\mathbf{x}}_t$. It is hard to ensure $\mathbf{b}_i'\,\mathbf{f}_t = c_i\bar{\mathbf{b}}'\,\mathbf{f}_t$ if $r > 1$. For instance, consider the case that $r = 2$, $\mathbf{b}_i = (1, 1)'$, $\bar{\mathbf{b}} = (2, 0)'$, $\mathbf{f}_t = (1, 1)'$, then $\mathbf{b}_i'\,\mathbf{f}_t = \bar{\mathbf{b}}'\,\mathbf{f}_t = 2$. However, if $\mathbf{f}_s = (2, 0)'$, then $\mathbf{b}_i'\,\mathbf{f}_s = 2$ while $\bar{\mathbf{b}}'\,\mathbf{f}_s = 4$. If $\mathbf{b}_i'\,\mathbf{f}_t = c_{it}\bar{\mathbf{b}}'\,\mathbf{f}_t$, (9.4.2) does not approximate (9.3.11) and (9.4.3) is not consistent if $\mathbf{f}_t$ is

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

correlated with $\mathbf{x}_{it}$. If $\mathbf{b}_i' \,\mathbf{f}_t = c_{it}\bar{\mathbf{b}}' f_t$, additional assumptions are needed to approximate $\mathbf{b}_i' \,\mathbf{f}_t$. For instance, Pesaran (2006) assumes that

$$\mathbf{x}_{it} = \Gamma_i \,\mathbf{f}_t + \boldsymbol{\epsilon}_{it}, \tag{9.4.4}$$

$$E(\boldsymbol{\epsilon}_{it} u_{it}) = \mathbf{0}. \tag{9.4.5}$$

Then

$$\begin{aligned} \mathbf{z}_{it} = \begin{pmatrix} y_{it} \\ \mathbf{x}_{it} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{\beta}'\Gamma_i + \mathbf{b}_i' \\ \Gamma_i \end{pmatrix} \mathbf{f}_t + \begin{pmatrix} \boldsymbol{\beta}'\boldsymbol{\epsilon}_{it} + u_{it} \\ \boldsymbol{\epsilon}_{it} \end{pmatrix} \\ &= C_i \,\mathbf{f}_t + \mathbf{e}_{it}. \end{aligned} \tag{9.4.6}$$

It follows that

$$\bar{\mathbf{z}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_{it} = \bar{C}\,\mathbf{f}_t + \bar{\mathbf{e}}_t, \tag{9.4.7}$$

where $\bar{C} = \frac{1}{N}\sum_{i=1}^{N} C_i$, $\bar{\mathbf{e}}_t = \begin{pmatrix} \boldsymbol{\beta}'(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\epsilon}_{it}) + \bar{u}_t \\ \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\epsilon}_{it} \end{pmatrix}$. If $r \leq k+1$, $\bar{C}$ is of rank r and $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\epsilon}_{it} \longrightarrow \mathbf{0}$ (or $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\epsilon}_{it} \longrightarrow$ a constant vector) as $N \longrightarrow \infty$, then

$$\mathbf{f}_t \simeq (\bar{C}'\bar{C})^{-1}\bar{C}'\bar{\mathbf{z}}_t. \tag{9.4.8}$$

Then model (9.3.11) is formally identical to (9.4.2) when $(C_i, \mathbf{d}_i') = \mathbf{b}_i'(\bar{C}'\bar{C})^{-1}\bar{C}'$.

However, under (9.4.4), (9.4.5), and the additional assumption that

$$\text{Cov}\,(\Gamma_i, \mathbf{b}_i) = \mathbf{0}, \tag{9.4.9}$$

one can simply obtain a consistent estimator of $\boldsymbol{\beta}$ by adding time dummies to (9.1.1). The least-squares dummy variable estimator of $\boldsymbol{\beta}$ is equivalent to the within (time) estimator of (see Chapter 3, Section 3.2)

$$(y_{it} - \bar{y}_t) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_t)'\boldsymbol{\beta} + (v_{it} - \bar{v}_t), \tag{9.4.10}$$

where

$$v_{it} - \bar{v}_t = (\mathbf{b}_i - \bar{\mathbf{b}})'\,\mathbf{f}_t + (u_{it} - \bar{u}_t),$$

$$\mathbf{x}_{it} - \bar{\mathbf{x}}_t = (\Gamma_i - \bar{\Gamma})\,\mathbf{f}_t + (\boldsymbol{\epsilon}_{it} - \bar{\boldsymbol{\epsilon}}_t),$$

$$\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it},\; \bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{it},\; \bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}_i,\; \bar{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \Gamma_i,$$

$$\bar{u}_t = \frac{1}{N}\sum_{i=1}^{N} u_{it},$$

and $\bar{\boldsymbol{\epsilon}}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\epsilon}_{it}$. Under (9.4.9),

$$\begin{aligned} &\text{Cov}\,(\mathbf{x}_{it} - \bar{\mathbf{x}}_t, v_{it} - \bar{v}_t) \\ &= E\left\{(\Gamma_i - \bar{\Gamma})(\mathbf{b}_i - \bar{\mathbf{b}})'\right\} \text{Cov}\,(\mathbf{f}_t)\,\text{Cov}\,(\boldsymbol{\epsilon}_{it}, u_{it}) = \mathbf{0}. \end{aligned} \quad (9.4.11)$$

Therefore, as $N \to \infty$, the least-squares estimator of (9.4.10),

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(y_{it} - \bar{y}_t)\right] \quad (9.4.12)$$

is consistent and asymptotically normally distributed with covariance matrix

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{cv}) = \sigma_u^2\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)'\right]^{-1}. \quad (9.4.13)$$

(Coakley, Fuertes, and Smith 2006; Sarafidis and Wansbeek 2012). However, if (9.4.9) does not hold, (9.4.12) exhibits large bias and large size distortion (Sarafidis and Wansbeek 2012).

9.5 TEST OF CROSS-SECTIONAL INDEPENDENCE

Many of the conventional panel data estimators that ignore cross-sectional dependence are inconsistent even when $N \to \infty$ if T is finite. Modeling cross-sectional dependence is much more complicated than modeling time series dependence. So is the estimation of panel data models in the presence of cross-sectional dependence. Therefore, it could be prudent to first test cross-sectional independence and only embark on estimating models with cross-sectional dependence if the tests reject the null hypothesis of no cross-sectional dependence.

9.5.1 Linear Model

Consider a linear model,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \quad (9.5.1)$$

The spatial approach assumes a known correlation pattern among cross-sectional units, W. Under the null of cross-sectional independence, $\theta = 0$ for any W. Therefore, a test for spatial effects is a test of the null hypothesis $H_0 : \theta = 0$ (or $\delta = 0$). Burridge (1980) derives the Lagrange multiplier test statistic for model (9.2.2) or (9.2.3),

$$\tau = \frac{[\hat{\mathbf{v}}'(W \otimes I_T)\hat{\mathbf{v}}/(\hat{\mathbf{v}}'\hat{\mathbf{v}}/NT)]^2}{tr[(W^2 \otimes I_T) + (W'W \otimes I_T)]} \quad (9.5.2)$$

which is χ^2 distributed with one degree of freedom, where $\hat{\mathbf{v}} = \mathbf{y} - X\boldsymbol{\beta}$.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

For error component spatial autoregressive model (9.2.19), Anselin (1988) derived the Lagrangian multiplier (LM) test statistic for $H_0 : \theta = 0$,

$$\tau^* = \frac{\left[\frac{1}{\sigma_u^2}\hat{\mathbf{v}}^{*'}(W \otimes I_T + \hat{k}(T\hat{k} - 2)\mathbf{e}_T\mathbf{e}_T')\hat{\mathbf{v}}^*\right]}{P}, \tag{9.5.3}$$

which is asymptotically χ^2 distributed with one degree of freedom, where $\mathbf{v}^* = \mathbf{y} - X\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^N X_i' V^{*-1} X_i\right)^{-1} \left(\sum_{i=1}^N X_i' V^{*-1} \mathbf{y}_i\right)$, the usual error component estimator, $\hat{k} = \hat{\sigma}_\alpha^2\big[\hat{\sigma}_u^2\big(1 + T\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2}\big)\big]^{-1}$, and $P = (T^2\hat{k}^2 - 2\hat{k} + T)(trW^2 + trW'W)$. Baltagi et al. (2007) consider various combination of error components and the spatial parameter test. Kelejian and Prucha (2001), and Pinkse (2000) have suggested tests of cross-sectional dependence based on the spatial correlation analogue of the Durbin–Watson/Box-Pierce tests for time series correlations.

Breusch and Pagan (1980) derived an LM test statistic for cross-sectional dependence:

$$LM = T \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij}^2, \tag{9.5.4}$$

where $\hat{\rho}_{ij}$ is the estimated sample cross-correlation coefficient between the least-squares residuals $\hat{v}_{it}$ and $\hat{v}_{jt}$, where $\hat{v}_{it} = y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_i$, and $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i\mathbf{y}_i$. When N is fixed and $T \to \infty$, (9.5.4) converges to a χ^2 distribution with $\frac{N(N-1)}{2}$ degrees of freedom under the null of no cross-sectional dependence. When N is large, the scaled Lagrangian multiplier statistic (SLM),

$$\text{SLM} = \sqrt{\frac{2}{N(N-1)}}LM \tag{9.5.5}$$

is asymptotically normally distributed with mean 0 and variance 1.

Many panel data sets have N much larger than T. Because $E(T\hat{\rho}_{ij}^2) \neq 0$ for all T, SLM is not properly centered. In other words, when $N > T$, the SLM tends to overeject, often substantially.

To correct for the bias in large N and finite T panels, Pesaran et al. (2008) propose a bias-adjusted LM test,

$$\text{LM}_B = \sqrt{\frac{2}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(T-k)\hat{\rho}_{ij}^2 - \mu_{ij}}{w_{ij}}, \tag{9.5.6}$$

where $\mu_{ij} = E[(T-k)\hat{\rho}_{ij}^2]$, $w_{ij}^2 = \text{Var}\,[(T-k)\hat{\rho}_{ij}^2]$, and k is the dimension of $\mathbf{x}_{it}$. They show that (9.5.6) is asymptotically normally distributed with mean 0 and variance 1 for all $T > k + 8$. The exact expressions for μ_{ij} and w_{ij}^2 when $\mathbf{x}_{it}$ is strictly exogenous and v_{it} are normally distributed are given by Pesaran et al. (2008).

Because the adjustment of the finite sample bias of the LM test is complicated, Pesaran (2004) suggests a CD test statistic for cross-sectional dependence:

$$\mathrm{CD} = \sqrt{\frac{2T}{N(N-1)}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}\right). \tag{9.5.7}$$

When both N and $T \to \infty$, the CD test converges to a normal distribution with mean 0 and variance 1 under the null of cross-sectional independence conditional on $\mathbf{x}$. The CD test can also be applied to the linear dynamic model:

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + u_{it}. \tag{9.5.8}$$

The Monte Carlo simulations conducted in Pesaran (2004) shows that the estimated size is very close to the nominal level for any combinations of N and T considered. However, the CD test has power only if $\frac{1}{N}\sum_{i=1}^{N}\rho_{ij} \neq 0$. On the other hand, the LM test has power even if the average of the correlation coefficient is equal to 0 as long as some pairs, $\hat{\rho}_{ij} \neq 0$.

As an alternative, Sarafidis, Yamagata, and Robertson (SYR) (2009) proposed a Sargan's (1958) difference test based on the GMM estimator of (9.5.8). As shown in Chapter 4, $\boldsymbol{\theta}' = (\gamma, \boldsymbol{\beta}')$ can be estimated by the GMM method (4.3.47). SYR suggest to split W_i into two separate sets of instruments,

$$W'_{1i} = \begin{bmatrix} y_{i0} & 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & y_{i0} & y_{i1} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & y_{i0} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & y_{i0} & \cdot & \cdot & y_{i,T-2} \end{bmatrix}, \tag{9.5.9}$$

and

$$W'_{2i} = \begin{bmatrix} \mathbf{x}'_i & \mathbf{0}' & \mathbf{0}' & \cdot & \cdot \\ \mathbf{0}' & \mathbf{x}'_i & \mathbf{0}' & \cdot\cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \mathbf{x}'_i \end{bmatrix}, \tag{9.5.10}$$

where $\mathbf{x}'_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})$, W'_{1i} is $(T-1) \times T(T-1)/2$, W'_{2i} is $(T-1) \times KT(T-1)$ and $\mathbf{x}_{it}$ is strictly exogenous.[7]

Under the null of no cross-sectional dependence, both sets of moment conditions

$$E[\mathbf{W}_{1i}\Delta\mathbf{u}_i] = \mathbf{0}, \tag{9.5.11}$$

[7] If $\mathbf{x}_{it}$ is predetermined rather than strictly exogenous, a corresponding W_2 can be constructed as

$$W_2 = \begin{bmatrix} \mathbf{x}'_1, & \mathbf{0}' & \mathbf{0}' & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \mathbf{x}'_1 & \mathbf{x}'_2 & \mathbf{0}' & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{x}'_1 & \cdot & \mathbf{x}'_{T-1} \end{bmatrix}$$

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

and

$$E[\mathbf{W}_{2i}\Delta\mathbf{u}_i] = \mathbf{0}, \tag{9.5.12}$$

hold. However, if there exists cross-sectional dependence, (9.5.11) may not hold. For instance, suppose u_{it} can be decomposed into the sum of two components, the impact of r time-varying common omitted factors and an idiosyncratic component, ϵ_{it},

$$u_{it} = \mathbf{b}_i' \, \mathbf{f}_t + \epsilon_{it}. \tag{9.5.13}$$

For simplicity, we assume ϵ_{it} is independently distributed over i and t. Then the first difference of u_{it},

$$\Delta u_{it} = \mathbf{b}_i' \Delta \, \mathbf{f}_t + \Delta\epsilon_{it}, \tag{9.5.14}$$

and

$$\begin{aligned} y_{it} = {} & \frac{1-\gamma^t}{1-\gamma}\alpha_i + \gamma^t y_{i0} + \sum_{j=0}^{t-1} \gamma^j \mathbf{x}_{i,t-j}' \boldsymbol{\beta} \\ & + \mathbf{b}_i' \sum_{j=0}^{t-1} \gamma^j \, \mathbf{f}_{t-j} + \sum_{j=0}^{t-1} \gamma^j \epsilon_{i,t-j}. \end{aligned} \tag{9.5.15}$$

Under the assumption that $\mathbf{f}_t$ are nonstochastic and bounded but $\mathbf{b}_i$ are random with mean $\mathbf{0}$ and covariance $E\mathbf{b}_i\mathbf{b}_i' = \sum_b$, $E(y_{i,t-j}\Delta u_{it})$ is not equal to 0, for $j = 2, \ldots, t$. Therefore, SYR suggest estimating γ and $\boldsymbol{\beta}$ by (4.3.45) first using both (9.5.11) and (9.5.12) moment conditions, denoted by $(\hat{\gamma}, \hat{\boldsymbol{\beta}}')$, construct estimated residuals $\Delta\mathbf{u}_i$ by $\Delta\hat{u}_i = \Delta\mathbf{y}_i - \Delta y_{i,-1}\hat{\gamma} - \Delta X_i\hat{\boldsymbol{\beta}}$, where $\Delta\mathbf{y}_i = (\Delta y_{i2}, \ldots, \Delta y_{iT})'$, $\Delta\mathbf{y}_{i,-1} = (\Delta y_{i1}, \ldots, \Delta y_{i,T-1})'$ and $\Delta X_i = (\Delta\mathbf{x}_{i1}, \ldots, \Delta\mathbf{x}_{iT})'$. Then estimate $(\gamma, \boldsymbol{\beta}')$ using moment conditions (9.5.12) only,

$$\begin{aligned} \begin{pmatrix} \tilde{\boldsymbol{\gamma}} \\ \tilde{\boldsymbol{\beta}} \end{pmatrix} = {} & \left\{ \left[\sum_{i=1}^N \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_{2i} \right] \hat{\Omega}^{-1} \left[\sum_{i=1}^N W_{2i}'(\Delta\mathbf{y}_{i,-1}, \Delta X_i) \right] \right\}^{-1} \\ & \cdot \left\{ \left[\sum_{i=1}^N \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_{2i} \right] \hat{\Omega}^{-1} \left[\sum_{i=1}^N W_{2i}'\Delta\mathbf{y}_i \right] \right\}, \end{aligned} \tag{9.5.16}$$

where $\hat{\Omega}^{-1} = N^{-1}\sum_{i=1}^N W_{2i}'\Delta\hat{\mathbf{u}}_i\Delta\hat{\mathbf{u}}_i' W_{2i}$. Under the null of cross-sectional independence both estimators are consistent. Under the alternative, $(\hat{\gamma}, \hat{\boldsymbol{\beta}}')$ may not be consistent but (9.5.16) remains consistent. Therefore, SYR, following the idea of Sargan (1958) and Hansen (1982), suggest using the test statistic

$$\begin{aligned} & N^{-1}\left(\sum_{i=1}^{N}\Delta\hat{\mathbf{u}}_i' W_i\right)\hat{\Psi}^{-1}\left(\sum_{i=1}^{N} W_i'\Delta\hat{\mathbf{u}}_i\right) \\ & \quad - N^{-1}\left(\sum_{i=1}^{N}\Delta\tilde{\mathbf{u}}_i' W_{2i}\right)\tilde{\Psi}^{-1}\left(\sum_{i=1}^{N} W_{2i}'\Delta\tilde{\mathbf{u}}_i\right) \end{aligned} \tag{9.5.17}$$

where $\Delta\tilde{\mathbf{u}}_i = \Delta\mathbf{y}_i - \Delta\mathbf{y}_{i,-1}\tilde{\gamma} - \Delta X_i\tilde{\boldsymbol{\beta}}$, $\hat{\Psi} = \frac{1}{N}\sum_{i=1}^{N} W_i'\Delta\hat{\mathbf{u}}_i\Delta\hat{\mathbf{u}}_i' W_i$ and $\tilde{\Psi} = \frac{1}{N}\sum_{i=1}^{N} W_{2i}'\Delta\tilde{\mathbf{u}}_i\Delta\tilde{\mathbf{u}}_i' W_{2i}$. SYR show that under the null of cross-sectional independence, (9.5.17) converges to a χ^2 distribution with $\frac{T(T-1)}{2}(1+K)$ degrees of freedom as $N \to \infty$.

The advantage of the SYR test is that the test statistic (9.5.17) has power even if $\sum_{j=1}^{N}\rho_{ij} = 0$. Monte Carlo studies conducted by SYR show that the test statistic (9.5.17) performs well if the cross-sectional dependence is driven by nonstochastic $\mathbf{f}_t$ but stochastic $\mathbf{b}_i$. However, if the cross-sectional dependence is driven by fixed $\mathbf{b}_i$ and stochastic $\mathbf{f}_t$, then the test statistic is unlikely to have power because $E(\Delta y_{i,t-j}\Delta u_i) = 0$ if $\mathbf{f}_t$ is independently distributed over time.[8]

9.5.2 Limited Dependent-Variable Model

Many limited dependent-variable models take the form of relating observed y_{it} to a latent y_{it}^*, (e.g., Chapters 7 and 8),

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{9.5.18}$$

through a link function $g(\cdot)$

$$y_{it} = g(y_{it}^*). \tag{9.5.19}$$

For example, in the binary choice model,

$$g(y_{it}^*) = I(y_{it}^* > 0), \tag{9.5.20}$$

and in the Tobit model,

$$g(y_{it}^*) = y_{it}^* I(y_{it}^* > 0), \tag{9.5.21}$$

where I(A) is an indicator function that takes the value 1 if A occurs and 0 otherwise.

There is a fundamental difference between the linear model and limited dependent-variable model. There is a one-to-one correspondence between v_{it}

[8] $E(\Delta y_{i,t-j}\Delta u_{it})$ is not equal to 0 if $\mathbf{f}_t$ is serially correlated. However, if $\mathbf{f}_t$ is serially correlated, then u_{it} is serially correlated and $y_{i,t-j}$ is not a legitimate instrument if the order of serially correlation is greater than j. Laggerd y can be legitimate instruments only if $E(\Delta u_{it} y_{i,t-s}) = 0$. Then the GMM estimator of (4.3.47) will have to be modified accordingly.

and y_{it} in the linear model, but not in limited dependent variable model. The likelihood for observing $\mathbf{y}_t = (y_{it}, \ldots, y_{Nt})'$,

$$P_t = \int_{A(\mathbf{v}_t \mid \mathbf{y}_t)} f(\mathbf{v}_t) d\mathbf{v}_t, \tag{9.5.22}$$

where $A(\mathbf{v}_t \mid \mathbf{y}_t)$ denotes the region of integration of $\mathbf{v}_t = (v_{it}, \ldots, v_{Nt})'$ which is determined by the realized $\mathbf{y}_t$ and the form of the link function. For instance, in the case of probit model, $A(\mathbf{v}_t \mid \mathbf{y}_t)$ denotes the region $(a_{it} < v_{it} < b_{it})$, where $a_{it} = -\mathbf{x}_{it}'\boldsymbol{\beta}$, $b_{it} = \infty$ if $y_{it} = 1$ and $a_{it} = -\infty$, $b_{it} = -\mathbf{x}_{it}'\boldsymbol{\beta}$ if $y_{it} = 0$.

Under the assumption that v_{it} is independently normally distributed across i, Hsiao, Pesaran, and Picks (2012) show that the Lagrangian multiplier test statistic of cross-sectional independence takes an analogous form:

$$LM = T \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \tilde{\rho}_{ij}^2, \tag{9.5.23}$$

where

$$\tilde{\rho}_{ij} = \frac{T^{-1} \sum\limits_{t=1}^{T} \tilde{v}_{it} \tilde{v}_{jt}}{\sqrt{T^{-1} \sum\limits_{t=1}^{T} \tilde{v}_{it}^2} \sqrt{T^{-1} \sum\limits_{t=1}^{T} \tilde{v}_{jt}^2}}, \tag{9.5.24}$$

and $\tilde{v}_{it} = E(v_{it} \mid y_{it})$, the conditional mean of v_{it} given y_{it}. For instance, in the case of probit model,

$$\tilde{v}_{it} = \frac{\phi(\mathbf{x}_{it}'\boldsymbol{\beta})}{\Phi(\mathbf{x}_{it}'\boldsymbol{\beta})[1 - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta})]} [y_{it} - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta})]. \tag{9.5.25}$$

In the case of the Tobit model

$$\tilde{v}_{it} = (y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}) I(y_{it} > 0) - \sigma_i \frac{\phi(\frac{\mathbf{X}_{it}'\boldsymbol{\beta}}{\sigma_i})}{\Phi(-\frac{\mathbf{X}_{it}'\boldsymbol{\beta}}{\sigma_i})} [1 - I(y_{it} > 0)], \tag{9.5.26}$$

where $\sigma_i^2 = \text{Var}\ (v_{it})$, $\phi(\cdot)$ and $\Phi(\cdot)$ denote standard normal and integrated standard normal. Under the null of cross-sectional independence, (9.5.23) converges to a χ^2 distribution with $\frac{N(N-1)}{2}$ degrees of freedom if N is fixed and $T \to \infty$. When N is also large

$$\sqrt{\frac{2}{N(N-1)}} LM \tag{9.5.27}$$

is asymptotically standard normally distributed.

When N is large and T is finite, the LM test statistically is not centered properly. However, for the nonlinear model, the bias correction factor is not easily derivable. Hsiao et al. (2012) suggest constructing Pesaran (2006) CD statistic using $\tilde{v}_{it}$.

Sometimes, the deviation of $\tilde{v}_{it}$ is not straightforward for a nonlinear model. Hsiao et al. (2012) suggest replacing $\tilde{v}_{it}$ by

$$v_{it}^* = y_{it} - E(y_{it} \mid \mathbf{x}_{it}) \tag{9.5.28}$$

in the construction of an LM or CD test statistic. Monte Carlo experiments conducted by Hsiao et al. (2012) show that there is very little difference between the two procedures to construct CD tests.

9.5.3 An Example – A Housing Price Model of China

Mao and Shen (2013) consider China's housing price model using 30 provincial-level quarterly data from the second quarter of 2001 to the fourth quarter of 2012 of the logarithm of seasonally adjusted real house price, y_{it}, as a linear function of the logarithm of seasonally adjusted real per capita wage income (x_{1it}); the logarithm of real long-term interest rate (x_{2it}); and the logarithm of the urban population (x_{3it}). Table 9.1 provides Mao and Shen (2013) estimates of the mean group estimator $\hat{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i$ for the cross sectionally independent heterogeneous model (MG),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + v_{it}; \tag{9.5.29}$$

the Pesaran (2006) common correlated effects heterogeneous model (CCEMG),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \bar{y}_t c_i + \bar{\mathbf{x}}_t'\mathbf{d}_i + v_{it}; \tag{9.5.30}$$

and the homogeneous common correlated effects model (CCEP),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{y}_t \tau_i + \bar{\mathbf{x}}_t'\mathbf{d}_i + v_{it}. \tag{9.5.31}$$

It can be seen from the results in Table 9.1 that (1) the estimated slope coefficients, $\boldsymbol{\beta}$, are very sensitive to the adjustment (CCEMG or CCEP) or nonadjustment of cross-sectional dependence, and (2) the suggested approach to control the impact of cross-sectional dependence works only if the observed data satisfy the assumptions underlying the approach. (See Remark 9.4.1 for the limitation of augmenting regression models by the cross-sectional mean.) As one can see from Table 9.1, the Pesaran (2004) CD tests (9.5.7) of the residuals of the (9.5.30) and (9.5.31) indicate that significant cross-sectional dependence remains. It is only by further adjusting the common correlated effects model residuals by a spatial model with the spatial weight matrix specified in terms of the geometric distance between region i and j that Mao and Shen (2013) can achieve cross-sectional independence to their model.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Table 9.1. *Common correlated effects estimation*

	MG			CCEMG			CCEP		
x1	1.088‡	1.089‡	0.979‡	0.264	0.313†	0.308^{+}	0.388†	0.467‡	0.449‡
	(0.058)	(0.056)	(0.114)	(0.176)	(0.173)	(0.170)	(0.169)	(0.165)	(0.170)
x2	–	−0.003	−0.052	–	−6.453	4.399	–	−4.796	4.387
	–	(0.058)	(0.057)	–	(2.927)	(2.839)	–	(3.943)	(3.401)
x3	–	–	0.718	–	–	−0.098	–	–	0.104
	–	–	(0.484)	–	–	(0.552)	–	–	(0.130)
CD	28.15‡	30.39‡	27.64‡	−4.257‡	−.4173‡	−4.073‡	−4.521‡	−4.494‡	−4.518‡

Symbols $^{+}$, †, and ‡ denote that the corresponding stastics are significant at 10%, 5%, and 1% level respectively. The values in parentheses are corresponding standard errors.

Source: Mao and Shen (2013, Table V).

9.6 A PANEL DATA APPROACH FOR PROGRAM EVALUATION

9.6.1 Introduction

Individuals are often given “treatments,” such as a drug trial, a training program, and so forth. If it is possible to simultaneously observe the same person in the treated and untreated states, then it is fairly straightforward to isolate the treatment effects in question. When it is not possible to simultaneously observe the same person in the treated and untreated states or the assignment of individuals to treatment is nonrandom, treatment effects could confound with the factors that would make people different on outcome measures or with the sample selection effects or both.

In this section we first review some basic approaches for measuring treatment effects with cross-sectional data, and then we show how the availability of panel data can substantially simplify the inferential procedure.

9.6.2 Definition of Treatment Effects

For individual i, let (y_i^{0*}, y_i^{1*}) be the potential outcomes in the untreated and treated state. Suppose the outcomes can be decomposed as the sum of the effects of observables, $\mathbf{x}$, $m_j(\mathbf{x})$, and unobservables, ϵ_j; $j = 0, 1$, in the form,

$$y_i^{0*} = m_0(\mathbf{x}_i) + \epsilon_i^0, \tag{9.6.1}$$

$$y_i^{1*} = m_1(\mathbf{x}_i) + \epsilon_i^1, \tag{9.6.2}$$

where ϵ_i^0 and ϵ_i^1 are the 0 mean unobserved random variables, assumed to be independent of $\mathbf{x}_i$. The treatment effect for individual i is defined as

$$\Delta_i = y_i^{1*} - y_i^{0*}. \tag{9.6.3}$$

The average treatment effect (ATE) (or the mean impact of treatment if people were randomly assigned to the treatment)[9] is defined as

$$\begin{aligned}\Delta^{\text{ATE}} &= E[y_i^{1*} - y_i^{0*}] = E\left\{[m_1(\mathbf{x}) - m_0(\mathbf{x})] + (\epsilon^1 - \epsilon^0)\right\} \\ &= E[m_1(\mathbf{x}) - m_0(\mathbf{x})].\end{aligned} \tag{9.6.4}$$

Let d_i be the dummy variable indicating an individual’s treatment status with $d_i = 1$ if the ith individual receives the treatment and 0 otherwise. The effect of treatment on the treated (TT) (or the mean impact of treatment of those who received treatment compared to what they would have been in the absence of

[9] See Heckman (1997), Heckman and Vytacil (2001), and Imbens and Angrist (1994) for the definitions of the marginal treat effect (MTE) and the local average treatment effect (LATE).

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

treatment) is defined as

$$\begin{aligned}\Delta^{\text{TT}} &= E(y^{1*} - y^{0*} \mid d = 1)\\ &= E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 1] + E[\epsilon^1 - \epsilon^0 \mid d = 1].\end{aligned} \tag{9.6.5}$$

Similarly, we can define the effect of treatment on untreated group as

$$\begin{aligned}\Delta^{\text{TUT}} &= E(y^{1*} - y^{0*} \mid d = 0)\\ &= E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 0] + E[\epsilon^1 - \epsilon^0 \mid d = 0].\end{aligned} \tag{9.6.6}$$

The ATE is of interest if one is interested in the effect of treatment for a randomly assigned individual or population mean response to treatment. The TT is of interest if the same selection rule for treatment continues in the future. The relation between ATE and TT is given by

$$\Delta^{\text{ATE}} = \text{Prob}\ (d = 1)\Delta^{\text{TT}} + \ \text{Prob}\ (d = 0)\Delta^{\text{TUT}}. \tag{9.6.7}$$

If $E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 1] = E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 0] = E[m_1(\mathbf{x}) - m_0(\mathbf{x})]$ and $E[\epsilon^1 - \epsilon^0 \mid d = 1] = E[\epsilon^1 - \epsilon^0 \mid d = 0] = E[\epsilon^1 - \epsilon^0]$, then $\Delta^{\text{ATE}} = \Delta^{\text{TT}} = \Delta^{\text{TUT}}$.

If we simultaneously observe y_i^{0*} and y_i^{1*} for a given i, then ATE and TT can be easily measured. However, for a given i, the observed outcome, y_i is either y_i^{0*} or y_i^{1*}, not both,

$$y_i = d_i y_i^{1*} + (1 - d_i)y_i^{0*}. \tag{9.6.8}$$

If we measure the treatment effect by comparing the mean difference between those receiving the treatment (the treatment group) and those not receiving the treatment (control group), $\frac{1}{n_d}\sum_{i\in\psi} y_i$ and $\frac{1}{N-n_d}\sum_{i\epsilon\bar{\psi}} y_i$, where $\psi = \{i \mid d_i = 1\}$, $\bar{\psi}_i = \{i \mid d_i = 0\}$ and $n_d = \sum_{i=1}^N d_i$,

$$\begin{aligned}&\frac{1}{n_d}\sum_{i\epsilon\psi} y_i - \frac{1}{N - n_d}\sum_{i\epsilon\bar{\psi}} y_i \longrightarrow E(y \mid d = 1) - E(y \mid d = 0)\\ &\quad = \{E[m_1(x) - m_0(x) \mid d = 1]\} \ + \ \{E[m_0(x) \mid d = 1] -\\ &\qquad E[m_0(x) \mid d = 0]\} \ + \ \{E(\epsilon^1 \mid d = 1) - E(\epsilon^0 \mid d = 0)\}.\end{aligned} \tag{9.6.9}$$

The average difference between the treatment group and control group ((9.6.9)) is the sum of three components, the treatment effect of the treated, Δ^{TT}, $E[m_1(x) - m_0(x) \mid d = 1]$, the effects of confounding variables being different between the treatment group and control group, $E[m_0(x) \mid d = 1] - E[m_0(x) \mid d = 0]$, and the participation (or selection effects, $E(\epsilon^1 \mid d = 1) - E(\epsilon^0 \mid d = 0)$). If participation of treatment is random, then $E(\epsilon^1 \mid d = 1) = E(\epsilon^1) = E(\epsilon^0 \mid d = 0) = E(\epsilon^0) = 0$. If $f(x \mid d = 1) = f(x \mid d = 0) = f(x)$, then $E[m_1(x) - m_0(x) \mid d = 1] = E[m_1(x) - m_0(x)]$

and $E[m_0(x) \mid d = 1] - E[m_0(x) \mid d = 0] = 0$, (9.6.9) provides an unbiased measure of $\Delta^{ATE}(\equiv \Delta^{TT})$.[10]

In an observational study, the treatment group and control group are often drawn from different populations (e.g., LaLonde 1986; Dehejia and Wahba 1999). For instance, the treatment group can be drawn from welfare recipients eligible for a program of interest while the control group can be drawn from a different population. If there are systematic differences between the treatment group and comparison group in observed and unobserved characteristics that affect outcomes, estimates of treatment effects based on the comparison of the difference between $\frac{1}{n_d}\sum_{i \in \psi} y_i - \frac{1}{N-n_d}\sum_{i \in \bar{\psi}} y_i$ are distorted. The distortion can come from either one or both of the following two sources:

(1) Selection bias due to observables, $E\{m_0(\mathbf{x}) \mid d = 1\} \neq E\{m_0(\mathbf{x}) \mid d = 0\}$ (or $E\{m_1(\mathbf{x}) \mid d = 1\} \neq E\{m_1(\mathbf{x}) \mid d = 0\}$), that is, bias due to differences in observed (conditional) variables between the two groups.

(2) Selection bias due to unobservables, that is, bias due to differences in unobserved characteristics between the two groups, $E(\epsilon^0 \mid d = 1) \neq E(\epsilon^0 \mid d = 0)$, (and $E(\epsilon^1 \mid d = 1) \neq E(\epsilon^1 \mid d = 0)$).

A variety of matching and statistical-adjustment procedures have been proposed to take account of discrepancies in observed and unobserved characteristics between treatment and control group members (e.g., Heckman and Robb 1985; Heckman, Ichimura, and Todd 1998; LaLonde 1986; Rosenbaum and Rubin 1983). We shall first review methods for the analysis of cross-sectional data, and then discuss the panel data approach.

9.6.3 Cross-Sectional Adjustment Methods

9.6.3.1 Parametric Approach

Suppose y_i^{1*} and y_i^{0*} ((9.6.1) and (9.6.2)) can be specified parametrically. In addition, if the participation of treatment is assumed to be a function of

$$d_i^* = h(\mathbf{z}_i) + v_i, \qquad (9.6.10)$$

where

$$d_i = \begin{cases} 1, & \text{if } d_i^* > 0, \\ 0, & \text{if } d_i^* \leq 0, \end{cases} \qquad (9.6.11)$$

and $\mathbf{z}$ denote the factors determining the selection equation that may overlap with some or all elements of $\mathbf{x}$. With a known joint distribution of $f(\epsilon^1, \epsilon^0, v)$, the mean response functions $m_1(\mathbf{x})$, $m_0(\mathbf{x})$ can be consistently estimated by the maximum-likelihood method and

$$A\hat{T}E(\mathbf{x}) = \hat{m}_1(\mathbf{x}) - \hat{m}_0(\mathbf{x}) \qquad (9.6.12)$$

(e.g., Damrongplasit, Hsiao, and Zhao 2010).

[10] Similarly, we can write $E(y \mid d = 1) - E(y \mid d = 0) = \Delta^{\text{TUT}} + \{E(m_1(\mathbf{x}) \mid d = 1) - E(m_1(\mathbf{x}) \mid d = 0\} + \{E(\epsilon^1 \mid d = 1) - E(\epsilon^1 \mid d = 0)\}$.

9.6.3.2 Nonparametric Approach

If $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ are unspecified, they can be estimated by nonparametric methods provided that conditional on a set of confounding variables, say $\mathbf{x}$, the distributions of (y^{1*}, y^{0*}) are independent of d (or d^*). In other words, conditional on $\mathbf{x}$, there is no selection on unobservables (conditional independence),

$$E(y^{1*} \mid d, \mathbf{x}) = E(y^{1*} \mid \mathbf{x}), \tag{9.6.13}$$

$$E(y^{0*} \mid d, \mathbf{x}) = E(y^{0*} \mid \mathbf{x}). \tag{9.6.14}$$

Then conditional on $\mathbf{x}$, the average treat effect, ATE($\mathbf{x}$),

$$\begin{aligned} \text{ATE}(\mathbf{x}) &= E(y^{1*} - y^{0*} \mid \mathbf{x}) \\ &= E(y \mid d = 1, \mathbf{x}) - E(y \mid d = 0, \mathbf{x}) \\ &= E(y^{1*} \mid \mathbf{x}) - E(y^{0*} \mid \mathbf{x}) \end{aligned} \tag{9.6.15}$$

9.6.3.2 (i) Matching Observables in Terms of Propensity Score Method (or Selection on Observables Adjustment)

However, if the dimension of $\mathbf{x}$ is large, the nonparametric method may suffer from "the curse of dimensionality." As a dimension reduction method, Rosenbaum and Rubin (1983) have suggested a propensity score method to match the observable characteristics of the treatment group and the control group. The Rosenbaum and Rubin (1983) propensity score methodology supposes unit i has observable characteristics $\mathbf{x}_i$. Let $P(\mathbf{x}_i)$ be the probability of unit i having been assigned to treatment, called the propensity score in statistics and choice probability in econometrics, defined as $P(\mathbf{x}_i) = \text{Prob}\,(d_i = 1 \mid \mathbf{x}_i) = E(d_i \mid \mathbf{x}_i)$. Assume that $0 < P(\mathbf{x}_i) < 1$ for all $\mathbf{x}_i$,[11] and $\text{Prob}\,(d_1, \ldots, d_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^N P(\mathbf{x}_i)^{d_i}[1 - P(\mathbf{x}_i)]^{1-d_i}$, for $i = 1, \ldots, N$. If the treatment assignment is ignorable given $\mathbf{x}$, then it is ignorable given $P(\mathbf{x})$; that is,

$$\{(y_i^{1*}, y_i^{0*}) \perp d_i \mid \mathbf{x}_i\} \Longrightarrow \{(y_i^{1*}, y_i^{0*}) \perp d_i \mid P(\mathbf{x}_i)\}, \tag{9.6.16}$$

where $\perp$ denotes orthogonality.

To show (9.6.16) holds, it is sufficient to show that

$$\begin{aligned} &\text{Prob}\,(d = 1 \mid y^{0*}, y^{1*}, P(\mathbf{x})) \\ &\quad = \text{Prob}\,(d = 1 \mid P(\mathbf{x})) \\ &\quad = P(\mathbf{x}) = \text{Prob}\,(d = 1 \mid \mathbf{x}) = \text{Prob}\,(d = 1 \mid y^{0*}, y^{1*}, \mathbf{x}) \end{aligned} \tag{9.6.17}$$

[11] The assumption that $0 < P(\mathbf{x}_i) < 1$ guarantees that for each $\mathbf{x}_i$, we obtain observations in both the treated and untreated states. This assumption can be relaxed as long as there are $\mathbf{x}$ such that $0 < P(\mathbf{x}) < 1$.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Eq. (9.6.17) follows from applying the ignorable treatment assignment assumption to

$$\begin{aligned}\text{Prob}\,(d = 1 \mid y^{0*}, y^{1*}, P(\mathbf{x}))&\\ &= E_x\left\{\text{Prob}\,(d = 1 \mid y_0^*, y_1^*, \mathbf{x}) \mid y^{0*}, y^{1*}, P(\mathbf{x})\right\}\\ &= E_x\left\{\text{Prob}\,(d = 1 \mid \mathbf{x}) \mid y^{0*}, y^{1*}, P(\mathbf{x})\right\}\\ &= E_x\left\{P(\mathbf{x}) \mid y^{0*}, y^{1*}, P(\mathbf{x})\right\}\\ &= E_x\left\{P(\mathbf{x}) \mid P(\mathbf{x})\right\} = P(\mathbf{x}),\end{aligned} \tag{9.6.18}$$

where E_x denotes taking the expectation with respect to $\mathbf{x}$.

It follows from (9.6.16) that

$$\mathbf{x}_i \perp d_i \mid P(\mathbf{x}_i). \tag{9.6.19}$$

To prove (9.6.19), it is sufficient to show that

$$\text{Prob}\,(d = 1 \mid \mathbf{x}, P(\mathbf{x})) = \text{Prob}\,(d = 1 \mid P(\mathbf{x})). \tag{9.6.20}$$

Equation (9.6.20) follows from $\text{Prob}(d = 1 \mid \mathbf{x}, P(\mathbf{x})) = \text{Prob}\,(d = 1 \mid \mathbf{x}) = P(\mathbf{x})$ and

$$\begin{aligned}\text{Prob}\,(d = 1 \mid P(\mathbf{x})) &= E_x\,\{\text{Prob}\,(d = 1 \mid \mathbf{x}, P(\mathbf{x})) \mid P(\mathbf{x})\}\\ &= E_x\,\{P(\mathbf{x}) \mid P(\mathbf{x})\} = P(x).\end{aligned}$$

Equation (9.6.19) implies that the conditional density of $\mathbf{x}$ given d and $P(\mathbf{x})$,

$$f(\mathbf{x} \mid d = 1, P(\mathbf{x})) = f(\mathbf{x} \mid d = 0, P(\mathbf{x})) = f(\mathbf{x} \mid P(\mathbf{x})). \tag{9.6.21}$$

In other words, Equation (9.6.19) implies that if a subclass of units or a matched treatment–control pair is homogeneous in $P(\mathbf{x})$, then the treated and control units in that subclass or matched pair will have the same distribution of $\mathbf{x}$. In other words, at any value of a propensity score, the mean difference between the treatment group and control group is an unbiased estimate of the average treatment effect at that value of the propensity score if treatment assignment is ignorable.

$$\Delta(P(\mathbf{x}))^{\text{TT}} = E\{E(y \mid d = 1, P(\mathbf{x})) - E(y \mid d = 0, P(\mathbf{x})) \mid d = 1\}, \tag{9.6.22}$$

where the outer expectation is over the distribution of $\{P(\mathbf{x}) \mid d = 1\}$.

The attraction of propensity score matching method is that in (9.6.15) we condition on $\mathbf{x}$ (intuitively, to find observations with similar covariates), while in (9.6.22) we condition just on the propensity score because (9.6.22) implies that observations with the same propensity score have the same distribution of the full vector of covariates, $\mathbf{x}$. Equation (9.6.19) asserts that conditional on $P(\mathbf{x})$, the distribution of covariates should be the same across the treatment and comparison groups. In other words, conditional on the propensity score,

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

each individual has the same probability of assignment to treatment as in a randomized experiment. Therefore, the estimation of average treatment effect for the treated[12] can be done in two steps. The first step involves the estimation of propensity score parametrically or nonparametrically (e.g., see Chapter 7). In the second step, given the estimated propensity score, one can estimate $E\{y \mid P(\mathbf{x}), d = j\}$ for $j = 0, 1$, take the difference between the treatment and control groups, then weight these by the frequency of treated observations or frequency of (both treated and untreated) observations in each stratum to get an estimate of TT or ATE ($E\{E[y \mid d = 1, P(\mathbf{x})] - E[y \mid d = 0, P(\mathbf{x})]\} = E\{E[y_1 - y_0 \mid P(\mathbf{x})]\}$), where the outer expectation is with respect to the propensity score, $P(\mathbf{x})$. For examples of using this methodology to evaluate the effects of training programs in nonexperimental studies, see Dehejia and Wahba (1999), and LaLonde (1986), Liu, Hsiao, Matsumoto, and Chou (2009), etc.

9.6.3.2 (ii) Regression Discontinuity Design

Let $\mathbf{x}_i = (w_i, \mathbf{q}_i')$ be k covariates, where w_i is a scalar and $\mathbf{q}_i$ is a $(k-1) \times 1$ vector. Both w_i and $\mathbf{q}_i$ are not affected by the treatment. The basic idea behind the regression discontinuity (RD) design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor w_i being on either side of a fixed threshold. This predictor, w_i, (together with $\mathbf{q}_i$), also affects the potential outcomes.

For notational ease, we shall assume $\mathbf{q}_i = 0$ for this subsection. In the sharp RD (SRD) designs, it is assumed that all units with the values of w at least c are assigned to the treatment group and participation is mandatory for these individuals, and with values of w less than c are assigned to the control groups and members of these group are not eligible for the treatment, then

$$\begin{aligned} \text{ATE}(c) &= \lim_{w \downarrow c} E(y \mid w) - \lim_{w \uparrow c} E(y \mid w), \\ &= E(y^1 - y^0 \mid w = c) \end{aligned} \tag{9.6.23}$$

(This approach although assumes unconfoundedness of Rosenbaum and Rubin (1983), however, it violates $0 < P(d = 1 \mid \mathbf{x}) < 1$).

This approach assumes either

(1) $E(y^0 \mid w)$ and $E(y^1 \mid w)$ are continuous in w or (2) $F_{y^0|w}(y \mid w)$ and $F_{y^1|w}(y \mid w)$ are continuous in w for all y.

In the fuzzy RD (FRD), we allow

$$\lim_{w \downarrow c} \text{Prob}(d = 1 \mid w) \neq \lim_{w \uparrow c} \text{Prob}(d = 1 \mid w), \tag{9.6.24}$$

then

$$\text{ATE}(c) = \frac{\lim_{w \downarrow c} E(y \mid w) - \lim_{w \uparrow c} E(y \mid w)}{\lim_{w \downarrow c} P(d = 1 \mid w) - \lim_{w \uparrow c} P(d = 1 \mid w)}. \tag{9.6.25}$$

[12] The measurement is of interest if future selection criteria for treatment are like past selection criteria.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Let

$$\lim_{w\downarrow c} \text{Prob}(d = 1 \mid w) - \lim_{w\uparrow c} \text{Prob } (d = 1 \mid w) = \triangledown. \tag{9.6.26}$$

$$P = \lim_{w\uparrow c} \text{Prob } (d = 1 \mid w)$$

Then

$$\begin{aligned} &\lim_{w\downarrow c} E(y \mid w) - \lim_{w\uparrow c} E(y \mid w) \\ &= \left\{(P + \triangledown)Ey^1 - (1 - P - \triangledown)Ey^0\right\} - [PEy^1 + (1 - p)Ey^0] \\ &= \triangledown E[y^1 - y^0] \end{aligned} \tag{9.6.27}$$

Both the SRD and FRD designs provide only estimates of the ATE for a subpopulation with $w_i = c$. The designs do not allow the estimation of the overall ATE.

Let $\psi = \{i \mid w_i < c\}$ and $\bar{\psi} = \{i \mid w_i \geq c\}$, then for the SRD, we may estimate the ATE(c) by the kernel method,

$$\widehat{\text{ATE}}(c) = \frac{\sum\limits_{i\in\bar{\psi}} y_i K(\frac{w_i - c}{h}) - \sum\limits_{i\in\psi} y_i K(\frac{w_i - c}{h})}{\sum\limits_{i\in\bar{\psi}} K(\frac{w_i - c}{h}) - \sum\limits_{i\in\psi} K(\frac{w_i - c}{h}).}, \tag{9.6.28}$$

where $K(\cdot)$ is a kernel function satisfying $K(0) \neq 0$, $K(v) \to 0$ as $v \to \pm\infty$. Or use the Fan and Gijbels (1992) local linear regression approach,

$$\min_{\alpha_0,\beta_0} \sum_{i:c-h<x_i<c} (y_i - \alpha_0 - \beta_0(w_i - c))^2 \tag{9.6.29}$$

and

$$\min_{\alpha_1,\beta_1} \sum_{i:c\leq x_i<c+h} (y_i - \alpha_1 - \beta_1(w_i - c))^2 \tag{9.6.30}$$

Since $E(y^1 \mid w = c) = \hat{\alpha}_1 + \beta_1(c - c) = \hat{\alpha}_1$ and $E(y^0 \mid w = c) = \hat{\alpha}_0 + \hat{\beta}_0(c - c) = \hat{\alpha}_0$, therefore

$$\widehat{\text{ATE}}(c) = \hat{\alpha}_1 - \hat{\alpha}_0. \tag{9.6.31}$$

For FRD,

$$\widehat{\text{ATE}}(c) = \frac{\hat{\alpha}_1 - \hat{\alpha}_0}{\hat{\gamma}_1 - \hat{\gamma}_0}, \tag{9.6.32}$$

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

where $(\hat{\gamma}_1, \hat{\delta}_1)$ is the solution of

$$\min \sum_{i:c \le x_i < c+h} (d_i - \gamma_1 - \delta_1(w_i - c))^2 \tag{9.6.33}$$

and $(\hat{\gamma}_0, \hat{\delta}_0)$ is the solution of

$$\min \sum_{i:c-h \le x_i < c} (d_i - \gamma_0 + \delta_0(w_i - c))^2. \tag{9.6.34}$$

(For a survey of RD, see Imbens, and Lemieux 2008.)

9.6.3.3 Summary of Cross-Sectional Approaches

The advantages of the parametric approach are that it can simultaneously take account of both selection on observables and selection on unobservables. It can also estimate the impact of each explanatory variable. The disadvantage is that it needs to impose both functional form and distributional assumptions. If the prior information is inaccurate, the resulting inferences are misleading. The advantages of the nonparametric approach are that there is no need to impose any assumption on the conditional mean functions or the effects of unobservables. The disadvantages are that some sort of conditional independence assumption have to hold conditional on some confounding variables. Hence it only takes into account the issues of selection on observables; neither is it feasible to estimate the impact of each observable factor. In other words, the advantages of the parametric approaches are the disadvantages of nonparametric approach. The disadvantages of the parametric approach are the advantages of the nonparametric approach.

9.6.4 Panel Data Approach

Panel data contains information over time for a number of individuals. Some of the observed individuals could be receiving treatment for part of the observed periods and no treatment for the rest of the observed periods. Some could be receiving treatment and some no treatment for the whole sample periods. The information on interindividual differences and intraindividual dynamics could lessen the restrictions imposed on the adjustment approaches using cross-sectional data alone.

9.6.4.1 Parametric Approach

One of the common assumptions using cross-sectional data is to assume that the observable factors, $\mathbf{x}$, are orthogonal to the impact of unobservable factors, ϵ^0 and ϵ^1 (e.g., (9.6.1) and (9.6.2)), even if it allows the joint dependence of $(\epsilon^1, \epsilon^0, d)$. However, the impact of unobservable factors could be correlated with observable factors, $\mathbf{x}$. Panel data allow us to control the correlation between (ϵ^0, ϵ^1) and $\mathbf{x}$, in addition to the correlation between (ϵ^0, ϵ^1) and d. For

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

instance, suppose that the outcome equations and participation equation are of the form

$$y_{it}^{1*} = \mathbf{x}_{it}'\boldsymbol{\beta}_1 + \epsilon_{it}^1, \tag{9.6.35}$$

$$y_{it}^{0*} = \mathbf{x}_{it}'\boldsymbol{\beta}_0 + \epsilon_{it}^0, \tag{9.6.36}$$

$$d_{it} = 1(\mathbf{x}_{it}'\gamma + v_{it} > 0), \tag{9.6.37}$$

where

$$\epsilon_{it}^j = \alpha_i^j + u_{it}^j, \quad j = 0, 1, \tag{9.6.38}$$

and u_{it}^j is i.i.d. with mean 0 and constant variance. If the correlations between ϵ_{it}^j and d_{it} are not confined to the individual specific components, α_i^j with d_{it}, but also the individual time-varying component u_{it}^j so $E(u_{it}^j v_{it}) \neq 0$, the panel data fixed-effects sample selection estimators of Kyriazidou (1997), Honoré (1992), etc. (as summarized in Chapter 8.4) can be used to control the impact of unobserved heterogeneity, α_i^1, α_i^0, and estimate the treatment effects (e.g., Hsiao, Shen, Wang, and Weeks (2007, 2008)).[13]

9.6.4.2 *Nonparametric Approach*

9.6.4.2 *(i) Difference-in-Difference Method*

As discussed in Section 9.6.3, one of the critical assumption using the nonparametric approach is to assume conditional independence between the outcomes (y_1, y_0) and participation, d, conditional on $\mathbf{x}$. To avoid the issue of the curse of dimensionality, Rosenbaum and Rubin (1983) propose a propensity score matching method. However, the propensity score matching adjustment to control the bias induced by selection on observables depends critically on the correct specification of the propensity score, Prob $(d_i = 1 \mid \mathbf{x}_i)$. With panel data, one can avoid the specification of the propensity score, Prob $(d_i = 1 \mid \mathbf{x}_i)$ if under the assumption that there is no selection bias and the impacts due to changes in x over time are the same between the treatment group (those who received the treatment) and the control group (those who did not receive the treatment) through a difference-in-difference method (Imbens and Angrist 1994).

Assume a panel begins with all the individuals in the control group (i.e., no treatment). At some time during the sample span, some individuals received treatment at time t and no treatment at time s, and some individuals neither received treatment at time t, nor at time s. Let $\psi = \{i \mid d_{it} = 1, d_{is} = 0\}$, and $\Psi = \{i \mid d_{it} = 0, d_{is} = 0\}$. Then the difference-in-difference estimate of the

[13] See Heckman and Hotz (1989) for other types of model specification tests.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

ATE is

$$\begin{aligned}\widehat{\text{ATE}} = {} & [E(y_{it} \mid i \in \psi) - E(y_{is} \mid i \in \psi)] \\ & - [E(y_{it} \mid i \in \Psi) - E(y_{is} \mid i \in \Psi)].\end{aligned} \tag{9.6.39}$$

For instance, the Northern Territory in Australia considered marijuana use a criminal act in 1995, but decriminalized it in 1996.[14] The Australian National Drug Strategy Household Surveys provide information about marijuana smoking behavior for residents of New Territories, New South Wales, Queensland Victoria, and Tasmania in 1995 and 2001; in all of them except New Territories (NT) it was nondecriminalized over this period. The percentage of smokers in NT in 1995 was 0.2342 and in 2001 it was 0.2845. The percentages of residents in nondecriminalized states in 1995 were 0.1423 and 0.1619 in 2001. The difference-in-difference estimate of the impact of decriminalization on marijuana usage is to raise the probability of smoking by

$$\begin{aligned}& \{(0.2845 - 0.2342) - (0.1619 - 0.1423)\} \\ & = 0.0503 - 0.0196 = 0.0307.\end{aligned} \tag{9.6.40}$$

9.6.4.2 (ii) Predicting Counterfactuals Using Control Group Information

The difference-in-difference method will provide a valid measurement of treatment effects under fairly restrictive assumptions. Namely, (1) there is no selection effect $E(\epsilon_i^0 \mid d_i) = E(\epsilon_i^0) = E(\epsilon_i^1) = E(\epsilon_i^1 \mid d_i) = 0$; (2) the marginal impacts of $\mathbf{x}$ are the same for those receiving treatment and not receiving treatment, $\frac{\partial E(y^{1*} \mid \mathbf{X})}{\partial \mathbf{X}} = \frac{\partial E(y^{0*} \mid \mathbf{X})}{\partial \mathbf{X}}$; and (3) changes in $\mathbf{x}$ for those in the treatment group and control group are the same, $E\{(x_{jt} - \mathbf{x}_{js}) \mid d_j = 1)\} = E\{(\mathbf{x}_{it} - \mathbf{x}_{is}) \mid d_i = 0\}$. However, with panel data, it is possible to relax these restrictive assumptions and still allow us to measure the treatment effects through the exploitation of correlations across individuals. Moreover, it also allows the treatment effects to vary over time.

Hsiao, Ching, and Wan (2012) propose to exploit the correlations across cross-sectional units to construct the counterfactuals. They assume the correlations across cross-sectional units are due to some common omitted factors. Decompose the outcomes of individual unit i into the sum of two components, the impact of K common factors that affect all individuals, $\mathbf{f}_t$, and an idiosyncratic component, $\alpha_i + \epsilon_{it}$, where α_i is fixed and ϵ_{it} is random with $E(\epsilon_{it}) = 0$ and $E(\epsilon_{it}\epsilon_{js}) = 0$ if $i \neq j$. The impact of common factors, $\mathbf{f}_t$, on individuals can be different for different individuals,

$$y_{it} = \alpha_i + \mathbf{b}_i' \mathbf{f}_t + \epsilon_{it}, \quad \begin{array}{c} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{9.6.41}$$

[14] Decriminalization does not mean smoking or possession of small amounts of marijuana is legal. It is still an offense to use or grow marijuana. An individual caught must pay a fine within a specified period to be eligible for the reduced penalty involving no criminal record or imprisonment (e.g., Damrongplasit and Hsiao 2009).

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Then the contemporaneous covariance between y_{it} and y_{jt} is given by

$$\text{Cov}\,(y_{it}, y_{jt}) = \mathbf{b}_i' E(\mathbf{f}_t\, \mathbf{f}_t')\mathbf{b}_j. \tag{9.6.42}$$

Stacking the $N \times 1$ y_{it} into a vector,

$$\mathbf{y}_t = B\,\mathbf{f}_t + \boldsymbol{\alpha} + \boldsymbol{\epsilon}_t, \tag{9.6.43}$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$, $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$, and B is the $N \times K$ factor loading matrix, $B = (\mathbf{b}_1, \ldots, \mathbf{b}_N)'$. Suppose all N units did not receive the treatment for $t = 1, \ldots, T_1$, that is, $\mathbf{y}_t = \mathbf{y}_t^{0*}$, but from time period $T_1 + 1$ onwards, the first unit received treatment, $y_{1t} = y_{1t}^{1*}$, $t = T_1 + 1, \ldots, T$, while the rest of individuals did not, $y_{it} = y_{it}^{0*}$, $t = 1, \ldots, T$ for $i = 2, \ldots, N$. As long as

$$E(\epsilon_{it} \mid d_{1t}) = 0, \quad i = 2, \ldots, N, \tag{9.6.44}$$

one can write

$$\begin{aligned} y_{1t}^{0*} &= E(y_{1t}^{0*} \mid \tilde{\mathbf{y}}_t) + \eta_{1t}, \quad t = 1, \ldots, T, \\ &= a + \mathbf{c}'\tilde{\mathbf{y}}_t + \eta_{1t}, \end{aligned} \tag{9.6.45}$$

where $\tilde{\mathbf{y}}_t = (y_{2t}, \ldots, y_{Nt})'$ and $E(\eta_{1t} \mid \tilde{\mathbf{y}}_t) = 0$. It is shown by Hsiao, Ching, and Wan (2012) that minimizing

$$\frac{1}{T_1}(\mathbf{y}_1^0 - \mathbf{e}a - Y\mathbf{c})' A (\mathbf{y}_1^0 - \mathbf{e}a - Y\mathbf{c}) \tag{9.6.46}$$

yields consistent estimates of a and $\mathbf{c}$, where $\mathbf{y}_1^0 = (y_{11}, \ldots, y_{1T_1})'$, $\mathbf{e}$ is a $T_1 \times 1$ vector of 1's, Y is a $T_1 \times (N-1)$ matrix of T_1 time series observations of $\tilde{\mathbf{y}}_t$, and A is a $T_1 \times T_1$ positive definite matrix. From the estimates $(\hat{a}, \hat{\mathbf{c}}')$, one can construct the predicted value of the first unit in the absence of treatment, y_{1t}^{0*}, by

$$\hat{y}_{1t}^{0*} = \hat{a} + \hat{\mathbf{c}}'\tilde{\mathbf{y}}_t, \quad t = T_1 + 1, \ldots, T. \tag{9.6.47}$$

The treatment effect on the first unit can then be estimated by

$$\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^{0*}, \quad t = T_1 + 1, \ldots, T. \tag{9.6.48}$$

The construction of the standard error of $\hat{y}_{1t}^{0*}$, $\sigma_{y_{1t}^0}$, follows from the standard prediction error formula. For instance, if η_{1t} is independently, identically distributed over time, then

$$\sigma_{y_{1t}^0}^2 = \sigma_{\eta_1}^2 [1 + (1, \tilde{\mathbf{y}}_t')(Y'Y)^{-1}(1, \tilde{\mathbf{y}}_t')']. \tag{9.6.49}$$

Hence, the confidence band for Δ_{1t} can be easily constructed as

$$\hat{\Delta}_{1t} \pm c\sigma_{y_{1t}^0}, \tag{9.6.50}$$

where c is chosen by the desired confidence level.

Cross-sectional data provide measurement of policy intervention as a once-and-for-all impact. Panel data allow the policy impact to be evolutionary. If Δ_{1t}

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

is serially correlated, but stationary, one can further model the time-varying treatment effects by an autoregressive moving average model using Box–Jenkins (1970) methodology

$$a(L)\Delta_{1t} = \mu + \theta(L)\eta_t \tag{9.6.51}$$

where L is the lag operator, η_t is an i.i.d. process with 0 mean and constant variance, and the roots of $\theta(L) = 0$ lie outside the unit circle. If the roots of $a(L) = 0$ all lie outside the unit circle, the treatment effect is stationary, and the long-term treatment effect is

$$\Delta_1 = a(L)^{-1}\mu = \mu^*. \tag{9.6.52}$$

Alternatively, one can estimate the long-run impact by taking the simple average of $\hat{\Delta}_{1t}$. When both T_1 and $(T - T_1)$ go to infinity,

$$\underset{(T-T_1)\to\infty}{\text{plim}} \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t} = \Delta_1 \tag{9.6.53}$$

The variance of (9.6.53) can be approximated by the heteroscedastic-autocorrelation consistent (HAC) estimator of Newey and West (1987).

Condition (9.6.44) makes no claim about the relationship between d_{1t} and ϵ_{1t}. They can be correlated. All we need is that the jth unit's idiosyncratic components ϵ_{jt} are independent of d_{1t} for $j \neq 1$. The approach can be viewed as a "measurement without theory" approach or a nonparametric approach.

The parameters, a and $\mathbf{c}$ can be obtained by regressing y_{1t} on $y_{it}, i = 2, \ldots, N$, for $t = 1, \ldots, T_1$. Often N is large. Using more or all cross-sectional units improves the within-sample fit, but does not necessarily yield more accurate post-sample prediction. One way to select the best combination of cross-sectional units to generate predicted y_{1t}^{0*} for $t = T_1 + 1, \ldots, T$ is to use one of the model selection criteria (e.g., AIC (Akaike (1973)), AICC (Hurvich and Tsai (1989)) or BIC (Schwarz (1978)). For instance, Hsiao, Ching, and Wan (2012) suggest the following two-step procedure:

Step 1: Selection the best predictor for y_{1t}^* using j cross-sectional units out of $(N - 1)$ cross-sectional units, denoted by $M(j)^*$ by R^2, for $j = 1, \ldots, N - 1$.

Step 2: From $M(1)^*, M(2)*, \ldots, M(N - 1)^*$, choose $M(m)^*$ in terms of some model selection criterion.

9.6.4.3 An Example – Measuring the Impact of the Closer Economic Partnership Arrangement on Hong Kong

Hong Kong signed Closer Economic Partnership Arrangement (CEPA) with Mainland China in June 2003 and started implementing its arrangement in January 2004. The CEPA aims to strengthen the linkage between Mainland China and Hong Kong by allowing Chinese citizens to enter Hong Kong as

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Table 9.2. *AICC selected model using data for the period 1993Q1–2003Q4*

	Beta	Std	T
Constant	−0.0019	0.0037	−0.524
Austria	−1.0116	0.1682	−6.0128
Italy	−0.3177	0.1591	−1.9971
Korea	0.3447	0.0469	7.3506
Mexico	0.3129	0.051	6.1335
Norway	0.3222	0.0538	5.9912
Singapore	0.1845	0.0546	3.3812
$R^2 = 0.931$			
AICC = −378.9427			

Source: Hsiao et al. (2012, Table 20).

individual tourists and liberalizing trade in services, enhancing cooperation in the area of finance, and promoting trade and investment facilitation and mutual recognition of professional qualifications. The implementation of CEPA started on January 1, 2004, where 273 types of Hong Kong products could be exported to the Mainland tariff free; another 713 types on January 1, 2005; 261 on January 1, 2006; and a further 37 on January 2007. Chinese citizens residing in selected cities are also allowed to visit Hong Kong as individual tourists, from 4 cities in 2003 to 49 cities in 2007, covering all 21 cities in Guangdong province.

Hsiao, Ching, and Wan (2012) tried to assess the impact of economic integration of Hong Kong with Mainland China on Hong Kong's economy by comparing what actually happened to Hong Kong's real GDP growth rates with what would have been if there were no CEPA with Mainland China in 2003. More specifically, they analyzed how these events have changed the growth rate of Hong Kong.

Because Hong Kong, by comparison, is tiny relative to other regions, Hsiao et al. (2012) believe that whatever happened in Hong Kong will have no bearing on other countries. In other words, they expect (9.6.44) to hold. Therefore, they use quarterly real growth rate of Australia, Austria, Canada, China, Denmark, Finland, France, Germany, Indonesia, Italy, Japan, Korea, Malaysia, Mexico, Netherlands, New Zealand, Norway, Philippines, Singapore, Switzerland, Taiwan Thailand, UK, and US to predict the quarterly real growth rate of Hong Kong in the absence of intervention. All the nominal GDP and CPI are from Organisation for Economic Co-operation and Development (OECD) Statistics, International Financial Statistics, and the CEIC database.

Using the AICC criterion, the countries selected are Austria, Italy, Korea, Mexico, Norway, and Singapore. Ordinary least-squares (OLS) estimates of the weights are reported in Table 9.2. Actual and predicted growth path from 1993Q1 to 2003Q4 are plotted in Figure 9.1. The availability of more

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

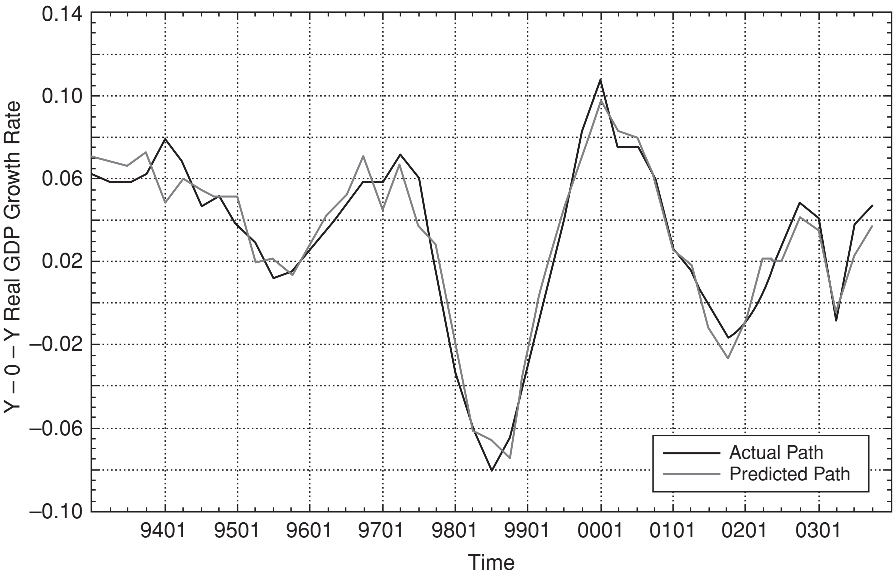

Figure 9.1. Actual and AICC predicted real GDP growth rate from 1993Q1 to 2003Q4. *Source:* Hsiao et al. (2012, Fig. 7).

Table 9.3. *Treatment effect for economic integration 2004Q1–2008Q1 based on AICC selected model*

	Actual	Control	Treatment
Q1-2004	0.077	0.0493	0.0277
Q2-2004	0.12	0.0686	0.0514
Q3-2004	0.066	0.0515	0.0145
Q4-2004	0.079	0.0446	0.0344
Q1-2005	0.062	0.0217	0.0403
Q2-2005	0.071	0.0177	0.0533
Q3-2005	0.081	0.0333	0.0477
Q4-2005	0.069	0.029	0.04
Q1-2006	0.09	0.0471	0.0429
Q2-2006	0.062	0.0417	0.0203
Q3-2006	0.064	0.025	0.039
Q4-2006	0.066	0.0009	0.0651
Q1-2007	0.055	−0.0101	0.0651
Q2-2007	0.062	0.0092	0.0528
Q3-2007	0.068	0.0143	0.0537
Q4-2007	0.069	0.0508	0.0182
Q1-2008	0.073	0.0538	0.0192
MEAN	0.0726	0.0323	0.0403
STD	0.0149	0.0213	0.016
T	4.8814	1.5132	2.5134

Source: Hsiao et al. (2012, Table 21).

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

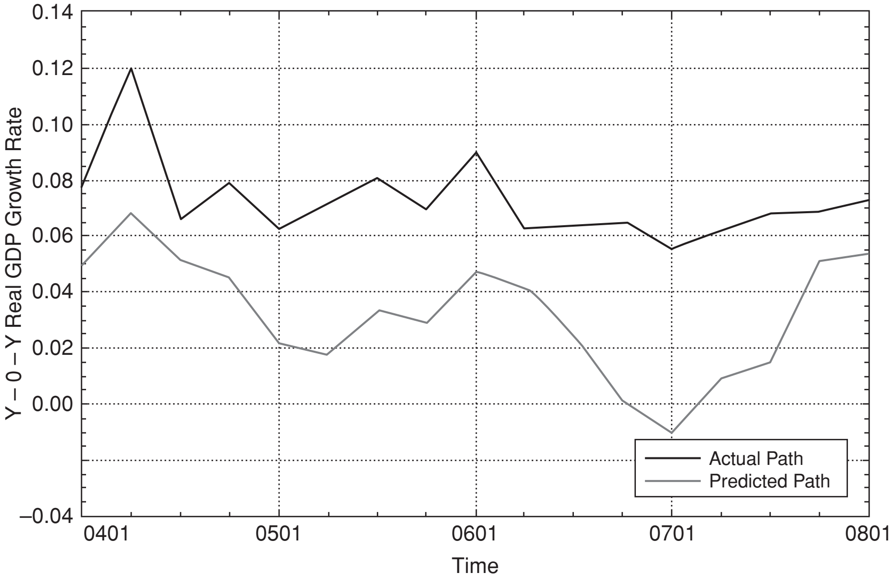

Figure 9.2. AICC – Actual and counterfactual real GDP growth rate from 2004Q1 to 2008Q1. *Source:* Hsiao et al. (2012, Fig. 8).

preintervention period data appears to allow more accurate estimates of the country weights and better tracing of the preintervention path. The estimated quarterly treatment effects are reported in Table 9.3. The actual and predicted counterfactual for the period 2004Q1 to 2008Q1 are presented in Figure 9.2.

Table 9.4. *AIC selected model using data for the period 1993Q1–2003Q4*

	Beta	Std	T
Constant	−0.003	0.0042	−0.7095
Austria	−1.2949	0.2181	−5.9361
Germany	0.3552	0.233	1.5243
Italy	−0.5768	0.1781	−3.2394
Korea	0.3016	0.0587	5.1342
Mexico	0.234	0.0609	3.8395
Norway	0.2881	0.0562	5.1304
Switzerland	0.2436	0.1729	1.4092
Singapore	0.2222	0.0553	4.0155
Philippines	0.1757	0.1089	1.6127
$R^2 = 0.9433$			
AIC = −385.7498			

Source: Hsiao et al. (2012, Table 22).

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

Table 9.5. *AIC–Treatment effect for economic integration 2004Q1–2008Q1 based on AIC selected model*

	Actual	Control	Treatment
Q1-2004	0.077	0.0559	0.0211
Q2-2004	0.12	0.0722	0.0478
Q3-2004	0.066	0.0446	0.0214
Q4-2004	0.079	0.0314	0.0476
Q1-2005	0.062	0.0121	0.0499
Q2-2005	0.071	0.0126	0.0584
Q3-2005	0.081	0.0314	0.0496
Q4-2005	0.069	0.0278	0.0412
Q1-2006	0.09	0.0436	0.0464
Q2-2006	0.062	0.0372	0.0248
Q3-2006	0.064	0.0292	0.0348
Q4-2006	0.066	0.0122	0.0538
Q1-2007	0.055	0.0051	0.0499
Q2-2007	0.062	0.0279	0.0341
Q3-2007	0.068	0.0255	0.0425
Q4-2007	0.069	0.0589	0.0101
Q1-2008	0.073	0.062	0.011
Mean	0.0726	0.0347	0.0379
Std	0.0149	0.0193	0.0151
T	4.8814	1.7929	2.5122

Source: Hsiao et al. (2012, Table 23).

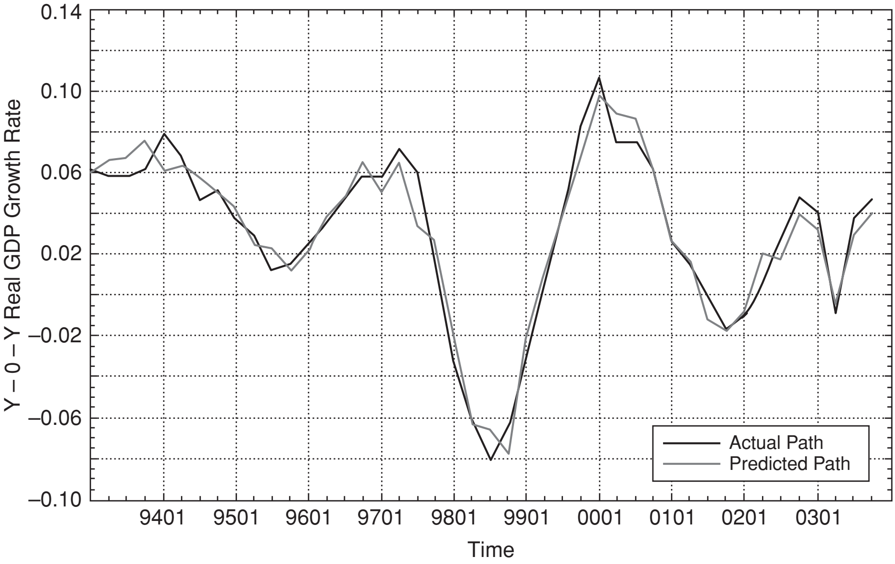

Figure 9.3. Actual and AIC predicted real GDP growth rate from 1993Q1 to 2003Q4. *Source:* Hsiao et al. (2012, Fig. 10).

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press

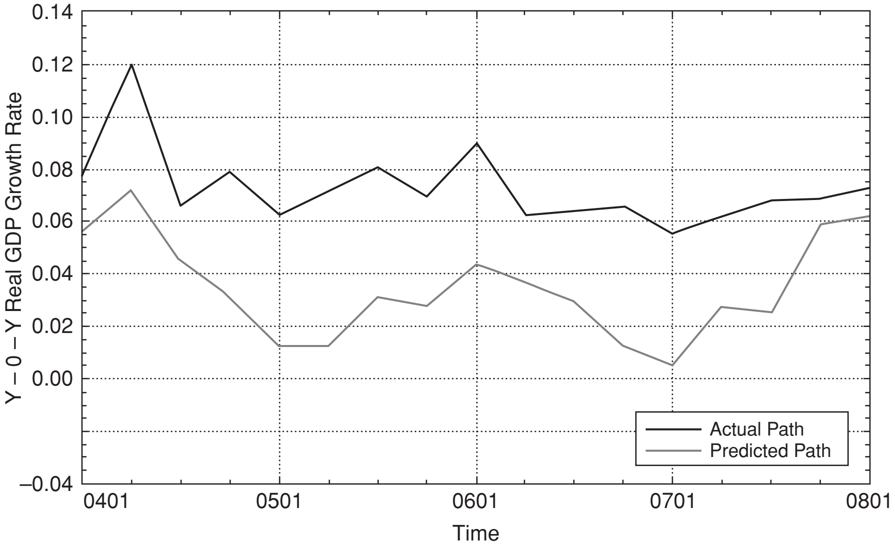

Figure 9.4. AIC – Actual and counterfactual real GDP growth rate from 2004Q1 to 2008Q1. *Source:* Hsiao et al. (2012, Fig. 11).

Using the AIC criterion, the selected group consists of Austria, Germany, Italy, Korea, Mexico, Norway, Philippines, Singapore, and Switzerland. The OLS estimates of the weights are in Table 9.4 and the estimated quarterly treatment effects are in Table 9.5. The pre- and post-intervention actual and predicted outcomes are plotted in Figures 9.3 and 9.4. It is notable that even though the two models use different combinations of countries, both groups of countries trace closely the actual Hong Kong path before the implementation of CEPA (with R^2 above .93). It is also quite remarkable that the post-sample predictions closely matched the actual turning points at a lower level for the treatment period even though no Hong Kong data were used. The CEPA effect at each quarter was all positive and appeared to be serially uncorrelated. The average actual growth rate from 2004Q1 to 2008Q1 is 7.26 percent. The average projected growth rate without CEPA is 3.23 percent using the group of countries selected by AICC and 3.47 percent using the group selected by AIC. The estimated average treatment effect is 4.03 percent with a standard error of 0.016 based on the AICC group and 3.79% with a standard error of 0.0151 based on the AIC group. The t-statistic is 3.5134 for the former group and 3.5122 for the latter group. Either set of countries yields similar predictions and highly significant CEPA effects. In other words, through liberalization and increased openness with Mainland China, the real GDP growth rate of Hong Kong is raised by more than 4 percent compared to the growth rate had there been no CEPA agreement with Mainland China.

https://doi.org/10.1017/CBO9781139839327.012 Published online by Cambridge University Press