Chapter 2

# Specification and Estimation

An important class of spatial models can be traced back to the work of Cliff and Ord (1973, 1981). The models they considered have since been generalized and extended in a variety of directions.[1] In Section 2.1 we present and discuss a general linear form of a model that is specified in terms of a spatial lag of the dependent variable, spatially lagged exogenous regressors, and a spatially autocorrelated error term. Section 2.1.1 introduces the important concept of triangular arrays which is a distinctive characteristic of spatial models compared to time series analysis. More details will be given below but, essentially, the idea is that the sequence of observations on the dependent variable, as well as the other model variables, must be indexed not only by the order of the observations, but also by the sample size. Triangular arrays rarely occur in time series models. Sections 2.1.2 and 2.1.3 discuss the definition of the parameter space and various forms of normalization of the model, whereas Section 2.1.4 introduces an important condition which is assumed in the derivation of many properties of spatial estimators. Sections 2.2, 2.3, and 2.4 are the core of this chapter since they deal with the estimation of cross-sectional spatial models. Section 2.2 deals with estimation issues in various special cases of the general model. A discussion of the model assumptions and the development of a generalized moments (GM) estimator of a parameter in the disturbance process are also provided in this section. Section 2.3 presents an instrumental variable estimator of the regression parameters of the general model. Finally, in Section 2.4 the maximum likelihood approach is reviewed. Sections 2.5 and 2.6 conclude the chapter. In Section 2.5 we show that when the spatial weighting matrices are the same in the general model and there are **no** regressors, there is an identification problem concerning the spatial parameters. We also show that if there are regressors in the models, there is no such problem. In Section 2.6 we make known that certain time series procedures should not be used in a spatial framework.

[1] Among others, see Kelejian and Prucha (2004, 2007a,b, 2010a), Lee (2003, 2004, 2007), Lee and Liu (2010), Lee et al. (2010), Baltagi and Li (2004), Baltagi et al. (2014a), Fingleton and Le Gallo (2008), Fingleton (2008a,b), Mutl and Pfaffermayr (2011), Piras (2013), Elhorst (2005), and references cited therein.

**Spatial Econometrics.** http://dx.doi.org/10.1016/B978-0-12-813387-3.00002-0

## 2.1 THE GENERAL MODEL

Let $y' = \left[y_1, ..., y_N\right]$, where $y_i,\ i = 1, ..., N$ is the dependent variable corresponding to the $i$th unit and consider the following model:

$$\begin{aligned} y_i &= a + X_{i.}B_1 + \rho_1\left(W_{i.}y\right) + \left(W_{i.}X\right)B_2 + u_i, \\ u_i &= \rho_2(W_{i.}u) + \varepsilon_i, \quad |\rho_1| < 1,\ |\rho_2| < 1 \end{aligned} \tag{2.1.1}$$

where $X_{i.}$ is a $1 \times k$ row vector of observations on exogenous explanatory variables whose values vary over the units, $W_{i.}$ is a $1 \times N$ row vector of observations on the $i$th row of an observable and exogenous $N \times N$ weighting matrix defined below,[2] $u_i$ is a disturbance term, $\varepsilon_i$ is a random term which is i.i.d. $\left(0, \sigma_\varepsilon^2\right)$, $B_1$ and $B_2$ are $k \times 1$ parameter vectors, and $a$, $\rho_1$, and $\rho_2$ are scalar parameters. The term $\varepsilon_i$ is typically refereed to as an innovation in the error process.

The model in (2.1.1) contains a fair amount of spatial spillovers. For example, the value of the dependent variable corresponding to the $i$th unit (which is typically a cross-sectional unit) depends upon the within unit effect of the regressor vector, $X_{i.}$, as well as the values of the dependent variable and the exogenous regressors in neighboring units. Similarly, the model also allows for spillovers in the disturbance terms by the second line of (2.1.1).

Denote the $N \times N$ weighting matrix as

$$W = \begin{bmatrix} W_{1.} \\ \vdots \\ W_{N.} \end{bmatrix}$$

so that the model in (2.1.1) can be written in stacked form as

$$\begin{aligned} y &= ae_N + XB_1 + \rho_1\left(Wy\right) + \left(WX\right)B_2 + u, \\ u &= \rho_2\left(Wu\right) + \varepsilon \end{aligned} \tag{2.1.2}$$

where $e_N$ is an $N \times 1$ vector of unit elements, and

$$\begin{aligned} X &= [X'_{1.}, ..., X'_{N.}]', \\ u &= [u_1, ..., u_N]', \\ \varepsilon &= [\varepsilon_1, ..., \varepsilon_N]'. \end{aligned}$$

For future reference, we note that the model in (2.1.2) is sometimes referred to as the spatial Durbin model; if $B_2 = 0$ and $\rho_2 = 0$, the model is sometimes

[2] In Chapter 13 we will consider the case in which the weighting matrix is endogenous.

called a spatial lag model; if $\rho_1 = 0$ and $B_2 = 0$, the model is sometimes called a spatial error model.

Denote the $i$th column of $X$, which is the $N \times 1$ vector of observations on the $i$th exogenous variable, as $X_{.i}, i = 1, ..., k$. By matrix multiplication, the term $WX$ in (2.1.2) can be expressed as

$$\begin{aligned} WX &= W[X_{.1}, ..., X_{.k}] \\ &= [WX_{.1}, WX_{.2}, ..., WX_{.k}]. \end{aligned} \tag{2.1.3}$$

In a time series setting, the time lag of a variable, say $z_t$, is the value of that same variable at an earlier point in time, e.g., $z_{t-j}$, $j > 0$. Now note that the $r$th element of the $N \times 1$ vector $WX_{.1}$ in (2.1.3), namely $(WX_{.1})_r$, is

$$(WX_{.1})_r = \sum_{j=1}^{N} w_{r,j} x_{j1} \tag{2.1.4}$$

where $x_{j1}$ is the $j$th element of $X_{.1}$. Note that $(WX_{.1})_r$ is a weighted sum of the values of that **same** first exogenous variable over neighboring units. Also recall that the diagonal elements of weighting matrices are zero. For these reasons, the vector $WX_{.1}$ is often referred to as the spatial lag of the first exogenous variable, $X_{.1}$; generalizing, $WX_{.j}$ is the spatial lag of the $j$th exogenous regressor.

The model (2.1.2) contains the spatial lags of all of the exogenous variables. Typically researchers consider the spatial lags of only a subset of the exogenous variables, e.g., perhaps only $WX_{.2}$ and $WX_{.5}$.

As an example, consider a model in which the dependent variable is the price of all the transactions involving single family houses in the Washington, DC area in a specific year. To explain the price of a single family house, the structural characteristics of the house (such as number of bedrooms, number of bathrooms, square footage of the living area, age, presence of a garage, whether the house has a fireplace, the construction material, etc.) along with the neighborhood characteristics (median income, percentage of population over 65, etc.) and amenities (distance from central business district, presence of parks, lakes, or rivers, accessibility, views, etc.) are generally considered good predictors. The model postulated above could include such additional variables: the spatial lag of the dependent variable as well as the spatial lag of some, or all, of the exogenous variables. A possible explanation for including the spatial lag of the price should be evident. In determining the market price of a house, economic agents look at the price of comparable houses in the same neighborhood. The rational to include the spatial lag of one (or more) of the explanatory variables is also quite intuitive. As an example consider the age variable. A new construction surrounded by older houses is likely to be, *ceteris paribus*, less expensive than

$$\begin{bmatrix} y_{11}, & y_{12}, & y_{13}, & y_{14}, & \text{etc.} \\ & y_{22}, & y_{23}, & y_{24}, & \\ & & y_{33}, & y_{34}, & \\ & & & y_{44}, & \end{bmatrix}$$

**FIGURE 2.1.1** Appearance of a triangular array

a new construction surrounded by newer houses. The spatial lag of the age variable could account for this.[3] The spatial lag of the disturbance term is supposed to capture remaining exogenous spatially correlated effects that are omitted from the model either because data on certain variables are not available, or simply they may not even be known.

### 2.1.1 Triangular Arrays

More formal assumptions of the model in (2.1.1) are presented later. In this section we review an important issue related to spatial models. We start by assuming that the model can be solved for $y$ in that $(I_N - aW)$ is nonsingular for all $|a| < 1$. More on this is said below. Given this nonsingularity, the reduced form of the model can be written as follows:

$$\begin{aligned} y &= (I_N - \rho_1 W)^{-1} [ae_N + XB_1 + WXB_2 + u] \qquad (2.1.1.1) \\ &= (I_N - \rho_1 W)^{-1} [ae_N + XB_1 + WXB_2] + \\ &(I_N - \rho_1 W)^{-1} (I_N - \rho_2 W)^{-1} \varepsilon. \end{aligned}$$

We now note something about spatial models which may be somewhat unexpected. The solution of the model in (2.1.2), given in (2.1.1.1), involves the elements of the $N \times N$ matrices $(I_N - \rho_1 W)^{-1}$ and $(I_N - \rho_2 W)^{-1}$. These elements will generally depend upon the size of the sample. Since, via (2.1.1.1), the elements of the dependent vector $y$ depend upon these inverses, the implication is that each of the elements of the vector $y$ will generally change as the sample size changes even though the model remains the same. In other words, the first value of $y$ when $N = 5$ will not necessarily be the same as when $N = 6$. For this reason, in a formal analysis, the elements of the dependent vector would have two indices: one denoting the observation number, and the other denoting the sample size. For example, $y_{i,j}$ would denote the $i$th element of $y$ when the sample size is $j$, and in general, $y_{i,j} \neq y_{i,s}$ if $j \neq s$. Since all of the elements of the dependent vector would be subscripted with the sample size, the dependent vector itself would be so subscripted, $y_N$. If these data are plotted, a triangular appearance, such as that in Fig. 2.1.1, is revealed.

[3] Many applied works also include the square of the age variable to capture a "vintage" effect.

Because of this triangular appearance, the values of the dependent variable will be described as a triangular array.

Here lies an important difference between spatial and time series data. In a time series study based on annual data, the sample increases as additional years pass. The implication is that data corresponding to earlier years do not change as data on later years become available except, perhaps, for error corrections, e.g., the unemployment rate in 2005 does not change when the data for 2006 becomes available. As will become evident in latter discussions, inference in many spatial models relates to results obtained from large sample theory. This means that one typically determines estimation results for the limiting case in which $N \to \infty$, and then assumes that those results are a reasonable approximation for the finite sample which is available. In this case, the sample size is often viewed as "increasing" with (hypothetically) increasingly larger random samples. In such a scenario, each observation on the elements of the regressor matrix, $X$, would also be expected to change in a triangular array fashion. For example, if the first value of a regressor is the income level of the first observed family, that "first" value need not be the same when $N = 20$ as when $N = 30$. For instance, if random samples are taken, the first family observed when $N = 20$ may not even be in the random sample taken when $N = 30$. Therefore, in formal studies, the regressor matrix would also be subscripted with the sample size, e.g., $X_N$.

At this point we do not index the variables in the model for simplicity of notation. We will, however, account for the triangular nature of the sample in spatial models.

### 2.1.2 Geršgorin's Theorem[4] and Weighting Matrices

As suggested in (2.1.1.1), the model in (2.1.2) is complete in that it can be solved for the dependent vector $y$ in terms of $X$, $WX$, and $\varepsilon$ **if** the matrix $(I_N - aW)$ is nonsingular for all $|a| < 1$. Consistent with this, the parameter space for $\rho_1$ and $\rho_2$ are often taken as $|\rho_1| < 1$ and $|\rho_2| < 1$. Note that this parameter space is continuous for both $\rho_1$ and $\rho_2$ in a known region. Below we show that, under very reasonable conditions, the weighting matrix can always be normalized in such a way that $(I_N - aW)$ is nonsingular for all $|a| < 1$. Furthermore, the normalization will not effect the size, sign, or significance of the model parameters. These "normalization" results are important because, under the very reasonable conditions, one can always work with a spatial model for which $(I_N - aW)^{-1}$ exists for all $|a| < 1$. This is important for a number of reasons. First, in many spatial models inference is based on large sample properties of the estimators

[4] A formal statement of this theorem is given in Horn and Johnson (1985, pp. 344–346). See also Kelejian and Prucha (2010a).

which depend upon the inverse of $(I_N - aW)$ in a continuous and known region for $a$, among other things. Second, important measures of spatial model spillovers, described in detail in Chapter 3, depend upon the existence of the inverse of $(I_N - aW)$ in a continuous interval for $a$. Third, many spatial models are estimated by procedures which would encounter considerable "complications" if the space for $a$ for which $(I_N - aW)^{-1}$ exists is not continuous.

We now consider conditions for the nonsingularity of a matrix $(I_N - aW)$, where $a$ is a constant.

**Note 1.** If $W$ is row normalized, $(I_N - aW)$ is singular at $a = 1$.

Indeed, let $e_N = (1, 1, ..., 1)'_{N\times 1}$. Then

$$(I_N - W)\,e_N = e_N - We_N = e_N - e_N = 0. \tag{2.1.2.1}$$

This is why in (2.1.1) the parameter space for $\rho_1$ and $\rho_2$ does not include 1.

**Note 2.** If $W$ is row normalized, then $(I_N - aW)^{-1}$ exits for all $|a| < 1$.

**Proof.** We prove this by relying on the theorem by Geršgorin and a matrix result concerning characteristic roots by Ord (1975). Let $A_{N\times N}$ have elements $a_{ij}$. Let $R_i, i = 1, ..., N$ be the sum of the absolute values of the elements in the $i$th row of $A$ with the exception of its diagonal elements; similarly, let $C_j$, $j = 1, ..., N$ be the sum of the absolute values in the $j$th column of $A$, again with the exception of the diagonal elements:

$$R_i = \sum_{j=1, j\neq i}^{N} \left|a_{ij}\right|, \quad C_j = \sum_{i=1, i\neq j}^{N} \left|a_{ij}\right|. \tag{2.1.2.2}$$

Then Geršgorin's theorem implies that each eigenvalue, $\lambda_i$, of $A$ satisfies at least one of the inequalities relating to the row sums:

$$|\lambda - a_{ii}| \leq R_i, \quad i = 1, ..., N. \tag{2.1.2.3}$$

Furthermore, each root satisfies at least one of the inequalities relating to the column sums:

$$\left|\lambda - a_{jj}\right| \leq C_j, \quad j = 1, ..., N. \tag{2.1.2.4}$$

Let us now apply this statement to a weighting matrix $W$ with diagonal elements $w_{ii} = 0$. Let

$$\begin{aligned} r &= \max_i \textstyle\sum_j \left|w_{ij}\right| & c &= \max_j \textstyle\sum_i \left|w_{ij}\right| \\ &= \max_i R_i, & &= \max_j C_j. \end{aligned} \tag{2.1.2.5}$$

Given (2.1.2.2)–(2.1.2.5), the roots of $W$ must satisfy the conditions

$$|\lambda_i| \leq r, |\lambda_i| \leq c, \quad i = 1, ..., N. \tag{2.1.2.6}$$

If $W$ is row normalized, $r = 1$. From (2.1.2.6) it then follows that when $W$ is row normalized $|\lambda_i| \leq 1, i = 1, ..., N$.

Now let $Q$ be the matrix that triangularizes $W$, i.e.,

$$QWQ^{-1} = D_\lambda, \quad D_\lambda = \begin{bmatrix} \lambda_1 & & \lambda_{ij} \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}. \tag{2.1.2.7}$$

Thus, applying the result suggested by Ord (1975), $|D_\lambda| = \lambda_1 \ldots \lambda_N$ and we have[5]

$$\begin{aligned} |I_N - aW| &= \left|QQ^{-1}(I_N - aW)\right| \\ &= \left|Q(I_N - aW)Q^{-1}\right| \\ &= |I_N - aD_\lambda| = (1 - a\lambda_1)\ldots(1 - a\lambda_N) \\ &\neq 0 \end{aligned} \tag{2.1.2.8}$$

if $|a| < 1$, then $|a\lambda_i| \leq |a| < 1$. Thus, $(I_N - aW)$ is nonsingular if $|a| < 1$.

**Note 3.** Finally, we note one more point, which appears in the literature, and corresponds to a special case; see Anselin (1988). Let $W$ again be a weighting matrix, with $w_{ii} = 0$, $i = 1, \ldots, N$, and assume that all of the roots of $W$ are real,[6] and that $W$ is not row normalized. Let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and smallest roots of $W$, respectively. Assume, as will typically be the case if all of the roots are real, that $\lambda_{\max} > 0$ and $\lambda_{\min} < 0$. Then $(I_N - aW)$ is nonsingular for all

$$\lambda_{\min}^{-1} < a < \lambda_{\max}^{-1}. \tag{2.1.2.9}$$

**Proof.** First note that $(I_N - aW)$ is nonsingular for $a = 0$. If $a \neq 0$, we have as before

$$\begin{aligned} |I_N - aW| &= |I_N - aD_\lambda| \\ &= (1 - a\lambda_1)(1 - a\lambda_2)\cdots(1 - a\lambda_N) \end{aligned} \tag{2.1.2.10}$$

so $I_N - aW$ is nonsingular unless $a$ is equal to the inverse of a root, i.e., $\lambda_1^{-1}, \ldots, \lambda_N^{-1}$, or $a^{-1}$ is equal to a root $\lambda_1, \ldots, \lambda_N$.

[5] Recall that if $A$ and $B$ are two $N \times N$ matrices, then $|AB| = |A||B|$.

[6] This is a very strong assumption which will be evident from the results below.

Thus if

$$a^{-1} < \lambda_{\min}$$

or if

$$a^{-1} > \lambda_{\max}$$

then

$$(I_N - aW) \quad \text{is nonsingular.}$$

But

$$\begin{aligned} &\text{if } a^{-1} > \lambda_{\max} \text{ then } a < \lambda_{\max}^{-1}; \\ &\text{if } a^{-1} < \lambda_{\min} \text{ then } a > \lambda_{\min}^{-1}, \end{aligned} \tag{2.1.2.11}$$

and so (2.1.2.9) holds.

Although this result appears in various parts of the literature, it is of limited importance because the roots of a nonsymmetric matrix will typically not all be real, e.g., some will be complex. Complex numbers are not ordered and so the maximum and minimum roots cannot be identified. For example, let $i = (-1)^{1/2}$. Then one cannot say that $5i > 4i$ since multiplying across by $i$ yields $-5 > -4$, which is not true. If multiplication by $i$ reverses the inequality, then multiplying $5i > 4i$ across by $i$ repeatedly yields $5 < 4$, which is also not true.[7]

## 2.1.3 Normalization to Ensure a Continuous Parameter Space

The discussion above indicates that if $W$ is row normalized, the parameter space for $\rho_1$ and $\rho_2$ specified in (2.1.1) is such that the inverses in (2.1.1.1) exist. If $W$ is *not* row normalized $(I_N - aW)$ will generally be singular for certain values of $|a| < 1$. As shown in Kelejian and Prucha (2010a), in this case the model can be easily transformed to obtain a reparameterized form of the weighting matrix $W$, say $W^*$, such that $(I_N - aW^*)$ is nonsingular for all $|a| < 1$. This transformed model is such that the interpretations, and estimated statistical significance of the regression parameters are not effected. This is demonstrated below. Thus for purposes of analysis, one need not consider the case for which the model has a weighting matrix, say $W^+$, which is such that $(I_N - aW^+)$ is singular for certain values of $|a| < 1$.

[7] Although one cannot say that $a + bi$ is greater than or less than $c + di$, one could order complex roots by their absolute values, $(a^2 + b^2)^{1/2}$ and $(c^2 + d^2)^{1/2}$. However, the above proof would not go through.

Let $\alpha = \min(r, c)$, where $r$ and $c$ are defined in (2.1.2.5) with respect to $W$ above. Then, assuming that the elements of $W$ are nonnegative we will first show that

$$(I_N - aW) \text{ is nonsingular for all } |a| < 1/\alpha. \tag{2.1.3.1}$$

Given (2.1.3.1), the parameter space for $\rho_1$ and $\rho_2$ could be taken as $|\rho_1| < 1/\alpha$ and $|\rho_2| < 1/\alpha$.

**Proof.** As we mentioned earlier, if $a = 0$, then $(I_N - aW)$ is nonsingular. Consider now the case in which $a \neq 0$. In this case $|I_N - aW| = 0$ implies

$$\begin{aligned} \left| \left(\frac{1}{a}\right) I_N - W \right| &= 0 \text{ or} \qquad (2.1.3.2) \\ \left| W - \left(\frac{1}{a}\right) I_N \right| &= 0. \end{aligned}$$

Therefore, $(I_N - aW)$ is singular if $\left(\frac{1}{a}\right)$ is equal to a root of $W$. In addition, observe that since the roots of $W$ are such that

$$|\lambda_i| \leq r\,, \quad |\lambda_i| \leq c, \tag{2.1.3.3}$$

then

$$|\lambda_i| \leq \min(r, c) = \alpha. \tag{2.1.3.4}$$

As a result,

$$\begin{aligned} |I_N - aW| &\neq 0 \text{ if} \qquad (2.1.3.5) \\ \left|\frac{1}{a}\right| &> \min(r, c) = \alpha \text{ or} \\ |a| &< \frac{1}{\alpha}. \end{aligned}$$

The result in (2.1.3.1) follows.

This is an important result because a model which has a weighting matrix which is not row normalized can always be normalized in such a way that the inverse needed to solve the model will exist in an easily established region. For example, suppose $W$ is not row normalized. Then the model[8]

$$\begin{aligned} y &= ae_N + XB + \rho_1 W y + \varepsilon \qquad (2.1.3.6) \\ &= ae_N + XB + (\rho_1 \alpha)\left(\frac{W}{\alpha}\right) y + \varepsilon \end{aligned}$$

[8] Note that $\alpha$ below will depend on $N$ and hence so will $\rho_1^*$ and $W^*$. For ease of notation, we do not indicate this dependence. We note, however, that for many models, $\alpha$ would be a known finite constant for all sample sizes $N$.

or

$$y = ae_N + XB + \rho_1^* W^* y + \varepsilon \tag{2.1.3.7}$$

$$\text{where } \rho_1^* = \rho_1 \alpha \ , \ \ W^* = \frac{W}{\alpha}, \text{ and}$$

$$\alpha = \min(c, r) : c = \max_j \sum_i |w_{ij}| ; \quad r = \max_i \sum_j |w_{ij}| ,$$

where $\alpha$, $c$, and $r$ relate to $W$. Note that $\alpha$ is easily determined. Also note that the maximum sum of the absolute values of the elements in the rows and columns of $(1/\alpha)W$ are $(1/\alpha)r$ and $(1/\alpha)c$, respectively. Thus, given (2.1.3.1),

$$|I_N - \rho_1^* W^*| \neq 0$$

since

$$\begin{aligned} |\rho_1^*| &< \frac{1}{\min\left(\frac{c}{\alpha}, \frac{r}{\alpha}\right)} \\ &< \frac{1}{\left(\frac{1}{\alpha}\right) \min(c, r)} = 1. \end{aligned}$$

Hence if the model is renormalized as

$$y = ae_N + XB + \rho_1^* W^* y + \varepsilon \tag{2.1.3.8}$$

and $\rho_1^*$ is taken to be the parameter, the inverse exists for all $|\rho_1^*| < 1$. One would then estimate $\rho_1^*$, and since $\rho_1^* = \rho_1 \alpha$, it is easy to estimate $\rho_1$ from

$$|\hat{\rho}_1| = \hat{\rho}_1^* / \alpha. \tag{2.1.3.9}$$

### 2.1.4 An Important Condition in Large Sample Analysis

The general model in (2.1.1) involves the parameters $\rho_1$ and $\rho_2$. Clearly, the estimation of this general model has to account for both of these parameters in some way. In the next sections we argue that, unless $\rho_2 = 0$, the disturbance terms are spatially correlated as well as heteroskedastic. Additionally, unless $\rho_1 = 0$, the model contains an endogeneity that must be accounted for. There are two ways of coping with these issues. One can use either maximum likelihood procedures or an instrumental variables procedure (IV). Hypothesis tests based on these procedures are based on large sample results. These large sample results typically rely on certain widely used matrix assumptions which may not be obvious to all readers and so they are discussed in this section. As we will indicate, these assumptions rule out certain models.

**1.** We will say that the row and column sums of an $N \times N$ matrix, $A$, are uniformly bounded in absolute value if

$$\max_i \sum_{j=1}^{N} \left|a_{ij}\right| \leq c_a, \tag{2.1.4.1}$$

$$\max_j \sum_{i=1}^{N} \left|a_{ij}\right| \leq c_a,$$

for all $N \geq 1$ where $c_a$ is a finite constant which does not depend on $N$.[9] We will also, on occasion, abbreviate reference to a matrix such as $A$ by saying that it is "absolutely summable." Note for future reference that a given column, say column $j$ in the matrix $A$ above, **cannot** satisfy the condition in (2.1.4.1) if the elements of that column are such that $\left|a_{ij}\right| > d \geq 0$ for all $i = 1, ..., N$ and $N \geq 1$. In this case the $j$th unit would be a central unit to which all units relate. Among other things, a unit could be central because of financial issues.

**2.** If $A$ and $B$ are $N \times N$ absolutely summable matrices, then so is $D = AB$. Because the proof of this statement is instructive it is given below.

**Proof.** The $(i, j)$th element of $D$ is, using evident notation,

$$d_{ij} = \sum_{r=1}^{N} a_{ir} b_{rj}. \tag{2.1.4.2}$$

Given (2.1.4.2), let $r_i$ be the $i$th row sum and note that

$$\begin{aligned} r_i = \sum_{j=1}^{N} \left|d_{ij}\right| &\leq \sum_{j=1}^{N} \sum_{r=1}^{N} |a_{ir}| \left|b_{rj}\right| \\ &= \sum_{r=1}^{N} \sum_{j=1}^{N} |a_{ir}| \left|b_{rj}\right| \\ &= \sum_{r=1}^{N} |a_{ir}| \sum_{j=1}^{N} \left|b_{rj}\right| \\ &\leq c_a c_b, \text{ for all } i = 1, ..., N \text{ and } N \geq 1. \end{aligned} \tag{2.1.4.3}$$

A similar demonstration will reveal that

$$\sum_{i=1}^{N} \left|d_{ij}\right| \leq c_a c_b, \text{ for all } j = 1, ..., N \text{ and } N \geq 1. \tag{2.1.4.4}$$

[9] As an illustration, if the maximum of the row and column sums are, respectively, 5 and 7 then they are both less than 7.

**3.** If $A$ is absolutely summable, its elements are bounded. This should be obvious.

**4.** If $A$ is absolutely summable and $Z_{N\times K}$ has **uniformly** bounded elements, then the elements of $Z'AZ$ are at most of order $N$, e.g., $O(N)$ see Section A.14 in Appendix A, which deals with large sample theory.

**Proof.** Let $Z_{ij}$ be the $(i, j)$th element of $Z$, and let $|Z_{ij}| \leq c_z$ for all $i, j$ and $N \geq 1$. Now consider the $(i, j)$th element of $Z'AZ$, say $\delta_{ij}$,

$$\begin{aligned}\delta_{ij} &= \sum_{r=1}^{N}\sum_{s=1}^{N} Z_{si}a_{sr}Z_{rj}, \qquad (2.1.4.5)\\ \left|\delta_{ij}\right| &\leq \sum_{r=1}^{N}\sum_{s=1}^{N} |Z_{si}|\,|a_{sr}|\,Z_{rj}| \\ &\leq c_z^2 \sum_{r=1}^{N}\sum_{s=1}^{N} |a_{sr}| \\ &\leq c_z^2 \sum_{r=1}^{N} c_a \\ &\leq c_z^2 c_a N = 0(N).\end{aligned}$$

## 2.2 ESTIMATION: VARIOUS SPECIAL CASES

In this section we turn to the estimation of various special cases of the general model presented above. For convenience, rewrite the model in (2.1.2) as

$$\begin{aligned} y &= ae_N + XB_1 + \rho_1 Wy + (WX)\,B_2 + u, \qquad (2.2.1)\\ u &= \rho_2\,(Wu) + \varepsilon, \quad |\rho_1| < 1,\ |\rho_2| < 1.\end{aligned}$$

Assume $(I_N - aW)^{-1}$ exits for $|a| < 1$. As we showed in the previous section, this is not a restrictive assumption since a model can always be normalized to ensure this condition.

### 2.2.1 Estimation When $\rho_1 = \rho_2 = 0$

In this case the model reduces to

$$\begin{aligned} y &= ae_N + XB_1 + (WX)\,B_2 + \varepsilon \qquad (2.2.1.1)\\ &= Z\gamma + \varepsilon,\\ Z &= (e_N, X, WX), \quad \gamma' = (a, B_1', B_2')\end{aligned}$$

where $X$ is $N \times k$, and we are assuming that the rank $(e_N, X, WX) = 2K + 1$ in order to rule out perfect multicolinearity. Clearly, in this framework $X$ does not include the constant term. Recall that in (2.1.1) that $\varepsilon_i$ are i.i.d. $\left(0, \sigma_\varepsilon^2\right)$ so that $\varepsilon \sim \left(0, \sigma_\varepsilon^2 I_N\right)$ and that $X$ and $W$ are nonstochastic. Note, however, that we have not assumed any particular distribution for $\varepsilon$, e.g., the normal.

Given the model in (2.2.1.1) the least squares estimator of $\gamma$ is

$$\begin{aligned} \hat{\gamma} &= (Z'Z)^{-1}Z'y \\ &= \gamma + (Z'Z)^{-1}Z'\varepsilon. \end{aligned} \tag{2.2.1.2}$$

Using (2.2.1.2) it follows that

$$\begin{aligned} E(\hat{\gamma}) &= \gamma + E\left[(Z'Z)^{-1}Z'\varepsilon\right] \\ &= \gamma + (Z'Z)^{-1}Z'E(\varepsilon) \\ &= \gamma \end{aligned} \tag{2.2.1.3}$$

so that $\hat{\gamma}$ is unbiased. Analogously, the variance–covariance matrix of $\hat{\gamma}$ is $VC_{\hat{\gamma}}$ where

$$\begin{aligned} VC_{\hat{\gamma}} &= E(\hat{\gamma} - \gamma)(\hat{\gamma} - \gamma)' \\ &= (Z'Z)^{-1}Z'E[\varepsilon\varepsilon']Z(Z'Z)^{-1} \\ &= (Z'Z)^{-1}Z'\sigma_\varepsilon^2 I_N Z(Z'Z)^{-1} \\ &= \sigma_\varepsilon^2 (Z'Z)^{-1}. \end{aligned} \tag{2.2.1.4}$$

However, the mean and variance covariance matrices given in (2.2.1.3) and (2.2.1.4) are not enough to test hypotheses concerning $\gamma$ because the distribution of $\hat{\gamma}$ is not known. Of course, in this simple model if one assumes that $\varepsilon$ is normally distributed, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_N)$, and also recalls that linear combinations of jointly normal variables are also normal, then via (2.2.1.2) $\hat{\gamma} \sim N(\gamma, \sigma_\varepsilon^2 (Z'Z)^{-1})$. However, since $\sigma_\varepsilon^2$ will typically not be known, this result is still not enough to test hypotheses.

A typical unbiased estimator of $\sigma_\varepsilon^2$ is $\hat{\sigma}_\varepsilon^2 = (y - Z\hat{\gamma})'(y - Z\hat{\gamma})/(N - 2K - 1)$. Given this, and the normality assumption, tests of hypotheses can then be based on the usual $t$-ratio, or $F$ tests. For example, let $\hat{\gamma}_i$ be the $i$th element of $\hat{\gamma}$, and let $\hat{\sigma}_{\hat{\gamma}_i}^2$ be the $i$th diagonal element of $\widehat{VC}_{\hat{\gamma}} = \hat{\sigma}_\varepsilon^2 (Z'Z)^{-1}$. Then, e.g., using evident notation the variable corresponding to $\hat{\gamma}_i$ would be significant at the (two tail) 5% level if

$$|\hat{\gamma}_i / \hat{\sigma}_{\hat{\gamma}_i}| > t_{N-2K-1}(0.975) \doteq 1.96,$$

where $t_{N-2K-1}(0.975)$ would be taken from a table of values on the $t$ distribution. Similarly, using evident notation, the joint hypothesis $H_0 : R\gamma = r$ against $H_1 : R\gamma \neq r$, where $R$ is a known $q \times 2K + 1$ matrix, $q < 2K + 1$, and $r$ is a known $q \times 1$ vector would be accepted at the 5% level if

$$(R\hat{\gamma} - r)[R\widehat{VC}_{\hat{\gamma}_i}R']^{-1}(R\hat{\gamma} - r) > F_{q,N-2K-1}(0.95).$$

In more realistic and complex models, the normality of the disturbance term will not be enough to determine the distribution of $\hat{\gamma}$. For this reason we give a simple illustration of the large sample approach. Specifically, we determine the large sample distribution of $\hat{\gamma}$ which, in turn, suggests an approximation to the small (or finite) sample distribution of $\hat{\gamma}$. Hypotheses can then be tested in terms of this small sample approximation. In the more complex models considered below, and in later chapters, the large sample procedures are essentially the same.

A complete set of formal assumptions for the general model in (2.2.1) is given in a later section. At this point we note that two of these assumptions, applied to (2.2.1.1), are that the elements of $Z$ are uniformly bounded in absolute value, and $N^{-1}Z'Z \to Q_{zz}$ where $Q_{zz}$ is a finite nonsingular matrix. In the analysis below recall that the elements in the error term in (2.2.1.1) are i.i.d. $(0, \sigma_\varepsilon^2)$.

The least squares estimator of $\gamma$ in (2.2.1.2) can be expressed as

$$N^{1/2}(\hat{\gamma} - \gamma) = N(Z'Z)^{-1}[N^{-1/2}Z'\varepsilon]. \tag{2.2.1.5}$$

Given the assumption that $N(Z'Z)^{-1} \to Q_{zz}^{-1}$, the central limit theorem given in Section A.15 of Appendix A yields

$$N^{-1/2}Z'\varepsilon \xrightarrow{D} N(0, \sigma_\varepsilon^2 Q_{zz}). \tag{2.2.1.6}$$

Given (2.2.1.5) and (2.2.1.6), and the result in (A.10.3) in Appendix A, it then follows that

$$\begin{aligned} N^{1/2}(\hat{\gamma} - \gamma) &\xrightarrow{D} N(0, \sigma_\varepsilon^2 Q_{zz}^{-1} Q_{zz} Q_{zz}^{-1}) \\ &= N(0, \sigma_\varepsilon^2 Q_{zz}^{-1}). \end{aligned} \tag{2.2.1.7}$$

Using the results in Section A.12 of Appendix A, small sample inference, and tests of hypotheses can be based on the approximation

$$\hat{\gamma} \simeq N[\gamma, \hat{\sigma}_\varepsilon^2 (Z'Z)^{-1}] \tag{2.2.1.8}$$

where $\hat{\sigma}_\varepsilon^2 = (y - Z\hat{\gamma})'(y - Z\hat{\gamma})/(N - \tau)$ and $\tau$ can be taken to be $(2K + 1)$, or zero. In both cases typical modeling assumptions will imply that $\hat{\sigma}_\varepsilon^2$ is consistent for $\sigma_\varepsilon^2$ since consistency is a large sample property. Note that the large sample distribution obtained in (2.2.1.7) does not require an assumed distribution of the random term $\varepsilon$.

In passing we note that the results above imply that $\hat{\gamma}$ is consistent. For example, (2.2.1.3) implies that $\hat{\gamma}$ is unbiased and its variance–covariance matrix in (2.2.1.4) is such that $\sigma_\varepsilon^2(Z'Z)^{-1} = \sigma_\varepsilon^2 N^{-1}[N(Z'Z)^{-1}] \to 0$ since $N(Z'Z)^{-1} \to Q_{ZZ}^{-1}$ which is a finite matrix. It then follows from Chebyshev's inequality in Section A.3 of Appendix A that $\hat{\gamma} \xrightarrow{P} \gamma$. This consistency result can also be obtained from (2.2.1.7) since $\hat{\gamma} - \gamma = N^{-1/2}[N^{1/2}(\hat{\gamma} - \gamma)] = N^{-1/2} 0_P(1) \to 0$; see the discussion of orders in probability in Section A.14 of Appendix A.

### Illustration 2.2.1.1: House value and its determinants

The data set used in this example is very well known among researchers in spatial econometrics since it has served as an example for many scientific contributions (see, e.g., Gilley and Pace, 1996, among others).[10] We estimate a model relating the logarithm of the median housing price in 506 Boston area communities (*price*) to various house and location characteristics such as the average number of *rooms*; the logarithm of the weighted distance (*dist*) of the community from five employment centers; a measure of per-capita crime in the community (*crime*); the concentration of nitrogen oxide (*nox*) in the air measured in parts per million; and a variable (*stratio*) measuring the average student–teacher ratio of schools in the community. Finally, we also include in the model the spatial lag of crime in neighboring communities (*wcrime*). The intuition is that being close to neighboring communities that have high level of crime reduces the median house value.

The OLS results from the estimated model are reported below:

$$\begin{aligned}\log\widehat{(price)} = {} & 2.049(0.159) - 0.875(0.101)\log(nox) - 0.272(0.039)\log(dist) \\ & - 0.036(0.005)\, stratio + 0.244(0.016)\, rooms \\ & - 0.009(0.002)\, crime - 0.016(0.002)\, wcrime\end{aligned}$$

where standard errors are in the parentheses.

The slope estimates all have the expected sign and are significantly different from zero. The coefficient estimate of the spatial lag of the crime variable is negative and statistically significant. It measures the semielasticity of price with respect to the per-capita crime in the neighboring communities.

[10] For simplicity, we only consider some of the variables generally included in the empirical specification.

### Illustration 2.2.1.2: A model of the murder rate

In the following example,[11] we relate the number of murders per 100,000 people (*mrdrte*) in each of the US 49 continental states to the state unemployment rate (*unem*) and the state total number of prisoners' executions over the past three years (*exec*). We also include the spatial lag of executions (*wexec*) in bordering states over the past three years. We are interested in seeing whether there is evidence for a deterrent effect of capital punishment. Our expectation on the coefficients is the following: The number of state executions should have a negative effect, and the unemployment rate should have a positive effect on the murder rate. We expect the coefficient of (*wexec*) to be negative. For example, the higher the number of executions in neighboring states, the lower the incentive to commit crime within the boundaries of those states.

The OLS results from the estimated model are reported below:

$$\begin{aligned}\widehat{mrdrte} &= -6.030(7.039) + 0.071(0.298)\, exec \\ &\quad + 2.384(1.061)\, unem - 0.182(0.529)\, wexec\end{aligned}$$

where standard deviations are in the parentheses.

We start by saying that the only significant variable is the rate of unemployment, and, therefore, the regression does not provide any evidence that capital punishment acts as a deterrent for murderers. Likewise, spatial effects play no roles when the spatial lag of the executions is included. The R-squared of the model is only 0.115. The lesson to be learned from this illustration is that "true model" for the murder rate is more complex.

### Illustration 2.2.1.3: A model of DUI arrests

The data in this example were used by Drukker et al. (2013c,d) in a series of papers to explain the spatial functions available from STATA. The same data set was also used by Bivand and Piras (2015) for a comparison of the implementation of spatial econometrics model estimation techniques across different software packages.[12] The simulated US Driving Under the Influence (DUI) county data set covers 3109 counties (omitting Alaska, Hawaii, and US territories), and uses simulations from variables used by Powers and Wilson (2004).[13] The dependent variable, *dui,* is defined as the alcohol-related arrest rate per 100,000

[11] The dataset has been taken from Wooldridge (2013). The original dataset is a panel data on 51 US states (including Alaska and Hawaii), but we only consider the most recent year (1993). Additionally, to create the row-standardized weighting matrix (with the queen criterion), we used a shape file that was obtained from the following address: http://www.arcgis.com/home/item.html?id=f7f805eb65eb4ab787a0a3e1116ca7e5.

[12] In particular, they focus on R, Stata, Matlab, and PySAL.

[13] The counties are taken from an ESRI Shapefile downloaded from the US Census.

daily vehicle miles traveled (DVMT). The explanatory variables include *police* (number of sworn officers per 100,000 DVMT); *nondui* (non-alcohol-related arrests per 100,000 DVMT); *vehicles* (number of registered vehicles per 1000 residents), and *dry* (a dummy for counties that prohibit alcohol sale within their borders, about 10% of counties). We also consider *wpolice,* which is the spatial lag of the number of officers in neighboring counties. Powers and Wilson (2004, p. 331) found that "there is no significant relationship between prohibition status and the DUI arrest rate when controlling for the proportionate number of sworn officers and the non-DUI arrest rate per officer" when examining data for 75 counties in Arkansas.

The OLS results from the estimated model are reported below:

$$\begin{aligned}\widehat{dui} &= -6.474(0.541) + 0.599(0.015)\ police + 0.000(0.001)\ nondui\\ &\quad + 0.016(0.001)\ vehicles + 0.107(0.035)\ dry + 0.034(0.016)\ wpolice.\end{aligned}$$

In this extended formulation of the model, there is some evidence, *ceteris paribus*, of the relationship between prohibition and arrest rate. Additionally, the number of officer in neighboring counties significantly affects the arrest rate. The simulated data has a higher $R^2$ value of 0.850, and the explanatory variables with the exception of *nondui* are all significant.

### 2.2.2 Estimation When $\rho_1 = 0$ and $\rho_2 \neq 0$

As a one step generalization, consider now the case in which the error terms are directly interrelated but the dependent variable is not:

$$\begin{aligned} y &= Z\gamma + u, \\ u &= \rho_2 W u + \varepsilon \end{aligned} \tag{2.2.2.1}$$

where again $Z$ and $\gamma$ are defined in (2.2.1.1). This model is often referred to as the spatial error model. In this case parameter $\gamma$ should be estimated by a GLS procedure which accounts for the specification of the disturbance terms.

**Properties of $u$**

Since

$$u = (I_N - \rho_2 W)^{-1}\,\varepsilon \tag{2.2.2.2}$$

and $\varepsilon \sim \left(0, \sigma_\varepsilon^2 I_N\right)$, it follows that

$$\begin{aligned} E\,(u) &= (I_N - \rho_2 W)^{-1}\, E(\varepsilon) = 0, \\ E\left(uu'\right) &= (I_N - \rho_2 W)^{-1}\, E[\varepsilon\varepsilon']\left(I_N - \rho_2 W'\right)^{-1} \end{aligned} \tag{2.2.2.3}$$

$$
\begin{aligned}
&= \sigma_\varepsilon^2 (I_N - \rho_2 W)^{-1} \left(I_N - \rho_2 W'\right)^{-1} \\
&= \sigma_\varepsilon^2 \Omega_u, \\
\Omega_u &= (I_N - \rho_2 W)^{-1} \left(I_N - \rho_2 W'\right)^{-1}.
\end{aligned}
$$

For most weighting matrices, the elements of $u$ will be both spatially correlated and heteroskedastic.

Assume for the moment that $\rho_2$ is known. Then, the GLS estimator of $\gamma$ is

$$\hat{\gamma}_{GLS} = \left(Z'\Omega_u^{-1}Z\right)^{-1} Z'\Omega_u^{-1} y. \tag{2.2.2.4}$$

Substituting $y$ from (2.2.1.1) into (2.2.2.4) it follows that

$$\hat{\gamma}_{GLS} = \gamma + \left(Z'\Omega_u^{-1}Z\right)^{-1} Z'\Omega_u^{-1} u. \tag{2.2.2.5}$$

Therefore, as easily shown,

$$
\begin{aligned}
E(\hat{\gamma}_{GLS}) &= \gamma, \qquad (2.2.2.6) \\
VC_{\hat{\gamma}_{GLS}} &= E(\hat{\gamma}_{GLS} - \gamma)(\hat{\gamma}_{GLS} - \gamma) \\
&= \sigma^2 \left(Z'\Omega_u^{-1}Z\right)^{-1}.
\end{aligned}
$$

In general, $\rho_2$ will not be known and so the GLS estimator in (2.2.2.4) is not feasible. There are, however, two feasible procedures. The first one, which we are going to see next, is based on maximum likelihood. The second procedure (discussed in a later section) is based on a generalized moments estimator of $\rho_2$.

### 2.2.2.1 Maximum Likelihood Estimation: $\rho_1 = 0$, $\rho_2 \neq 0$[14]

Typically, maximum likelihood estimation is based on the normality assumption, $\varepsilon \sim N\left(0, \sigma_\varepsilon^2 I_N\right)$.[15] Assuming normality of $\varepsilon$, it follows from (2.2.2.2) that

[14] There is a large body of literature which focuses on the maximum likelihood procedure, and procedures associated with it; see, e.g., Anselin (1988), Baltagi and Li (2004), Beron and Vijverberg (2004), Bolduc et al. (1995), Burridge (2012), Case (1991), Case et al. (1993), Dubin (1995), Florax and de Graaff (2004), Mur and Angulo (2006), Lee (2004), Lee and Yu (2012b), Yu et al. (2008).

[15] Let $\Psi$ be a random $r \times 1$ vector with mean $\mu$ and nonsingular (and therefore positive definite) $VC$ matrix $\sigma^2\Omega$. Then, if $\Psi$ is multivariate normal its density is

$$f(\psi) = \frac{e^{-\frac{1}{2\sigma^2}(\psi-\mu)'\Omega^{-1}(\psi-\mu)}}{(2\pi)^{r/2}(\sigma^2)^{r/2}\, |\Omega|_+^{1/2}}, \quad -\infty < \psi < \infty$$

where $|\Omega|_+^{1/2}$ is the positive square root of $|\Omega|$, and $-\infty < \psi < \infty$ is meant to hold for each of the $r$ elements of $\psi$.

$$u \sim N\left(0, \sigma_\varepsilon^2 \Omega_u\right), \tag{2.2.2.1.1}$$

$$\Omega_u = \sigma_\varepsilon^2 \left(I_N - \rho_2 W\right)^{-1} \left(I_N - \rho_2 W'\right)^{-1},$$

and then from (2.2.2.1) we get

$$y \sim N\left(Z\gamma, \sigma_\varepsilon^2 \Omega_u\right). \tag{2.2.2.1.2}$$

Since the likelihood function is the joint distribution of the dependent variable,

$$L = \frac{e^{-\frac{1}{2\sigma_\varepsilon^2}[y - Z\gamma]'\Omega_u^{-1}[y - Z\gamma]}}{\left(\sigma_\varepsilon^2\right)^{\frac{N}{2}} \sqrt{2\pi}^N \left|I_N - \rho_2 W\right|_+^{-1}}. \tag{2.2.2.1.3}$$

Typically, it is easier to maximize the logarithm of the likelihood than the likelihood itself. The logarithm of $L$ defined in (2.2.2.1.3) can be expressed and simplified as follows:

$$\begin{aligned}
\ln(L) &= -\frac{1}{2\sigma_\varepsilon^2}\left[y - Z\gamma\right]'\left[I_N - \rho_2 W'\right]\left[I_N - \rho_2 W\right]\left[y - Z\gamma\right] \\
&- \frac{N}{2}\ln\left(\sigma_\varepsilon^2\right) + \ln\left|I_N - \rho_2 W\right|_+ - N\ln\left(\sqrt{2\pi}\right) \\
&= -\frac{1}{2\sigma_\varepsilon^2}\left(\left[I_N - \rho_2 W\right]\left[y - Z\gamma\right]\right)'\left(\left[I_N - \rho_2 W\right]\left[y - Z\gamma\right]\right) \\
&- \frac{N}{2}\ln\left(\sigma_\varepsilon^2\right) + \ln\left(\left|I_N - \rho_2 W\right|_+\right) - N\ln\left(\sqrt{2\pi}\right) \\
&= -\frac{1}{2\sigma_\varepsilon^2}\left(y^*(\rho_2) - Z^*(\rho_2)\gamma\right)'\left(y^*(\rho_2) - Z^*(\rho_2)\gamma\right) \\
&- \frac{N}{2}\ln\left(\sigma_\varepsilon^2\right) + \ln\left(\left|I_N - \rho_2 W\right|_+\right) - N\ln\left(\sqrt{2\pi}\right)
\end{aligned} \tag{2.2.2.1.4}$$

where

$$y^*(\rho_2) = y - \rho_2 W y, \tag{2.2.2.1.5}$$

For future reference note that, in general, since $\Omega$ is positive definite, there exists a matrix $V$ such that $V'\Omega V = I_r$ where $V$ is nonsingular. Therefore, $\Omega = (V')^{-1}V^{-1}$, and so in the density above

$$\begin{aligned}
|\Omega|_+^{1/2} &= |\sigma^2(V')^{-1}V^{-1}|^{1/2} \\
&= (\sigma^2)^{r/2}|V^{-1}|_+.
\end{aligned}$$

Note that in (2.2.2.1.3) the VC matrix has this "decomposition".

Also for future reference, we note that the marginal distributions of a multivariate normal, say of $\Psi$ above, are themselves normal. The means and $VC$ matrices of these marginal distributions are given by the respective elements of $\mu$ and $\sigma^2\Omega$, e.g., the mean and variance of the first element of $\Psi$ are respectively the first element of $\mu$ and the first diagonal element of $\sigma^2\Omega$.

$$Z^{*}(\rho_2) = Z - \rho_2 WZ.$$

Note $y^{*}(\rho_2)$ and $Z^{*}(\rho_2)$ are spatial counterparts to the Cochrane–Orcutt transformation; see Cochrane and Orcutt (1949).

At this point we note a potentially serious problem in maximizing $\ln(L)$. This problem relates to the term $\ln\left(|I_N - \rho_2 W|_+\right)$ In fact, this term must:

**(a)** Be evaluated repeatedly for each trial value of $\rho_2$. If $N$ is large this will indeed be "tedious" for certain weighting matrices. For example, cross-sectional units could relate to counties, and there are over 3000 counties in the US. In other cross-sectional studies, the number of cross-sectional units could be families and so it could be the case that $N$ gets large very easily.

**(b)** Using Ord's (1975) suggestion for evaluating the determinant in terms of its characteristic roots, we have

$$\begin{aligned}\ln|I_N - \rho_2 W|_+ &= \ln\left[(1-\rho_2\lambda_1)\cdots(1-\rho_2\lambda_N)\right] \\ &= \sum_{i=1}^{N}\ln(1-\rho_2\lambda_i)\,.\end{aligned} \qquad (2.2.2.1.6)$$

If the roots can be accurately evaluated, $\ln|1-\rho_2 W|_+$ can be evaluated in terms of the sum for each trial value of $\rho_2$. This will be far simpler than the method proposed in (a). The problem is that if $N \geq 450$ both (a) and (b) could involve computational accuracy problems. For example, Kelejian and Prucha (1999) found that the calculation of roots for even a 400 × 400 non-symmetric matrix involved accuracy problems. Despite this, many other solutions have been proposed, both exact (e.g., Cholesky decomposition for symmetric matrices) and approximated such as the Chebyshev decomposition (Pace and LeSage, 2004) or the Monte Carlo approach (Barry and Pace, 1999). A complete treatment of these alternative methods is beyond the scope of our book and it can be found, for example, in LeSage and Pace (2009). Here we just want to mention that most of the methods generally used to approximate the log-Jacobian term (or simply speed up its computations) are both very efficient and reliable particularly when the sample size is large and the weighting matrix is particularly sparse.[16]

**(c)** Another problem with the maximum likelihood procedure itself is that many models considered in practice are such that the regressor matrix $X$ contains some regressors which are endogenous. In this case the maximum likelihood procedure cannot be implemented unless the joint distribution of those endogenous regressors and the model dependent variable is known.

[16] Modern spatial econometrics software, such as R or Matlab, has the capability of dealing with many of those methods.

This will typically not be the case, and so other procedures are suggested below!

### The Form of the MLE for $\gamma$ and $\sigma_\varepsilon^2$

Based on (2.2.2.1.4), we have

$$\frac{\partial \ln L}{\partial \gamma} = -\frac{1}{2\sigma_\varepsilon^2}\left(-2Z^*(\rho_2)' y^*(\rho_2) + 2Z^*(\rho_2)' Z^*(\rho_2)\gamma\right), \tag{2.2.2.1.7}$$
$$\frac{\partial \ln L}{\partial \sigma_\varepsilon^2} = \frac{1}{2\sigma_\varepsilon^4}\left[y^*(\rho_2) - Z^*(\rho_2)\gamma\right]'\left[y^*(\rho_2) - Z^*(\rho_2)\gamma\right] - \frac{N}{2\sigma_\varepsilon^2}.$$

It follows that

$$\hat{\gamma}_{ML} = \left[Z^*(\hat{\rho}_2)' Z^*(\hat{\rho}_2)\right]^{-1} Z^*(\hat{\rho}_2)' y^*(\hat{\rho}_2), \tag{2.2.2.1.8}$$
$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N}\left[y^*(\hat{\rho}_2) - Z^*(\hat{\rho}_2)\hat{\gamma}_{ML}\right]'\left[y^*(\hat{\rho}_2) - Z^*(\hat{\rho}_2)\hat{\gamma}_{ML}\right].$$

The interpretation of the MLEs is straightforward. Premultiplying (2.2.2.1) by $(I_N - \rho_2 W)$ yields

$$y^*(\rho_2) = Z^*(\rho_2)\gamma + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 I_N\right). \tag{2.2.2.1.9}$$

If $\rho_2$ were known, the model in (2.2.2.1.9) would be a standard linear regression model with a disturbance term that is normally distributed. The efficient estimator of $\gamma$ would be the least squares estimator defined in (2.2.2.1.8) with $\hat{\rho}_2$ replaced by $\rho_2$, e.g., the MLE is just the feasible counterpart to that estimator.

In an important article, Lee (2004) gave a formal demonstration of conditions that ensure consistency and asymptotic normality of the ML estimators for the general spatial model considered. In applied studies, dealing with a variety of regression models, it is (almost) always assumed that the usual results hold. For the model at hand, we have

$$\sqrt{N}\left(\hat{P} - P\right) \xrightarrow{D} N(0, V), \tag{2.2.2.1.10}$$
$$V^{-1} = -\lim E\left[N^{-1}\left(\frac{\partial^2 \ln L}{\partial P \partial P'}\right)\right], \quad P = \begin{bmatrix} \gamma \\ \rho_2 \\ \sigma_\varepsilon^2 \end{bmatrix}.$$

The matrix $-E\left(\frac{\partial^2 \ln L}{\partial P \partial P'}\right)$ is referred to as the information matrix, denoted by $I(P)$, and

$$I(P) = -E\left(\frac{\partial^2 \ln L}{\partial P \partial P'}\right) = E\left(\frac{\partial \ln(L)}{\partial P}\right)\left(\frac{\partial \ln(L)}{\partial P}\right)'. \tag{2.2.2.1.11}$$

Small sample inference is typically based on the approximations

$$\hat{P} = \begin{bmatrix} \hat{\gamma}_{ML} \\ \hat{\rho}_{2,ML} \\ \hat{\sigma}^2_{\varepsilon,ML} \end{bmatrix}, \quad \hat{P} \mathrel{\dot{\sim}} N\left(P, -\frac{\hat{V}}{N}\right), \tag{2.2.2.1.12}$$

$$\frac{\hat{V}}{N} = \left[\frac{\partial^2 \ln L}{\partial P \partial P'}\right]^{-1}_{\hat{P}}.$$

### A Computational Note on ML

The estimators in (2.2.2.1.8) are functions of $\hat{\rho}_2$. Computationally, a concentrated likelihood approach is generally taken. The idea is to concentrate out of the likelihood function both $\gamma$ and $\sigma^2_\varepsilon$. The resulting function only depends on the parameter $\rho_2$. Then, from a computational perspective, the following steps are undertaken:

**1.** A simple linear regression model of $y$ over $Z$ allows the computation of an initial set of OLS residuals.

**2.** Using these residuals, one can maximize the concentrated likelihood and find an initial estimate of $\rho_2$.

**3.** Next, a feasible GLS procedure can be implemented to estimate $\gamma$, and another set of GLS residuals can be computed.

**4.** Finally, one should iterate back and forth between FGLS and ML until a convergence criterion is met.

Virtually all the available software use a procedure like the one described above.

### Illustration 2.2.2.1.1: House value, air pollution, and crime in Boston

Let us consider again the housing price dataset in Boston. There are many additional variables that were not included in the original specification that might well be spatially correlated. In this illustration we account for these "left-out" variables in terms of a disturbance term that follows a spatial autoregressive model of order one.[17]

[17] We use the expression "spatial autoregressive model of order one" to indicate that we only consider the first spatial lag. However, other models could be specified including more than one lag (or even lags including different spatial weighting matrix). The estimation theory for these models was considered by Lee and Liu (2010) and Badinger and Egger (2011), among others. Evident estimators should be suggested by problems 1–3 at the end of this chapter.

The results from the estimated model that include an autocorrelated disturbance term are reported below:

$$\begin{aligned}\log(\widehat{price}) = {} & 2.306(0.170) - 0.588(0.132)\log(nox) - 0.151(0.058)\log(dist)\\ & - 0.032(0.006)\, stratio + 0.193(0.013)\, rooms\\ & - 0.008(0.001)\, crime - 0.014(0.002)\, wcrime.\end{aligned}$$

The model was estimated by ML using the Ord (1975) procedure to calculate the Jacobian term. The value of $\rho_2$ is equal to 0.681 and is statistically significant at the 1% level (with an asymptotic standard error of 0.034). The coefficient measuring the elasticity of price with respect to the level of pollution is lower than the one estimated in Examples 2.2.1.1. This can be interpreted as a sign of misspecification of the simple OLS model without the spatially autocorrelated error term.

Let us now consider the same model estimated by ML but using a different method for the computation of the Jacobian term.[18] The results from this model are reported below[19]:

$$\begin{aligned}\log(\widehat{price}) = {} & 2.305(0.170) - 0.589(0.132)\log(nox) - 0.151(0.058)\log(dist)\\ & - 0.032(0.006)\, stratio + 0.193(0.013)\, rooms\\ & - 0.008(0.001)\, crime - 0.014(0.0025)\, wcrime\end{aligned}$$

with $\rho_2 = 0.680$ again strongly significant. As can be noted, there is virtually no difference between the exact and approximated computation of the log-Jacobian term on this data.

## 2.2.3 Assumptions of the General Model

Before considering the development of an IV procedure put forth by Kelejian and Prucha (1998, 1999) for the general model in (2.2.1), we review and interpret its basic assumptions.

[18] Of the various methods available, we choose the Monte Carlo approach (Barry and Pace, 1999, LeSage and Pace, 2009). This method is based on the simple intuition: Since the matrix logarithm has a simple infinite series expansion in terms of the powers of $W$ and that the trace is a linear operator, the log-determinant can be expressed as a weighted series of traces of the powers of $W$:

$$\ln|I_N - \rho_2 W| = -\sum_{i=1}^{\infty}\frac{\rho_2^i Tr(W^i)}{i} = -\sum_{i=1}^{k}\frac{\rho_2^i Tr(W^i)}{i} - R.$$

For small $R$, the expression above can be used to approximate the log-determinant.

[19] Note that the Monte Carlo approach is an approximated method therefore the results are subject to change (slightly) over different trials.

First note that the general model in (2.2.1) can be expressed as

$$\begin{aligned} y &= Z\beta + u, \\ u &= \rho_2 (Wu) + \varepsilon, \quad |\rho_1| < 1, \ |\rho_2| < 1, \end{aligned} \tag{2.2.3.1}$$

where

$$\begin{aligned} Z &= (e_N, X, Wy, WX), \\ \beta &= (a, B_1', \rho_1, B_2'). \end{aligned}$$

As we already observed in previous sections of this chapter, the presence of $Wy$ introduces an endogeneity into the model. Therefore, the model needs to be estimated by a procedure which accounts for this endogeneity. In this section we discuss an instrumental variable procedure.

As shown in Kelejian and Prucha (1998), the ideal instruments are simply $E(Wy)$. Assume that the characteristic roots of the weighting matrix $W$ are less than or equal to 1 in absolute value, which they would be if $W$ is row normalized. In this case

$$(I_N - \rho_1 W)^{-1} = I_N + \rho_1 W + \rho_1^2 W^2 + \dots . \tag{2.2.3.2}$$

Given this inverse, the solution of the model for $y$ as described in (2.2.3.1) is

$$\begin{aligned} y &= (I_N - \rho_1 W)^{-1} [a e_N + X B_1 + W X B_2 + u] \\ &= [I_N + \rho_1 W + \rho_1^2 W^2 + \dots] [a e_N + X B_1 + W X B_2 + u] . \end{aligned} \tag{2.2.3.3}$$

Since $E(u) = 0$, it follows from (2.2.3.3) that the mean of $y$ is linear in $e_N, We_N, W^2 e_N, \dots, X, WX, W^2 X, \dots$.

Our instrument matrix for this model is $H$, where $H$ is an $N \times k_H$ nonstochastic matrix which consists of the linearly independent $(LI)$ columns of $(e_N, X, We_N, WX, W^2 e_N, W^2 X, \dots, W^r X)$ and is denoted as

$$H = (e_N, X, We_N, WX, W^2 e_N, W^2 X, \dots, W^r X)_{LI} \tag{2.2.3.4}$$

where $r$ is a finite constant. In many cases the weighting matrix is row normalized and so the terms $We_N, W^2 e_N, \dots$ are not considered. The reason for this is that $We_N = e_N$, $W^2 e_N = We_N = e_N$, and so on, by repeated substitution $W^r e_N = e_N$. For estimation, researchers often take $r = 2$.[20]

[20] For further details, see, e.g., Kelejian and Prucha (1999), Lee (2003, 2007), Kelejian et al. (2004), and Das et al. (2003), among others.

For the error term in (2.2.3.1), let $\bar{u} = Wu$, $\bar{\bar{u}} = W^2u$, and $\bar{\varepsilon} = W\varepsilon$, and let

$$g = \frac{1}{N}\begin{bmatrix} u'u \\ \bar{u}'\bar{u} \\ u'\bar{u} \end{bmatrix}, \quad G = \frac{1}{N}\begin{bmatrix} -\bar{u}'\bar{u} & 2u'\bar{u} & N \\ -\bar{\bar{u}}'\bar{\bar{u}} & 2\bar{\bar{u}}'\bar{u} & Tr(W'W) \\ -\bar{u}'\bar{\bar{u}} & u'\bar{\bar{u}} + \bar{u}'\bar{u} & 0 \end{bmatrix}. \qquad (2.2.3.5)$$

Using this notation, the assumptions for the general model are given below. Their interpretations follow.

**Assumption 2.1.** The elements of the innovation vector $\varepsilon$ are independent and identically distributed with mean zero and variance $\sigma_\varepsilon^2$, and have finite fourth moments. That is, $\varepsilon_i$ is i.i.d. $\left(0, \sigma_\varepsilon^2\right)$, $E\left(\varepsilon_i^4\right) < \infty$.[21]

**Assumption 2.2.** $|\rho_1| < 1$, $|\rho_2| < 1$.

**Assumption 2.3.** Let $P_{(a)} = (I_N - aW)$. Then $P_{(a)}$ is nonsingular for all $|a| < 1$.

**Assumption 2.4.** The diagonal elements of the spatial weighting matrix $W$ are set to zero: $w_{ii} = 0$, $i = 1, \ldots, N$.

**Assumption 2.5.** The row and column sums of $W$ and $P_{(a)}^{-1}$ are uniformly bounded in absolute value for all $|a| < 1$.

**Assumption 2.6.** (a) $g \xrightarrow{P} g^*$ where the elements of $g^*$ are finite; (b) $G \xrightarrow{P} G^*$ where $G^*$ is a finite nonsingular matrix.

**Assumption 2.7.** The matrix of regressors $X$ and the matrix of instruments $H$ are nonstochastic matrices whose elements are uniformly bounded in absolute value.

**Assumption 2.8.** (a) The matrix of the exogenous regressors has full column rank: $rank(e_N, X, WX) = 2k + 1$; (b) The instrument matrix $H$ has full column rank and contains at least the columns $(e_N, X, WX, W^2X)$ and $rank(H) \geq 2 + 2k$.

**Assumption 2.9.** The following matrix products are such that:

$$(A)\ p \lim_{N\to\infty} N^{-1}H'Z = Q_{HZ},$$

$$(B)\ p \lim_{N\to\infty} N^{-1}H'WZ = Q_{HWZ},$$

[21] This statement of the assumption does not allow for triangular arrays. A more formal statement of the assumption which does allow for triangular arrays is the central limit theorem in Section A.15 of the appendix.

$$(C)\ p\lim_{N\to\infty} N^{-1}H'H = Q_{HH}$$

where $Q_{HZ}$, $Q_{HWZ}$, $Q_{HH}$, and $(Q_{HZ} - \rho_2 HWZ)$ are finite full column rank matrices, and so $Q_{HH}$ is positive definite.

### Interpretations of Assumptions

On an intuitive level, Assumption 2.1 indicates that the elements of $\varepsilon$ are independent and identically distributed with zero mean and finite fourth moments. As stated, this assumption does not account for triangular arrays which (most) spatial models imply. The assumption is given in this simple form for ease of presentation and interpretation. A formal version of this assumption which accounts for triangular arrays is given in Section A.15 of the appendix on large sample theory. In that formal version the dependence of the disturbance vector $\varepsilon$ and its elements on the sample size is denoted as $\varepsilon_N$ and $\varepsilon_{iN}$, respectively.

Note that in the formal assumption in Section A.15, for each given value of $N \geq 1$, the terms $\varepsilon_{1,N}, \varepsilon_{2,N}, ..., \varepsilon_{N,N}$ are assumed to be identically and jointly independently distributed. It does **not** say, e.g., that $\varepsilon_{5,N}$ is independent of $\varepsilon_{5,N^*}$ if $N^* \neq N$. These two terms may, or may not, be independent; indeed, they may not even relate to the same unit; see the discussion in Section 2.1.1.

Assumptions 2.2 and 2.3 imply that the general model in (2.2.1) is complete in that it can be solved for the dependent vector $y$ in terms of the exogenous variables, the weighting matrix, and the stochastic term $\varepsilon$. In particular, as in (2.1.1.1),

$$y = (I_N - \rho_1 W)^{-1}[ae_N + XB_1 + (WX)B_2 + (I_N - \rho_2 W)^{-1}\varepsilon]. \qquad (2.2.3.6)$$

Assumption 2.4 is typical in that it is virtually assumed in all applied works. It states that the diagonal elements of $W$ are taken to be zero. This simply implies that no unit is viewed as its own neighbor.

Although Assumption 2.5 is typically made in formal large sample analysis,[22] it does place restrictions which may not hold in certain models. For example, in a spatial framework if one unit, say the fifth unit, is central in that all other units relate to it in a nonnegligible fashion, then the sum of the absolute values of the elements in the fifth column of $W$ would be expected to diverge as $N \to \infty$. Thus, Assumption 2.5 rules out models which have a "central" unit which all units are related to with weights that are bounded away from zero.

Parts (a) and (b) of Assumption 2.6 relate to the consistency of one of our suggested estimators of $\rho_2$ described below. This is demonstrated in the appendix. The consistency of our second suggested estimator of $\rho_2$, also described

[22] See, e.g., Kelejian and Prucha (1998, 1999, 2004), Lee (2004, 2007), and Kelejian and Piras (2011), among others.

below, then follows on an intuitive level. A formal, but tedious proof of consistency for the second suggested estimator is given in Kelejian and Prucha (1999).

Given Assumptions 2.3 and 2.5, we give results in Section 2.2.4 below, which demonstrate how to easily determine the probability limits described in Assumption 2.6.

Assumption 2.7 and part (a) of Assumption 2.8 rule out perfect multicolinearity. Part (b) of Assumption 2.8 implies that the number of linearly independent instruments is at least as large as the number of parameters to be estimated in the model. Assumption 2.9 is standard in large sample analysis and will, along with our other assumptions, ensure consistency and asymptotic normality of our regression parameter estimator.[23]

### 2.2.4 A Generalized Moments Estimator of $\rho_2$

The IV estimation of our general model in (2.2.3.1) which accounts for the spatial correlation requires an estimator of $\rho_2$. In this section we suggest two generalized moments estimators (GMM) given in Kelejian and Prucha (1999) for $\rho_2$. The first one of these two estimators is linear, and the second one is nonlinear. In the appendix to this chapter we give a somewhat high-level proof that the linear estimator is consistent under very reasonable conditions. That proof will strongly suggest that the nonlinear estimator is also consistent. Our high-level proof is not overly tedious, is straightforward, and instructive in that its format can be applied in other situations. A more complex low-level proof that both of our estimators are consistent is given by Kelejian and Prucha (1999).

As one might expect, our estimator of $\rho_2$ will be based on an estimated disturbance vector, $u$, which in turn requires an initial estimator of the regression parameters, $\beta$. In this section we also describe this estimator.

#### A Preliminary Result on Limits of Stochastic Quadratic Forms

Let $S$ be an $N \times N$ nonstochastic matrix whose row and column sums are uniformly bounded in absolute value. Let $\nu' = (\nu_1, \ldots, \nu_N)$ where $\nu_i$ are i.i.d. $\left(0, \sigma_\nu^2\right)$ and $E\left(\nu_i^4\right) = \mu_4 < \infty$. Then,

$$\frac{\nu' S \nu}{N} - E(\frac{\nu' S \nu}{N}) \xrightarrow{P} 0 \tag{2.2.4.1}$$

where, of course,

$$E(\frac{\nu' S \nu}{N}) = \sigma_\nu^2 \frac{Tr\,(S)}{N}.$$

[23] For example, assumptions similar to our Assumption 2.8 have been used in various frameworks, see, e.g., Kelejian and Prucha (1998, 2004), Kapoor et al. (2007), Kelejian and Piras (2011), and Lee (2004).

If the limit of $Tr\,(S)\,/N$ exists, and is given by

$$\lim_{N\to\infty} \frac{Tr\,(S)}{N} = S^*, \tag{2.2.4.2}$$

then from (2.2.4.1) we have

$$\frac{\nu' S \nu}{N} \xrightarrow{P} \sigma_\nu^2 S^*. \tag{2.2.4.3}$$

The proof of (2.2.4.1) is given in the appendix to this chapter.

The result in (2.2.4.1) is important, and somewhat general, and will be used at various points in this book. As a simple application of (2.2.4.1)–(2.2.4.3), the reader should convince himself/herself that

$$(\frac{\nu' I_N \nu}{N}) = (\frac{\nu' \nu}{N}) \xrightarrow{P} \sigma_\nu^2. \tag{2.2.4.4}$$

Consider again Assumption 2.6 which relates to the probability limits of $g$ and $G$. Since $\bar{u} = Wu$ and $\bar{\bar{u}} = W^2 u$, every element of $g$ and $G$, except for those in the third column of $G$, can be expressed as a quadratic form in the disturbance vector $u$. Since $u = (I_N - \rho_2 W)^{-1}\varepsilon$ and $\bar{\varepsilon} = W\varepsilon$, these terms can also be expressed as quadratic forms in the innovation vector $\varepsilon$. Finally, note that the matrices in these quadratic forms in $\varepsilon$ involve products of $W$ and $(I_N - \rho_2 W)^{-1}$. Since $|\rho_2| < 1$, Assumption 2.5 implies that the row and column sums of $W$ and $(I_N - \rho_2 W)^{-1}$ are uniformly bounded in absolute value. Furthermore, since the row and column sums of the product of such matrices are also uniformly bounded in absolute value, each term in $g$ and all the terms in $G$, other than the constants in the third column, can be expressed as $\varepsilon' S \varepsilon$ where, again, $S$ is a matrix whose rows and column sums are uniformly bounded in absolute value.[24] The probability limits in Assumption 2.6 can be determined using the result in (2.2.4.3).

### A Preliminary, but Consistent Estimator of $\beta$

Our preliminary consistent estimator of $\beta$ is the two stage least squares (henceforth, 2SLS) estimator using the instrument matrix $H$ given in (2.2.3.4). Specifically, let $\tilde{Z} = H(H'H)^{-1}H'Z$; then our preliminary estimators of $\beta$ in (2.2.3.1) and of the disturbance vector $u$ are

$$\tilde{\beta} = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'y, \tag{2.2.4.5}$$
$$\hat{u} = y - Z\tilde{\beta}.$$

[24] Let $a$ and $b$ be any two finite constants. Then the row and column sums of $aS + bS$ are clearly also uniformly bounded in absolute value, i.e., $aS + bS = S$.

To see the consistency of $\tilde{\beta}$, first note that $H(H'H)^{-1}H'$ is symmetric idempotent and so $\tilde{Z}'Z = \tilde{Z}'\tilde{Z}$. Thus, replacing $y$ by its expression in (2.2.3.1) we have

$$\begin{aligned}\tilde{\beta} &= \beta + (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'u \\ &= \beta + \left([N^{-1}Z'H][N(H'H)^{-1}][N^{-1}H'Z]\right)^{-1} * \\ &[N^{-1}Z'H][N^{-1}H'H]^{-1}[N^{-1}H'u].\end{aligned} \tag{2.2.4.6}$$

Assumptions 2.9 and 2.6 imply that

$$(\tilde{\beta} - \beta) - (Q'_{HZ}Q^{-1}_{HH}Q_{HZ})^{-1}Q'_{HZ}Q^{-1}_{HH}[N^{-1}H'u] \xrightarrow{P} 0. \tag{2.2.4.7}$$

Since, by Assumption 2.9, $(Q'_{HZ}Q^{-1}_{HH}Q_{HZ})^{-1}Q'_{HZ}Q^{-1}_{HH}$ is a finite matrix, it should be clear from (2.2.4.7) that $\tilde{\beta}$ is consistent: $\tilde{\beta} \xrightarrow{P} \beta$ if $N^{-1}H'u \xrightarrow{P} 0$. This follows from Chebyshev's inequality. To see that this is the case, let $\phi = N^{-1}H'u$, and note that $E(\phi) = 0$ and $E(\phi\phi') = \sigma^2_\varepsilon N^{-2}H'\Omega H$, where $\Omega = (I_N - \rho_2 W)^{-1}(I_N - \rho_2 W')^{-1}$. Assumption 2.5 implies that the row and column sums of $\Omega$ are uniformly bounded in absolute value; it then follows from statement 4 following (2.1.4.4) that $N^{-2}H'\Omega H \to 0$. Therefore, by Chebyshev's inequality $N^{-1}H'u \xrightarrow{P} 0$ and so $\tilde{\beta} \xrightarrow{P} \beta$.

### Development of the GMM of $\rho_2$

We now develop a GMM estimator for $\rho_2$. This GMM estimator will be based on $\tilde{\beta}$ and $\hat{u}$ in (2.2.4.5).

First note from (2.2.3.1) that

$$u - \rho_2 Wu = \varepsilon. \tag{2.2.4.8}$$

Premultiplying across by $W$ yields

$$Wu - \rho_2 W^2 u = W\varepsilon. \tag{2.2.4.9}$$

Again, let $\bar{u} = Wu$, $\bar{\bar{u}} = W^2 u$, $\bar{\varepsilon} = W\varepsilon$, and denote their $i$th elements by $\bar{u}_i$, $\bar{\bar{u}}_i$, and $\bar{\varepsilon}_i$, respectively. Then

$$\begin{aligned} u_i - \rho_2\bar{u}_i &= \varepsilon_i, \\ \bar{u}_i - \rho_2\bar{\bar{u}}_i &= \bar{\varepsilon}_i, \quad i = 1, \ldots, N.\end{aligned} \tag{2.2.4.10}$$

Square the first line in (2.2.4.10), then sum and divide by $N$ to get

$$\frac{\sum u_i^2}{N} + \rho_2^2\frac{\sum \bar{u}_i^2}{N} - 2\rho_2\frac{\sum u_i\bar{u}_i}{N} = \frac{\sum \varepsilon_i^2}{N}. \tag{2.2.4.11}$$

Square the second line in (2.2.4.10), then sum and divide by $N$ to get

$$\frac{\sum \bar{u}_i^2}{N} + \rho_2^2 \frac{\sum \bar{\bar{u}}_i^2}{N} - 2\rho_2 \frac{\sum \bar{u}_i \bar{\bar{u}}_i}{N} = \frac{\sum \bar{\varepsilon}_i^2}{N}. \tag{2.2.4.12}$$

Finally, multiply the first line in (2.2.4.10) by the second line in (2.2.4.10), then sum and divide by $N$ to get

$$\frac{\sum u_i \bar{u}_i}{N} + \rho_2^2 \frac{\sum \bar{u}_i \bar{\bar{u}}_i}{N} - \rho_2 \left[ \frac{\sum u_i \bar{\bar{u}}_i}{N} + \frac{\sum \bar{u}_i^2}{N} \right] = \frac{\sum \varepsilon_i \bar{\varepsilon}_i}{N}. \tag{2.2.4.13}$$

Now consider the right-hand sides of (2.2.4.11)–(2.2.4.13). Since $\varepsilon_i$ are i.i.d. $\left(0, \sigma_\varepsilon^2\right)$ and $E\left(\varepsilon_i^4\right) < \infty$, it follows from (2.2.4.1)–(2.2.4.3) that the right-hand side term in (2.2.4.11) is such that

$$\frac{\sum \varepsilon_i^2}{N} = \sigma_\varepsilon^2 + \delta_1 \text{ where } \delta_1 \xrightarrow{P} 0. \tag{2.2.4.14}$$

Now consider the right-hand side term in (2.2.4.12) and note that

$$\frac{\sum \bar{\varepsilon}_i^2}{N} = \frac{\varepsilon' W' W \varepsilon}{N}. \tag{2.2.4.15}$$

Since, by Assumption 2.5, the row and column sums of $W$ are uniformly bounded in absolute value, the row and column sums of $WW'$ are also bounded in absolute value. It then follows from (2.2.4.1) that

$$\frac{\sum \bar{\varepsilon}_i^2}{N} - \sigma_\varepsilon^2 \frac{Tr\left(W'W\right)}{N} \xrightarrow{P} 0. \tag{2.2.4.16}$$

Assumption 2.6 implies that the limit of $Tr\left(W'W\right)/N$ exits, and so we will express (2.2.4.16) as

$$\frac{\sum \bar{\varepsilon}_i^2}{N} = \sigma_\varepsilon^2 \frac{Tr\left(W'W\right)}{N} + \delta_2 \text{ where } \delta_2 \xrightarrow{P} 0. \tag{2.2.4.17}$$

Finally, the term on the right-hand side of (2.2.4.13) can be expressed as

$$\frac{\sum \varepsilon_i \bar{\varepsilon}_i}{N} = \frac{\varepsilon' W \varepsilon}{N}. \tag{2.2.4.18}$$

By Assumption 2.4, the diagonal elements of $W$ are zero. Given this, it follows that

$$\begin{aligned} E(\varepsilon' W \varepsilon / N) &= \frac{1}{N} \sum_{i=1}^{N} E[\varepsilon_i^2] w_{ii} + \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} E[\varepsilon_i \varepsilon_j] \\ &= 0 \end{aligned} \tag{2.2.4.19}$$

since $w_{ii}$ is zero and $E(\varepsilon_i\varepsilon_j)$ equal zero if $i \neq j$. Since the row and column sums of $W$ are uniformly bounded in absolute value, it follows from (2.2.4.1)–(2.2.4.3) that

$$\frac{\varepsilon' W' \varepsilon}{N} \xrightarrow{P} 0. \tag{2.2.4.20}$$

For consistency of notation, we express $\varepsilon' W' \varepsilon / N$ as

$$\frac{\varepsilon' W' \varepsilon}{N} = \delta_3 \text{ where } \delta_3 \xrightarrow{P} 0. \tag{2.2.4.21}$$

At this point we bring together the three equations which will be used to estimate $\rho_2$. These equations are (2.2.4.11), (2.2.4.12), and (2.2.4.13) with the right-hand side terms involving $\varepsilon_i$, $i = 1, ..., N$ replaced by their expressions in (2.2.4.14), (2.2.4.17), and (2.2.4.21):

$$\begin{aligned}\frac{\sum u_i^2}{N} + \rho_2^2 \frac{\sum \bar{u}_i^2}{N} - 2\rho_2 \frac{\sum u_i \bar{u}_i}{N} &= \sigma_\varepsilon^2 + \delta_1, \\ \frac{\sum \bar{u}_i^2}{N} + \rho_2^2 \frac{\sum \bar{\bar{u}}_i^2}{N} - 2\rho_2 \frac{\sum \bar{u}_i \bar{\bar{u}}_i}{N} &= \sigma_\varepsilon^2 \frac{Tr\left(W'W\right)}{N} + \delta_2, \\ \frac{\sum u_i \bar{u}_i}{N} + \rho_2^2 \frac{\sum \bar{u}_i \bar{\bar{u}}_i}{N} - \rho_2 \left[\frac{\sum u_i \bar{\bar{u}}_i}{N} + \frac{\sum \bar{u}_i^2}{N}\right] &= \delta_3.\end{aligned} \tag{2.2.4.22}$$

For the moment, we reparameterize these three equations by replacing $\rho_2^2$ by a "new" parameter, say $\eta$, i.e., $\eta = \rho_2^2$. In this case, the three equations in (2.2.4.22) can be viewed as a regression model with three parameters, $\eta$, $\rho_2$, and $\sigma_\varepsilon^2$, and three observations, i.e., the information that $\eta = \rho_2^2$ is ignored. To see this, let $\zeta' = \left[\eta, \rho_2, \sigma_\varepsilon^2\right]$ and $\delta' = [\delta_1, \delta_2, \delta_3]$. Note that the left-most terms in these three equations do not involve one of the three parameters, and all the other terms do, with the exception of the error terms $\delta_1$, $\delta_2$, and $\delta_3$.

Let

$$y_* = \begin{bmatrix} \frac{\sum u_i^2}{N} \\ \frac{\sum \bar{u}_i^2}{N} \\ \frac{\sum u_i \bar{u}_i}{N} \end{bmatrix}, \quad X_* = \begin{bmatrix} -\frac{\sum \bar{u}_i^2}{N} & 2\frac{\sum u_i \bar{u}_i}{N} & 1 \\ -\frac{\sum \bar{\bar{u}}_i^2}{N} & 2\frac{\sum \bar{u}_i \bar{\bar{u}}_i}{N} & \frac{Tr(W'W)}{N} \\ -\frac{\sum \bar{u}_i \bar{\bar{u}}_i}{N} & \left[\frac{\sum u_i \bar{\bar{u}}_i}{N} + \frac{\sum \bar{u}_i^2}{N}\right] & 0 \end{bmatrix}. \tag{2.2.4.23}$$

Given this notation, the three equations in (2.2.4.22) can be expressed as

$$y_* = X_* \zeta + \delta, \quad \delta \xrightarrow{P} 0, \tag{2.2.4.24}$$

$$\delta' = (\delta_1, \delta_2, \delta_3).$$

Finally, let $\hat{y}_*$ and $\hat{X}_*$ be respectively identical to $y_*$ and $X_*$ except that $u$ in (2.2.4.23) is replaced everywhere by $\hat{u}$ in (2.2.4.5). Defining $\hat{\delta} = \hat{y}_* - \hat{X}_* \zeta$, the three equations in (2.2.4.22), with $u$ replaced by $\hat{u}$, can be expressed as

$$\hat{y}_* = \hat{X}_* \zeta + \hat{\delta}. \tag{2.2.4.25}$$

Note that $\hat{y}_*$ and $\hat{X}_*$ are observable and are respectively a $3 \times 1$ vector and a $3 \times 3$ matrix. The model in (2.2.4.25) can be viewed as a regression model.

The linear generalized moments estimator given in Kelejian and Prucha (1999) for $\rho_2$ is the least squares estimator of $\rho_2$ based on (2.2.4.25), i.e., the second element of $\hat{\zeta}$ is

$$\hat{\zeta} = \hat{X}_*^{-1} \hat{y}_*. \tag{2.2.4.26}$$

The nonlinear generalized moments estimator in Kelejian and Prucha (1999) for $\rho_2$ is also based on the model in (2.2.4.25) except that $\eta$ is replaced by $\rho_2^2$ and then nonlinear least squares is applied, i.e., $\hat{\delta}'\hat{\delta}$ is minimized with respect to the remaining parameters, namely $\rho_2$ and $\sigma_\varepsilon^2$. It should be clear that the linear estimator of $\rho_2$ will be inefficient relative to the nonlinear estimator which uses the information that $\eta = \rho_2^2$. Results given in Kelejian and Prucha (1999) suggest that this is indeed the case.

In the appendix to this chapter the linear estimator $\hat{\zeta}$ in (2.2.4.26) is shown to be consistent. In addition, an argument is given which strongly suggests that the nonlinear estimator is also consistent. A more complex and tedious low-level proof that both the linear and nonlinear estimators are consistent is given in Kelejian and Prucha (1999).

### A Note on the Spatial Error Model

In the spatial error model $\rho_1 = 0$ and $\rho_2 \neq 0$ and so the general solution for $y$ in (2.2.3.3) reduces to

$$\begin{aligned} y &= a e_N + X B_1 + W X B_2 + u \\ &= MC + u \end{aligned} \tag{2.2.4.27}$$

where $M = (e_N, X, WX)$ and $C' = (a, B_1', B_2')$. In this case there is no endogeneity, and so the above GMM estimator of $\rho_2$ would be based on

$$\begin{aligned} \tilde{u} &= y - M\hat{C}, \\ \hat{C} &= (M'M)^{-1} M' y. \end{aligned} \tag{2.2.4.28}$$

## 2.3 IV ESTIMATION OF THE GENERAL MODEL

We are now ready to develop the IV estimation for the general model. As a preview, there are two complications that must be considered when estimating this model. The first complication is that $Wy$ is endogenous since it is correlated with the error term. This should be evident since $y$ depends directly on the error term, $u$. The second problem is that the error term is spatially correlated as well as heteroskedastic; see (2.2.2.3). If this second complication is not accounted for, the resulting regression parameter estimator will not be efficient.

Consider again the general model as expressed in (2.2.1), namely

$$\begin{aligned} y &= ae_N + XB_1 + \rho_1 Wy + (WX)B_2 + u, \\ u &= \rho_2 (Wu) + \varepsilon, \quad |\rho_1| < 1, \; |\rho_2| < 1, \end{aligned} \tag{2.3.1}$$

and its more compact expression in (2.2.3.1), namely

$$\begin{aligned} y &= Z\beta + u, \\ Z &= (e_N, X, Wy, WX), \\ \beta &= (a, B_1', \rho_1, B_2'). \end{aligned} \tag{2.3.2}$$

In our estimation procedure we first use the GMM procedure to estimate $\rho_2$, and then transform the model in a way that is similar to the Cochrane–Orcutt (1946) procedure. Finally, we estimate the transformed model by an IV procedure, namely two stage least squares. The procedure was first suggested by Kelejian and Prucha (1998), and is called general spatial two stage least squares, henceforth abbreviated a GS2SLS.

### The IV Estimator

Note first that since $\hat{\rho}_2$ is consistent it can be expressed as

$$\hat{\rho}_2 = (\rho_2 + \Delta_{1,N}), \qquad \Delta_{1,N} \xrightarrow{P} 0. \tag{2.3.3}$$

Therefore, from (2.2.3.1) it follows that

$$\begin{aligned} (I_N - \hat{\rho}_2 W)u &= (I_N - (\rho_2 + \Delta_{1,N})W)u \\ &= (I_N - \rho_2 W)u - \Delta_{1,N} Wu \\ &= \varepsilon - \Delta_{1,N} Wu. \end{aligned} \tag{2.3.4}$$

Therefore, in a manner somewhat similar to the time series Cochrane–Orcutt (1946) approach, if (2.3.1) or its more stacked form in (2.3.2) is multiplied across by $(I_N - \hat{\rho}_2 W)$, we have

$$(I_N - \hat{\rho}_2 W)y = (I_N - \hat{\rho}_2 W)Z\beta + \varepsilon - \Delta_{1,N} Wu. \tag{2.3.5}$$

We will show that the transformed model in (2.3.5) accounts for spatial correlation in the sense that the term $\Delta_{1,N}Wu$ is asymptotically negligible in the IV estimation. However, since $Wy$ is one of the regressors in $Z$, there is still an endogenous regressor problem. The IV approach we take to estimate (2.3.5) is 2SLS. The instruments we use in our empirical illustrations below are the columns of $H$ in (2.2.3.4) with $r = 2$. This accounts for the endogeneity, as well as for the exogenous variables in the model.

Let

$$\begin{aligned} y(\hat{\rho}_2) &= (I_N - \hat{\rho}_2 W)y, \\ Z(\hat{\rho}_2) &= (I_N - \hat{\rho}_2 W)Z, \\ \hat{Z}(\hat{\rho}_2) &= H(H'H)^{-1}H'Z(\hat{\rho}_2) \end{aligned} \tag{2.3.6}$$

where we have indicated the dependence of the expressions on the estimator $\hat{\rho}_2$. Then our IV estimator of $\beta$ is

$$\hat{\beta}(\hat{\rho}_2) = [\hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2)]^{-1}\hat{Z}'(\hat{\rho}_2)y(\hat{\rho}_2). \tag{2.3.7}$$

In the appendix to this chapter we show

$$N^{1/2}[\hat{\beta}(\hat{\rho}_2) - \beta] \xrightarrow{D} N(0, \sigma_\varepsilon^2 p\lim N([\hat{Z}'(\rho_2)\hat{Z}(\rho_2)]^{-1}) \tag{2.3.8}$$

where

$$p\lim[N^{-1}\hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2)] = (Q'_{HZ} - \rho_2 Q'_{HWZ})Q_{HH}^{-1}(Q_{HZ} - \rho_2 Q_{HWZ}). \tag{2.3.9}$$

Following the procedure in Sections A.2 and A.12 of Appendix A, our small sample approximation to the distribution of $\hat{\beta}(\hat{\rho}_2)$ is

$$\hat{\beta}(\hat{\rho}_2) \simeq N(\beta, \hat{\sigma}_\varepsilon^2[\hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2)]^{-1}) \tag{2.3.10}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N-\delta}[y(\hat{\rho}_2) - Z(\rho_2)\hat{\beta}(\hat{\rho}_2)]'[y(\hat{\rho}_2) - Z(\rho_2)\hat{\beta}(\hat{\rho}_2)]$$

and $\delta$ can be taken to be zero or **any** finite constant because, asymptotically, $\frac{N}{N-\delta} \to 1$. Some researchers take $\delta$ as the number of regression parameters in the model, which in this case is $(2k + 2)$.

As an illustration, suppose one wanted to test hypothesis $H_0$ that $R\beta = q$, where $R$ is a known $\varphi \times 2k+2$ matrix with rank of $\varphi < 2k+2$, and $q$ is a known $\varphi \times 1$ vector. This test could be carried out in terms of $R\hat{\beta}(\hat{\rho}_2)$. For example, given $H_0$, the approximate small sample distribution of $R\hat{\beta}(\rho_2)$ suggested by (2.3.10) would be

$$R\hat{\beta}(\rho_2) \simeq N(q, R\,\hat{\sigma}_{\varepsilon}^2[\hat{Z}'(\rho_2)\hat{Z}(\rho_2)]^{-1}R'). \tag{2.3.11}$$

Hypothesis $H_0$ would be rejected at the 5% level if

$$[R\hat{\beta}(\rho_2) - q]'\left[R\,\hat{\sigma}_{\varepsilon}^2[\hat{Z}'(\rho_2)\hat{Z}(\rho_2)]^{-1}R'\right]^{-1}[R\hat{\beta}(\rho_2) - q] > \chi_{\varphi}^2(0.95). \tag{2.3.12}$$

### Illustration 2.3.1: A larger model of housing prices

We again consider the Boston dataset, but now we fit two models to the data. In the first model the error term is taken to be spatially correlated, $\rho_2 \neq 0$, and in the other model it is not, $\rho_2 = 0$.

In real estate markets, it is good practice for realtors always to look at comparable houses not too far from the property they are looking at with their clients. In our context, something like this would correspond to adding a spatial lag of the price variable based on neighboring houses.

The results from the 2SLS estimator for the model in which $\rho_2 = 0$ are reported below. Standard deviations are in the parentheses:

$$\begin{aligned}\widehat{\log(price)} = {} & 0.603(0.189) - 0.457(0.089)\log(nox) - 0.145(0.030)\log(dist) \\ & - 0.021(0.004)\,stratio + 0.181(0.014)\,rooms \\ & - 0.008(0.001)\,crime + 0.526(0.053)\,w\log(price).\end{aligned}$$

All the variables have the expected sign and are strongly significant. The significance of the estimator of $\rho_1$ suggests that housing prices indeed partially depend on neighboring house prices. There are interesting spillover effects associated with this coefficient, which will be discussed in Chapter 3.

The results for the model in which $\rho_2 \neq 0$ are

$$\begin{aligned}\widehat{\log(price)} = {} & 0.571(0.203) - 0.448(0.098)\log(nox) - 0.140(0.034)\log(dist) \\ & - 0.022(0.005)\,stratio + 0.185(0.014)\,rooms \\ & - 0.007(0.001)\,crime + 0.532(0.055)\,w\log(price);\;\; \hat{\rho}_2 = 0.198.\end{aligned}$$

In this case the results are quite similar. Specifically, all of the signs of the estimated coefficients are the same, their magnitudes are similar, and they are all significant. This similarity is not always the case!

### Illustration 2.3.2: A larger model of DUI arrests

In Illustration 2.2.1.3 we used the simulated US driving under the influence data from Drukker et al. (2013c). In the present example we will consider a variation of that model with the same data set but we estimate the full model. In this model we assume spatial correlation of the error term, and consider the spatial lag of the dependent variable, but do not consider the spatial lag of the police variable. The results are reported below:

$$\begin{aligned}\widehat{dui} &= -6.410(0.418) + 0.598(0.015)\, police + 0.000(0.001)\, nondui \\ &\quad + 0.016(0.000)\, venichles + 0.106(0.035)\, dry + 0.047(0.017)\, wdui.\end{aligned}$$

The evidence is consistent with our earlier example described in Illustration 2.2.1.3. The explanatory variables, with the exception of *nondui*, are all significant. The results also highlight that the coefficient relating to the variable *wdui* is positive and statistically significant. Finally, the estimated value of $\rho_2 = 0.0009$. This suggests that the true value of $\rho_2$ may be zero, and hence the absence of spatial correlation. Formal tests for spatial correlation are given in Chapter 11.

## 2.4 MAXIMUM LIKELIHOOD ESTIMATION OF THE GENERAL MODEL

Consider again the general model in (2.2.3.1), or (2.2.1),

$$\begin{aligned} y &= ae_N + XB_1 + \rho_1 Wy + (WX)\,B_2 + u, \\ u &= \rho_2\,(Wu) + \varepsilon, \quad |\rho_1| < 1,\ |\rho_2| < 1, \end{aligned} \tag{2.4.1}$$

but now assume normality, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_N)$. Let $X_+ = (e_N, X, WX)$ and $\gamma' = (a, B_1', B_2')$. Then the model in (2.4.1) can be expressed as

$$y = X_+\gamma + \rho_1 Wy + u. \tag{2.4.2}$$

Again, assuming the inverses exist, the solution for $y$ in terms of $X_+$ and $\varepsilon$ is

$$y = (I_N - \rho_1 W)^{-1}\, X_+\gamma + (I_N - \rho_1 W)^{-1}\,(I_N - \rho_2 W)^{-1}\,\varepsilon. \tag{2.4.3}$$

It follows that

$$\begin{aligned} y &\sim N(\mu_y, \sigma_\varepsilon^2 V_y), \\ \mu_y &= (I_N - \rho_1 W)^{-1}\, X_+\gamma, \\ V_y &= G^{-1}G^{-1\prime}, \\ G &= (I_N - \rho_2 W)\,(I_N - \rho_1 W)\,. \end{aligned} \tag{2.4.4}$$

The likelihood function is therefore

$$L = \frac{e^{-\frac{1}{2\sigma_\varepsilon^2}[y-\mu_y]'G'G[y-\mu_y]}}{(2\pi)^{N/2}(\sigma_\varepsilon^2)^{N/2}|V_y|_+^{1/2}} \tag{2.4.5}$$

where $|V_y|_+^{1/2}$ is the positive square root of $|V_y|$. Since the determinant of the product of square matrices is equal to the product of the determinants, $|V_y|_+^{1/2} = [|G^{-1}||G^{-1}|]_+^{1/2} = |G|_+^{-1}$, given the likelihood in (2.4.5),[25] the log-likelihood is

$$\begin{aligned} &\ln(L) \\ &= -\frac{1}{2\sigma_\varepsilon^2}[y-\mu_y]'G'G[y-\mu_y] - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_\varepsilon^2) + \ln(|G|_+) \\ &= -\frac{1}{2\sigma_\varepsilon^2}[Gy-G\mu_y]'[Gy-G\mu_y] - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_\varepsilon^2) + \ln(|G|_+). \end{aligned} \tag{2.4.6}$$

The MLEs of $a, B_1, \rho_1, B_2$, and $\sigma_\varepsilon^2$ are obtained by maximizing the log-likelihood in (2.4.6). Again, one would base inferences on asymptotic normality of those estimators, and the small sample approximations as was done in Section 2.2.2.1. Recall that computational difficulties may be involved.

The form of the maximum likelihood estimators of $\gamma$ and $\sigma_\varepsilon^2$ have an intuitive appeal. For example, let $\hat{\Psi} = (\hat{a}, \hat{B}_1', \hat{\rho}_1, \hat{B}_2', \hat{\rho}_2, \hat{\sigma}_\varepsilon^2)'$ be the vector of maximum likelihood estimators. Note that $G(I_N - \rho_1 W)^{-1} = (I_N - \rho_2 W)$. Then the first order condition corresponding to $\gamma$ is

$$\frac{\partial \ln(L)}{\partial \gamma}|_{\hat{\Psi}} = 0,$$

or

$$\begin{aligned} X_+'(I_N - \hat{\rho}_2 W)'(I_N - \hat{\rho}_2 W)(I_N - \hat{\rho}_1 W)y = \\ X_+'(I_N - \hat{\rho}_2 W)'(I_N - \hat{\rho}_2 W)X_+'\hat{\gamma}. \end{aligned} \tag{2.4.7}$$

[25] We obtained the likelihood directly in terms of the joint density of $y$. Another way to determine the likelihood is to obtain the joint density of $\varepsilon$, say $f_\varepsilon(\varepsilon)$, and then the joint density of $y$, say $f_y(y)$, as

$$f_y(y) = f_\varepsilon(\varepsilon)\,|\frac{\partial \varepsilon}{\partial y}|_+$$

where $\varepsilon$ is replaced by its expression in $y$, which in our case is obtained by solving (2.4.3) for $\varepsilon$ in terms of $y$. Recalling that $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_N)$, and noting that $|\frac{\partial \varepsilon}{\partial y}|_+ = |G|_+$, we leave it to the reader to demonstrate that the result is identical to (2.4.5).

Let $X_+(\hat{\rho}_2) = (I_N - \hat{\rho}_2 W)X_+$ and $y(\hat{\rho}_2) = (I_N - \hat{\rho}_2 W)y$. Then, it follows from (2.4.7) that the maximum likelihood estimator of $\gamma$ is

$$\hat{\gamma} = [X_+(\hat{\rho}_2)'X_+(\hat{\rho}_2)]^{-1}X_+(\hat{\rho}_2)'(I_N - \hat{\rho}_2 W)(I_N - \hat{\rho}_1 W)y. \tag{2.4.8}$$

There is a straightforward interpretation of $\hat{\gamma}$. For example, the model in (2.4.2) implies that

$$(I_N - \rho_1 W)y = X_+\gamma + u. \tag{2.4.9}$$

The implication is that if $\rho_1$ were known, the model in (2.4.2) could be transformed so that the only regressors are those in the exogenous matrix $X_+$. If $\rho_2$ were also known, (2.4.9) could be multiplied across by $(I_N - \rho_2 W)$ to obtain

$$(I_N - \rho_2 W)(I_N - \rho_1 W)y = (I_N - \rho_2 W)X_+\gamma + \varepsilon. \tag{2.4.10}$$

The model in (2.4.10) does not have an endogeneity problem or a spatially correlated error term. Clearly, $\gamma$ would be estimated by least squares. Using evident notation,

$$\tilde{\gamma} = [X_+(\rho_2)'X_+(\rho_2)]^{-1}X_+(\rho_2)'(I_N - \rho_2 W)(I_N - \rho_1 W)y. \tag{2.4.11}$$

Comparing (2.4.8) to (2.4.11), the maximum likelihood estimator of $\gamma$ is just the feasible counterpart to $\tilde{\gamma}$ in (2.4.11).

In a similar light, the first order condition corresponding to $\sigma_\varepsilon^2$ is

$$\begin{aligned} &\frac{\partial \ln(L)}{\partial \sigma_\varepsilon^2}|_{\hat{\Psi}} = 0 \\ &\qquad = \frac{1}{2\hat{\sigma}_\varepsilon^4}[\hat{G}y - \hat{G}\hat{\mu}_y]'[\hat{G}y - \hat{G}\hat{\mu}_y] - \frac{N}{2\hat{\sigma}_\varepsilon^2} = 0 \end{aligned} \tag{2.4.12}$$

where $\hat{G} = \left(I_N - \hat{\rho}_2 W\right)\left(I_N - \hat{\rho}_1 W\right)$ and $\hat{\mu}_y = \left(I_N - \hat{\rho}_1 W\right)^{-1} X_+\hat{\gamma}$. The result in (2.4.12) implies

$$\hat{\sigma}_\varepsilon^2 = \frac{[\hat{G}(y - \hat{\mu}_y)]'[\hat{G}(y - \hat{\mu}_y)]}{N}. \tag{2.4.13}$$

For purposes of interpretation, note that if $\hat{\rho}_2$, $\hat{\rho}_1$, and $\hat{\gamma}$ were replaced by $\rho_2$, $\rho_1$, and $\gamma$, respectively, the MLE $\hat{\sigma}_\varepsilon^2$ would reduce to

$$\hat{\sigma}_\varepsilon^2 = \varepsilon'\varepsilon/N.$$

### Illustration 2.4.1: House values and crime in Boston: MLE

Let us consider the same model as in Illustration 2.3.1, but now consider the results obtained by ML. Those results are shown in the following equation:

$$\begin{aligned}\widehat{\log(price)} &= 0.449(0.202) - 0.412(0.094)\log(nox) - 0.138(0.031)\log(dist)\\ &\quad - 0.019(0.005)\,stratio + 0.178(0.016)\,rooms\\ &\quad - 0.007(0.001)\,crime + 0.578(0.061)\,w\log(price)\end{aligned}$$

with $\rho_2 = 0.098$ having standard error of 0.112. An interesting issue here is that the results obtained by ML are very close to those that we highlighted in Illustration 2.3.1. This is even more surprising if one considers that the sample size of the Boston data set is only 506.

## 2.5 AN IDENTIFICATION FALLACY

Consider the model

$$\begin{aligned} y &= X\beta + \rho_1 W y + u, \\ u &= \rho_2 M u + \varepsilon \end{aligned} \tag{2.5.1}$$

where $X$ is exogenous, $\varepsilon \sim N(0, \sigma^2 I_N)$, and $W$ and $M$ are two weighting matrices. Now consider a special case of this model in which $\beta = 0$ and $W = M$. Assuming both inverses exist, in this case

$$\begin{aligned} & y \sim N(0, \sigma^2\Omega), \\ & \Omega = (I_N - \rho_1 W)^{-1}(I_N - \rho_2 W)^{-1}(I_N - \rho_2 W')^{-1}(I_N - \rho_1 W')^{-1} \end{aligned} \tag{2.5.2}$$

and so

$$\begin{aligned} \Omega^{-1} &= (I_N - \rho_1 W')(I_N - \rho_2 W')(I_N - \rho_2 W)(I_N - \rho_1 W) \\ &= GG', \\ G &= [I_N - (\rho_1 + \rho_2)W' + \rho_1\rho_2 W'W']. \end{aligned} \tag{2.5.3}$$

It should be clear from (2.5.3) that the likelihood is perfectly symmetric in $\rho_1$ and $\rho_2$, and so these two parameters are not identified under the stated conditions. This is a known result in the literature. Note carefully what we have stated. If in (2.5.1) $\beta = 0$ and $W = M$, there is an identification problem concerning $\rho_1$ and $\rho_2$.

In practice it is typically assumed that in a model such as (2.5.1) $W = M$. However, models in which $\beta = 0$ are not typically considered in practice. Results given below will demonstrate that if in a model such as (2.5.1), $\beta \neq 0$ there is no identification problem concerning the parameters of the model even if $W = M$. This is important to note, and unfortunately has not been noted by all researchers. For instance, as we have previously noted, if $\rho_1 = 0$, the model in (2.5.1) is often referred to as the spatial error model; if in (2.5.1) $\rho_1 \neq 0$ but

$\rho_2 = 0$, the model is often referred to as the spatial lag model. There have been quite a number of studies in which researchers (still) tried to determine whether the true model is a spatial error model, or a spatial lag model because it has been assumed that the identification condition restricts the consideration of the general model in (2.5.1) in which neither $\rho_1$ nor $\rho_2$ are zero. This is unfortunate because the spatial patterns implied by the more general model in which $\rho_1 \neq 0$ and $\rho_2 \neq 0$ are "richer" in their correlation patterns than that implied by either the spatial error model or the spatial lag model.

## 2.6 TIME SERIES PROCEDURES DO NOT ALWAYS CARRY OVER

In this section we illustrate that certain time series procedures should not be used in a spatial framework. We consider this issue because some time series procedures have been suggested to estimate spatial models.

In order to avoid unnecessary complications, consider the simple model

$$\begin{aligned} y &= X\beta + u, \\ u &= \rho_2 W u + \varepsilon \end{aligned} \tag{2.6.1}$$

where $X$ is an $N \times k$ exogenous regressor matrix, and $\varepsilon \sim N(0, \sigma^2 I_N)$. Following a time series approach, replace $u$ by $(y - X\beta)$ in the second line in (2.6.1) to get

$$\begin{aligned} y &= X\beta + \rho_2 W(y - X\beta) + \varepsilon \\ &= X\beta + \rho_2 W y + W X\gamma + \varepsilon, \quad \gamma = -\beta\rho_2. \end{aligned} \tag{2.6.2}$$

If the information that $\gamma = -\beta\rho_2$ is not recognized, the model on the second line of (2.6.2) is overparameterized. Its apparent benefit is that its error term is not spatially correlated. Since the model contains the endogenous variable $Wy$, it cannot be consistently estimated by ordinary least squares. One might therefore attempt to estimate the model by an IV procedure, namely 2SLS.

Suppose $D$ is **any** $N \times s$, $s \geq 2k + 1$ nonstochastic instrument matrix, e.g., $D$ could be, among other things, $(X, WX, W^2X, W^3X, ..., W^rX)_{LI}$, where $r$ is any finite constant such that $s \geq 2k + 1$. The resulting 2SLS estimator of the model's $2k + 1$ parameters, namely $(\beta', \rho_2, \gamma')$, will **not** be consistent. The reason for this is that $E(Wy) = WX\beta$, and (2.6.2) already contains $WX$ as a regressor matrix; therefore there will be no "informative" instrument for $Wy$. For example, in (2.6.2), let $Z$ be the $2k + 1$ regressor matrix $Z = (X, Wy, WX)$. Then, under typical conditions and notation such as that in Section 2.2.3,

$$p \lim_{N \to \infty} N^{-1} D' Z = p \lim_{N \to \infty} N^{-1} D' Z \tag{2.6.3}$$

$$\begin{aligned} &= p \lim_{N\to\infty} N^{-1}(D'X, D'WX\beta, D'WX) \\ &\equiv \Psi. \end{aligned}$$

Since $D'WX\beta$ is linear in the columns of $D'WX$, the rank of $\Psi$ will be less than $2k+1$, which is the number of its columns. The implication of (2.6.3) is that the second stage regressor matrix will be singular in large samples and so the 2SLS estimator will not be consistent or defined in the large sample. To see this, let the second stage regressor matrix in the 2SLS procedure be

$$\hat{Z} = D(D'D)^{-1}D'Z.$$

Then, using evident notation,

$$\begin{aligned} p \lim_{N\to\infty} N^{-1}\hat{Z}'\hat{Z} &= p \lim_{N\to\infty} [N^{-1}Z'D]\,[N(D'D)^{-1}]\,[N^{-1}D'Z] \\ &= \Psi' Q_{DD}^{-1}\Psi \end{aligned} \tag{2.6.4}$$

and note that the rank of the $2k+1 \times 2k+1$ matrix $\Psi' Q_{DD}^{-1}\Psi$ will be less than $2k+1$.

One would think that (2.6.2) can be consistently estimated by nonlinear 2SLS, using the instruments $D$ described above. In this procedure the restriction $\gamma = -\beta\rho_2$ would be used and so the only parameters would be $\beta$ and $\rho_2$. Unfortunately, this procedure is also not consistent. The reason for this is similar to that given above. To see the issue involved rewrite (2.6.2) as

$$\begin{aligned} y_{N\times 1} &= P_{N\times 1} + \varepsilon_{N\times 1}, \\ P_{N\times 2k+1} &= X\beta + \rho_2 Wy - WX\rho_2\beta. \end{aligned} \tag{2.6.5}$$

Suppose we again use the instrument matrix $D$. Then, a condition given by Amemiya (1985, pages 110 and 246) for consistency of the nonlinear 2SLS estimator is that

$$p \lim_{N\to\infty} N^{-1}D'\left(\frac{\partial P}{\partial(\rho_2, \beta')}\right) = L \tag{2.6.6}$$

where $L$ has full column rank. Now

$$\begin{aligned} \frac{\partial P}{\partial(\rho_2, \beta')} &= \left[Wy - WX\beta,\ X - \rho_2 WX\right] \\ &= \left[W(y - X\beta),\ X - \rho_2 WX\right] \\ &= [Wu,\ X - \rho_2 WX]. \end{aligned} \tag{2.6.7}$$

The reader should be able to demonstrate that the first column of

$$p \lim_{N\to\infty} \left[N^{-1}D'\frac{\partial P}{\partial(\rho_2, \beta')}\right]$$

is a column of zeros if, as is typically assumed,

$$N^{-1}D'D \to Q_{DD}$$

where $Q_{DD}$ is a finite invertible matrix. It follows that Amemiya's condition will not hold.

On a simpler scale, to see that "something is wrong" note that

$$\begin{aligned} E[P] &= X\beta + \rho_2 WX\beta - \rho_2 WX\beta \\ &= X\beta \end{aligned} \tag{2.6.8}$$

only involves $K$ variables which can be used as instruments. However, the nonlinear model in (2.6.5) has $K+1$ parameters. It follows that, at least for the spatial error model, the considered times series procedure does not lead to consistent estimators.

## APPENDIX A2 PROOFS FOR CHAPTER 2

### Proof of (2.2.4.1)

Let the $(i, j)$th element of $S$ be $s_{ij}$. Note that

$$\frac{v'Sv}{N} = \frac{\sum_{i=1}^{N} v_i^2 s_{ii}}{N} + \frac{\sum_{i<j=2}^{N} v_i v_j \left[s_{ij} + s_{ji}\right]}{N}. \tag{A2.1}$$

Note that $E\left(v_i^2 v_r v_s\right) = 0$, unless $r = s$ which is ruled out in (A2.1). Therefore every term in the double sum in (A2.1) is uncorrelated with every squared term. Also all the squared terms are uncorrelated with each other, as are all of the cross-product terms since $E\left[v_i v_j v_r v_s\right] = 0$ unless $i = j$ and $r = s$, or $i = r$ and $j = s$, or $i = j = r = s$. All of these conditions are ruled out. Thus since $E\left(v_i v_j\right)^2 = E\left(v_i^2\right) E\left(v_j^2\right) = \sigma_v^4$, if $i \neq j$, we have from (A2.1)

$$\begin{aligned} Var\left(\frac{v'Sv}{N}\right) &= \frac{1}{N^2}\left[\sum_{i=1}^{N} s_{ii}^2 Var\left(v_i^2\right) + \sum_{i<j=2}^{N} \sigma_v^4 \left[s_{ij} + s_{ji}\right]^2\right] \\ &\leq \frac{1}{N^2}\left[\sum_{i=1}^{N} s_{ii}^2 h + \sigma_v^4 \sum_{i<j=2}^{N} \left[\left|s_{ij}\right| + \left|s_{ji}\right|\right]^2\right] \end{aligned} \tag{A2.2}$$

where $h = E\left(v_i^4\right) - \sigma_v^4$. Let $c_s$ be the bound on the row and column sums of the absolute values of the elements of $S$, and therefore also on the absolute values of the elements of $S$. It then follows from (A2.2) that

$$\begin{aligned}
Var\left(\frac{v'Sv}{N}\right) &\leq \frac{1}{N^2}\sum_{i=1}^{N} s_{ii}^2 h + \frac{\sigma_v^4}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[|s_{ij}|^2 + |s_{ji}|^2 + 2|s_{ij}s_{ji}|] \quad \text{(A2.3)}\\
&\leq \frac{c_s}{N^2}\sum_{i=1}^{N}|s_{ii}|h + \frac{\sigma_v^4 c_s}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[|s_{ij}| + |s_{ji}| + 2|s_{ij}|]\\
&\leq \frac{hc_s}{N^2}\sum_{i=1}^{N}|s_{ii}| + \frac{\sigma_v^4 c_s}{N^2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}[|s_{ij}| + |s_{ji}| + 2|s_{ij}|]\right)\\
&\leq \frac{hc_s}{N} + \frac{\sigma_v^4 c_s}{N^2}\sum_{i=1}^{N} 4c_s\\
&\leq \frac{hc_s}{N} + \frac{4\sigma_v^4 c_s^2}{N} \to 0.
\end{aligned}$$

Therefore $Var\left(\frac{v'Sv}{N}\right) \to 0$ as $N \to \infty$, and the result in (2.2.4.1) follows from Chebyshev's inequality.

### The Generalized Moments Estimator of $\rho_2$. Proof of Consistency

Our proof has two parts. First we prove consistency of the estimator of $\rho_2$ if the disturbance vector $u$ is observed. We then show that $u$ can be replaced in the generalized moments estimator for $\rho_2$ by $\hat{u}$ described in Section 2.2.4.

If $u$ were observed, the estimation of $\zeta$ would be based on (2.2.4.24) in the text. In this case both $y_*$ and $X_*$ are observed. Note that

$$\begin{aligned}
u &= (I_N - \rho_2 W)^{-1}\varepsilon, \quad \text{(A2.4)}\\
\bar{u} &= W(I_N - W)^{-1}\varepsilon,\\
\bar{\bar{u}} &= WW(I_N - \rho_2 W)^{-1}\varepsilon.
\end{aligned}$$

If $u$, $\bar{u}$, and $\bar{\bar{u}}$ are replaced in (2.2.4.23) by their expressions in (A2.4), $X_*$ turns out to be identical to $G$ in (2.2.4.5). Therefore, by part (b) of Assumption 2.6, $X_*$ is nonsingular for $N$ large enough. In addition, by part (b) of Assumption 2.6 and (2.2.4.1)–(2.2.4.3),

$$p\lim_{N\to\infty} X_* = G^*. \quad \text{(A2.5)}$$

Also taking probability limits across in (2.2.4.24) yields

$$\begin{aligned}
p\lim_{N\to\infty} y_* &= p\lim_{N\to\infty} X_*\zeta \quad \text{(A2.6)}\\
&= G^*\zeta.
\end{aligned}$$

Given this, and if, preliminarily, the disturbance vector $u$ were observed, the linear GMM of $\rho_2$, say $\tilde{\rho}_2$, would just be the second element of the least squares estimator of $\zeta$ based on (2.2.4.24), namely

$$\begin{aligned}\tilde{\zeta} &= (X_*' X_*)^{-1} X_*' y_* \\ &= X_*^{-1} y_*\end{aligned} \tag{A2.7}$$

since $X_*$ is a $3 \times 3$ matrix which is nonsingular.

Taking probability limits across (A2.7) and using (A2.5)–(A2.6) implies

$$p \lim_{N \to \infty} (\tilde{\zeta}) = \zeta. \tag{A2.8}$$

Thus, if the disturbance vector $u$ were observed, the linear estimator would be consistent

$$\tilde{\rho}_2 \xrightarrow{P} \zeta. \tag{A2.9}$$

Now consider the estimator of $\zeta$ based on $\hat{u}$ as given by the second element of $\hat{\zeta}$ in (2.2.4.26). Since the estimator of $\beta$ in (2.2.4.5) is consistent, it can be expressed as

$$\tilde{\beta} = \beta + \Delta_N, \quad \Delta_N \xrightarrow{P} 0. \tag{A2.10}$$

Given the model in (2.2.3.1), the estimator of $u$ is

$$\begin{aligned}\hat{u} &= y - Z(\beta + \Delta_N) \\ &= u - Z\Delta_N.\end{aligned} \tag{A2.11}$$

Recall (A2.4) and that the row and column sums of $W$ and $(I_N - \rho_2 W)^{-1}$ are uniformly bounded in absolute value. As suggested in the text, and should be clear from (2.2.4.23), with the exception of the constants in the third column of $X_*$, every element of $X_*$ and $y_*$ can be expressed as a quadratic of the form $\varepsilon' S \varepsilon / N$ where $S$ is an $N \times N$ matrix whose row and column sums are uniformly bounded in absolute value. Given (A2.11),

$$\frac{\hat{u}' S \hat{u}}{N} = \frac{u' S u}{N} - \frac{2\Delta_N' Z' S u}{N} + \frac{\Delta_N' Z' S Z \Delta_N}{N}. \tag{A2.12}$$

We will show that

$$\frac{\Delta_N' Z' S u}{N} \xrightarrow{P} 0, \quad \frac{\Delta_N' Z' S Z \Delta_N}{N} \xrightarrow{P} 0, \tag{A2.13}$$

and so

$$\frac{\hat{u}' S \hat{u}}{N} - \frac{u' S u}{N} \xrightarrow{P} 0. \tag{A2.14}$$

Given (A2.13),

$$\hat{y}_* - y_* \xrightarrow{P} 0, \quad \hat{X}_* - X_* \xrightarrow{P} 0, \tag{A2.15}$$

and so, by (A2.5) and (A2.6),

$$\hat{y}_* \xrightarrow{P} G^*\zeta, \quad \hat{X}_* \xrightarrow{P} G^*. \tag{A2.16}$$

The consistency of $\hat{\zeta}$ then follows from (A2.5), (A2.6), and (A2.16).

### The Limits of $\Delta_N' Z'Su/N$ and $\Delta_N' Z'SZ\Delta_N/N$

Consider first $\Delta_N' Z'Su/N$ and recall that $Z = (e_N; X; Wy; WX)$. Recalling from (A2.10) that $\Delta_N \xrightarrow{P} 0$, the result in (A2.13) relating to $\Delta_N' Z'Su/N$ will hold if $Z'Su/N$ is $0_P(1)$, i.e., if $Z'Su/N^{1+\delta} \xrightarrow{P} 0$, for all $\delta > 0$.

Again, let matrix $S$ denote any $N \times N$ matrix whose row and column sums are uniformly bounded in absolute value. This is the only property of $S$ that is assumed. As an example, given Assumption 2.5,

$$\begin{aligned} (I_N - \rho_2 W) &= S, \\ W(I_N - \rho_1 W)^{-1}(I_N - \rho_1 W') &= S, \text{ etc.} \end{aligned}$$

Given this notation, and the solution of the model for $y$ in (2.2.3.6) can be expressed as

$$\begin{aligned} y &= S[ae_N + XB_1 + (WX)B_2 + (I_N - \rho_2 W)^{-1}\varepsilon] \\ &= aSe_N + SXB_1 + SXB_2 + S\varepsilon. \end{aligned} \tag{A2.17}$$

Note that $Wy$ is also expressible as

$$Wy = aSe_N + SXB_1 + SXB_2 + S\varepsilon. \tag{A2.18}$$

It follows that

$$Z'Su/N = \begin{bmatrix} e_N' S\varepsilon/N \\ X'S\varepsilon/N \\ y'S\varepsilon/N \\ X'S\varepsilon/N \end{bmatrix}. \tag{A2.19}$$

Using Chebyshev's inequality, the reader should have no difficulty in showing that

$$p \lim_{N\to\infty} e_N' S\varepsilon/N = 0 \quad \text{and} \quad p \lim_{N\to\infty} X_N' S\varepsilon/N = 0. \tag{A2.20}$$

Consider the remaining term, namely $y'S\varepsilon/N$, in (A2.19). Given (A2.17) and (A2.20), $Z'Su/N$ is $0_P(1)$ if $\varepsilon'S\varepsilon/N$ is $0_P(1)$; see Section A.14 in the appendix on large sample theory. Given Assumptions 2.1–2.5, and (2.2.4.1)–(2.2.4.3)

$$[\varepsilon'S\varepsilon/N - \sigma_\varepsilon^2 Tr(S)/N] \xrightarrow{P} 0. \tag{A2.21}$$

Since the elements of $S$ are all uniformly bounded in absolute value, so is $Tr(S)/N$ and hence $\varepsilon'S\varepsilon/N = 0_P(1)$. The result in (A2.13) relating to $\Delta_N' Z'Su/N$ follows. Using similar manipulations, the result in (A2.13) that relates to $\Delta_N' Z'SZ\Delta_N/N$ is left as an exercise.

The consistency of the nonlinear GMM should be evident. It is based on the same model, namely (2.2.4.25), except that it imposes the true condition $\eta = \rho_2^2$. A formal, and tedious proof of consistency is given in Kelejian and Prucha (1999).

### Derivation of the Large Sample Result in (2.3.8)

Let $\hat{\rho}_2$ be any consistent estimator of $\rho_2$. Premultiplying the model in (2.2.3.1) by $(I_N - \hat{\rho}_2 W)$, we have

$$\begin{aligned} y(\hat{\rho}_2) &= Z(\hat{\rho}_2)\beta + (I_N - \hat{\rho}_2 W)u \\ &= Z(\hat{\rho}_2)\beta + \varepsilon - \Delta_N W u. \end{aligned} \tag{A2.22}$$

Since $H(H'H)^{-1}H'$ is idempotent, it should be clear that $\hat{Z}'(\hat{\rho}_2)Z(\hat{\rho}_2) = \hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2)$ in (2.3.7). Thus, substituting (A2.22) into (2.3.7), we have

$$\hat{\beta}(\hat{\rho}_2) = \beta + [\hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2)]^{-1}\hat{Z}'(\hat{\rho}_2)[\varepsilon - \Delta_N W u], \tag{A2.23}$$

or

$$N^{1/2}[\hat{\beta}(\hat{\rho}_2) - \beta] = [N^{-1}\hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2)]^{-1} N^{-1/2}\hat{Z}'(\hat{\rho}_2)[\varepsilon - \Delta_N W u]. \tag{A2.24}$$

We now consider each of the components in (A2.24) in turn. Recalling that $H(H'H)^{-1}H'$ is idempotent, we have from Assumption 2.9 that

$$\begin{aligned} &N^{-1}\hat{Z}'(\hat{\rho}_2)\hat{Z}(\hat{\rho}_2) \\ &= N^{-1}Z'(\hat{\rho}_2)H(H'H)^{-1}H'Z'(\hat{\rho}_2) \\ &= N^{-1}[Z'(I_N - \hat{\rho}_2 W')H][N(H'H)^{-1}][N^{-1}H'(I_N - \hat{\rho}_2 W)Z] \\ &\xrightarrow{P} (Q_{HZ}' - \rho_2 Q_{HWZ}')Q_{HH}^{-1}(Q_{HZ} - \rho_2 Q_{HWZ}). \end{aligned} \tag{A2.25}$$

Now consider the next term in (A2.24), namely $N^{-1/2}\hat{Z}'(\hat{\rho}_2)\varepsilon$. Given (2.3.6), the consistency of $\hat{\rho}_2$, and Assumption 2.9, we have

$$N^{-1/2}\hat{Z}'(\hat{\rho}_2)\varepsilon = N^{-1}[Z'(I_N - \hat{\rho}_2 W')H]\,[N(H'H)^{-1}]\,N^{-1/2}H'\varepsilon \tag{A2.26}$$

and so

$$N^{-1/2}\hat{Z}'(\hat{\rho}_2)\varepsilon - [Q'_{HZ} - \rho_2 Q'_{HWZ}]Q_{HH}^{-1}\,[N^{-1/2}H'\varepsilon] \xrightarrow{P} 0. \tag{A2.27}$$

By Assumptions 2.7 and 2.9, the elements of $H$ are uniformly bounded in absolute value and $p\lim_{N\to\infty} N^{-1}H'H = Q_{HH}$ which is nonsingular. Therefore by Assumption 2.1 and the central limit theorem in Section A.15 of the large sample theory appendix,

$$N^{-1/2}H'\varepsilon \xrightarrow{D} N(0, \sigma_\varepsilon^2 Q_{HH}). \tag{A2.28}$$

Thus, by (A2.27), (A2.28), and the continuous mapping result in (A.10.3) in the large sample theory appendix,

$$N^{-1/2}\hat{Z}'(\hat{\rho}_2)\varepsilon \xrightarrow{D} N(0, \sigma_\varepsilon^2[Q'_{HZ} - \rho_2 Q'_{HWZ}]Q_{HH}^{-1}[Q_{HZ} - \rho_2 Q_{HWZ}]). \tag{A2.29}$$

For the moment, assume the absence of the term in (A2.24) involving $\Delta_N$. In this case the results in (2.3.8) and (2.3.9) follow from (A2.24), (A2.25), (A2.29), and the continuous mapping result given in (A.10.3) in the large sample theory appendix.

We now show that the term involving $\Delta_N$ in (A2.24) can be ignored since its probability limit is zero. In (A2.24) let $F = N^{-1/2}\hat{Z}'(\hat{\rho}_2)\Delta_N Wu$. Again, given (2.3.6) and noting that $\Delta_N$ is a scalar,

$$F = \Delta_N[N^{-1}[Z'(I_N - \hat{\rho}_2 W')H]\,[N(H'H)^{-1}]\,N^{-1/2}H'Wu], \tag{A2.30}$$

or

$$F - \Delta_N[Q'_{HZ} - \rho_2 Q_{HWZ}]Q_{HH}^{-1}[N^{-1/2}H'Wu] \xrightarrow{P} 0.$$

Let $F_1 = N^{-1/2}H'Wu$. Then

$$\begin{aligned} E(F_1) &= 0, & \text{(A2.31)}\\ E(F_1F_1') &= \sigma_\varepsilon^2 N^{-1}H'W\Omega_u W'H \\ &= \sigma_\varepsilon^2 N^{-1}H'SH \end{aligned}$$

where $\Omega_u = (I_N - \rho_2 W)^{-1}\left(I_N - \rho_2 W'\right)^{-1}$, and $S$ is again a matrix whose row and column sums are uniformly bounded in absolute value. It then follows from (2.1.4.5) that the elements of the product $N^{-1}H'SH$ are $0(1)$, i.e., are uniformly bounded in absolute value. Given this, it follows from Chebyshev's inequality that $F_1 = 0_P(1)$ and so, since $\Delta_N \to 0$, we have $F \xrightarrow{P} 0$ because

$$\Delta_N[Q'_{HZ} - \rho_2 Q_{HWZ}]Q_{HH}^{-1}[N^{-1/2}H'Wu] \xrightarrow{P} 0. \tag{A2.32}$$

## SUGGESTED PROBLEMS

**1.** For a sample size $N$, consider the spatial moving average model for the error term

$$(P.1) \quad u = \varepsilon + \rho W \varepsilon, \ |\rho| < 1$$

where $\varepsilon$ has mean and VC matrix of 0 and $\sigma^2 I_N$, respectively.
Determine the VC matrix of $u$, and then determine equations for a GMM approach you would use to estimate $\rho$ and $\sigma^2$.

**2.** Consider the spatial model for the error term

$$(P.2) \quad u = \rho W u + \varepsilon + \lambda W \varepsilon$$

where $\varepsilon$ has mean and VC matrix of 0 and $\sigma^2 I_N$, respectively.
Determine equations which could be used in a GMM approach to estimate $\rho$, $\lambda$, and $\sigma^2$.

**3.** Consider the model

$$\begin{aligned}(P.3) \quad y &= X\beta + \lambda_1 W_1 y + \lambda_2 W_2 y + u,\\ u &= \rho M u + \varepsilon\end{aligned}$$

where $\varepsilon$ has mean and VC matrix of 0 and $\sigma^2 I_N$, respectively, and $W_1$, $W_2$, and $M$ are observed exogenous weighting matrices.

**(a)** Suggest an instrumental variable estimation procedure for this model which accounts for the endogeneity of $W_1 y$ and $W_2 y$, as well as for the spatially correlated error term.

**(b)** Let $\gamma = (\beta', \lambda_1, \lambda_2)$. If $\rho = 0$ in the above model $(P.3)$, describe the large sample distribution of your instrumental variable estimator given in part (a).

**4.** Assuming $\rho \neq 0$ in $(P.3)$, obtain the likelihood function, and then determine the first order conditions for $\beta$.

**5.** Consider the model

$$(P.4) \quad y = X\beta + \lambda_1 W_1 y + \lambda_2 W y + \varepsilon$$

where $W_1$ and $W_2$ are row normalized. Give a condition which is sufficient for the model $(P.4)$ to be solved for $y$ in terms of $X$ and $\varepsilon$.

**6.** Consider the model

$$\begin{aligned}(P.5) \quad y &= X\beta + \lambda W y + u,\\ u &= \varepsilon + \rho W \varepsilon,\end{aligned}$$

and assume the elements of $\varepsilon$ are i.i.d. with mean and variance of 0 and $\sigma^2$, respectively, as well as satisfying the remaining conditions in Section 2.2.4. Suggest an instrumental variable estimator and determine whether or not it is consistent.