©*INTERNATIONAL REGIONAL SCIENCE REVIEW* **20**, 1 & 2: 103–111 (1997)

ESTIMATION OF SPATIAL REGRESSION MODELS WITH AUTOREGRESSIVE ERRORS BY TWO-STAGE LEAST SQUARES PROCEDURES: A SERIOUS PROBLEM

HARRY H. KELEJIAN
Department of Economics, University of Maryland, College Park, MD 20742 USA
(kelejian@econ.umd.edu)

INGMAR R. PRUCHA
Department of Economics, University of Maryland, College Park, MD 20742 USA
(prucha@econ.umd.edu)

Time series regression models that have autoregressive errors are often estimated by two-stage procedures which are based on the Cochrane-Orcutt (1949) transformation. It seems natural to also attempt the estimation of spatial regression models whose error terms are autoregressive in terms of an analogous transformation. Various two-stage least squares procedures suggest themselves in this context, including an analog to Durbin's (1960) procedure. Indeed, these procedures are so suggestive and computationally convenient that they are quite "tempting." Unfortunately, however, as shown in this paper, these two-stage least squares procedures are generally, in a typical cross-sectional spatial context, not consistent and therefore should not be used.

INTRODUCTION

The spatial autoregressive model studied by Cliff and Ord (1973, 1981), which is a variant of the model considered by Whittle (1954), is widely used to describe the properties of the error terms in spatial regressions. As typically specified, the error terms of a spatial autoregressive model depend on two unknown parameters. One is an autoregressive parameter, say ρ, and the other is a variance, say σ^2. Interest often focuses on ρ as a measure of spatial dependence, and also because it is a component of the generalized least estimator of the regression parameters. However, consistent estimation of both ρ and σ^2 is important for making inferences based on the regression model.

Based on an analogy with the Cochrane-Orcutt (1949) transformation in a linear time series model with autocorrelated error terms, one might think that, in a spatial context, the parameter ρ can be estimated consistently by two-stage least squares (2SLS) procedures. In particular, one might consider the estimation of the parameter ρ by a procedure that is analogous to that suggested by Durbin (1960) for linear time series models, referred to in the spatial literature as the spatial Durbin procedure. Unfortunately, however, as shown below, under typical

Luc Anselin and Serge Rey provided helpful comments.

Received January 1997; revised April 1997.

assumptions these procedures are, in general, not consistent. This point is important, especially since these 2SLS procedures are computationally convenient and therefore their use is "tempting."

In this paper, the basic model is first specified, then results concerning the inconsistency of the 2SLS procedures are presented, and finally some concluding remarks are given in the last section. Technical details are relegated to the Appendix.

THE MODEL

In this section, the regression model is specified, along with its assumptions. Those assumptions are then discussed. The following concept will be needed for the discussion. Let a_{ij} denote the (i, j)-th element of an n by n matrix A. Then, the row and column sums of A are said to be uniformly bounded in absolute value if

$$\sum_{j=1}^{n} |a_{ij}| \leq c_a \quad \text{for all} \quad i = 1, \ldots, n;\ n \geq 1$$

$$\sum_{i=1}^{n} |a_{ij}| \leq c_a \quad \text{for all} \quad j = 1, \ldots, n;\ n \geq 1$$

where c_a is a finite constant.[1]

The model considered is

$$y = X\beta + \varepsilon, \tag{1}$$

$$\varepsilon = \rho W\varepsilon + u, \tag{2}$$

where y is the n by 1 vector of observations on the dependent variable, X is the n by k matrix of observations on k exogenous regressors, β is the k by 1 vector of regression parameters, ε is the n by 1 vector of regression disturbances, ρ is the scalar autoregressive parameter, W is an n by n weights matrix, and u is an n by 1 vector of innovation error terms.

Let u_i be the i-th element of u, let Z be an n by q, $q \geq k$, matrix of instruments, and let $P = (Z, W'Z)$. Then, assume the following:

ASSUMPTION 1: *The u_i's are i.i.d. with mean 0 and finite variance σ^2.*

ASSUMPTION 2: *The elements of the weights matrix W are known constants, and rank $(I - \rho W) = n$ for all $|\rho| < 1$.*

[1] It can be shown that if two matrices, say A and B, are conformable for multiplication and their row and column sums are uniformly bounded in absolute value, then the row and column sums of the product matrix AB are also uniformly bounded in absolute value (see, e.g., Kelejian and Prucha 1995). Of course, if the row or column sums of a matrix are uniformly bounded in absolute value, then this is also the case for each element.

ASSUMPTION 3: *The row and column sums of W and* $(I-\rho W)^{-1}(I-\rho W')^{-1}$ *are uniformly bounded in absolute value.*

ASSUMPTION 4: *The elements of the regressor matrix X are nonstochastic, and X has full column rank.*

ASSUMPTION 5: *The elements of the instrument matrix Z are nonstochastic and bounded in absolute value, and Z has full column rank.*

ASSUMPTION 6: $\lim_{n\to\infty} n^{-1}X'X = Q_x$ *and* $\lim_{n\to\infty} n^{-1}P'P = Q_p$ *where* Q_x *and* Q_p *are finite and nonsingular. Furthermore,* $\lim_{n\to\infty} n^{-1}Z'X$ *and* $\lim_{n\to\infty} n^{-1}Z'WX$ *are finite.*

Assumptions 1 and 2 imply that $\varepsilon = (I-\rho W)^{-1}u$ and furthermore that $E(\varepsilon\varepsilon') = \Omega_\varepsilon$, where

$$\Omega_\varepsilon = \sigma^2(I-\rho W)^{-1}(I-\rho W')^{-1}. \tag{3}$$

These two assumptions are typical in spatial autoregressive models unless special complications are considered[2] (e.g., Cliff and Ord 1981: 198–9). Assumption 3 is reasonable and should hold for most weights matrix specifications. For example, the row and column sums of W will be uniformly bounded if W becomes a sufficiently sparse matrix as $n \to \infty$. Another example where this condition is satisfied is the case in which the elements of W are row normalized and the maximum number of nonzero elements in any given column remains bounded as $n \to \infty$. Next observe from (3) that, except for the scale factor σ^2, $(I-\rho W)^{-1}(I-\rho W')^{-1}$ is the variance-covariance matrix of ε. The assumption that the row and column sums of this matrix are uniformly bounded therefore restricts the extent of correlations relating to the elements of ε. In particular, the assumption implies, as is easily seen, that there exists some finite constant, say c_ω, such that

$$n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\text{corr}\,(\varepsilon_i,\varepsilon_j)\right| \le c_\omega < \infty \qquad \text{for all } n \ge 1,$$

where corr $(\varepsilon_i, \varepsilon_j)$ denotes the correlation between ε_i and ε_j. Virtually all large sample analyses restrict the extent of correlations in some way (see, e.g., Amemiya 1985, Ch. 3, 4; Pötscher and Prucha 1997, Ch. 5, 6; Anselin and Kelejian 1997). Assumption 4 is a standard condition in the context of the general linear regression model. Essentially, Assumption 4 rules out perfect multicollinearity. Assumption 5 maintains that the instruments are nonstochastic. One interpretation of this assumption is that the instruments are exogenous variables, and that the analysis is conditional upon their realized values. Assumption 6 relates to second order sample moments and is similar to those typically made

[2]Among other things, these complications could relate to heteroskedasticity concerning the innovation error terms, more general patterns of spatial correlation, and parametric specifications of the weights matrix (see, e.g., Case 1991; Anselin 1990; Dubin 1988).

in large sample analyses involving instrumental variable estimators (e.g., Judge et al. 1985: 167–9).

TWO-STAGE LEAST SQUARES PROCEDURES

Applying the analog of a Cochrane-Orcutt (1949) transformation to (1) and (2) and rearranging terms in analogy to Durbin's (1960) approach yields

$$y = \rho Wy + (X - \rho WX)\beta + u\,, \tag{4}$$

which can also be written in an over-parameterized form as

$$y = \rho Wy + X\beta + WX\gamma + u \tag{5}$$

where the restriction $\gamma = -\rho\beta$ is not considered. Note that the model formulations (4) and (5) have been called the spatial Durbin model (see, e.g., Anselin 1988).[3] The model in (1) implies that $Wy = WX\beta + W\varepsilon$. It then follows from (2) and Assumptions 1 and 2 that

$$E(Wyu') = \sigma^2(I - \rho W)^{-1} \neq 0\,.$$

Therefore, as noted in Anselin (1988: 58), the spatially lagged regressor, Wy, is correlated with the error term, u. One implication of this is that the parameters of (5) cannot be consistently estimated by ordinary least squares, nor can the parameters of (4) be consistently estimated by nonlinear least squares.

In light of the correlation between Wy and u, one might think of estimating (4) by nonlinear 2SLS, or (5) by (linear) 2SLS. However, as will be demonstrated, these procedures are, in general, not consistent. For this discussion, it proves convenient to denote with $\theta = (\rho, \beta')'$ the stacked vector of the true model parameters in (4). Furthermore, let $\bar{\theta} = (\bar{\rho}, \bar{\beta}')'$ denote some arbitrary a priori permissible parameter vector (of corresponding dimensions). Rewrite (4) as

$$y = f(\theta) + u$$

with

$$f(\theta) = \rho Wy + (X - \rho WX)\beta\,. \tag{6}$$

The function $f(\theta)$ is often referred to as the response function. The nonlinear 2SLS estimator of $\theta = (\rho, \beta')'$, say $\hat{\theta} = (\hat{\rho}, \beta')'$, based on the instruments Z is now defined as the minimizer of

$$R_n(\bar{\theta}) = n^{-1}[y - f(\bar{\theta})]'Z(Z'Z)^{-1}Z'[y - f(\bar{\theta})]\,. \tag{7}$$

[3]These model formulations have also been considered by Burridge (1981) and Blommestein (1983) and have been referred to in the spatial literature as the spatial common factor model.

Amemiya (1985: 246) gives conditions under which the nonlinear 2SLS estimator is consistent. In terms of the model presented in this paper, one of Amemiya's conditions for the consistency of $\hat{\theta} = (\hat{\rho}, \beta')'$ is that the matrix

$$H = \operatorname*{plim}_{n\to\infty} n^{-1} Z' \left.\frac{\partial f(\bar{\theta})}{\partial \bar{\theta}}\right|_{\bar{\theta}=\theta} \tag{8}$$

has full column rank. For purposes of interpretation, if a model were linear in the parameters, then the derivative of the response function with respect to the parameters would be the regressor matrix, say S. In this case, H would then correspond to the probability limit of $n^{-1}Z'S$.[4]

From (6),

$$\begin{aligned} \left.\frac{\partial f(\bar{\theta})}{\partial \bar{\theta}}\right|_{\bar{\theta}=\theta} &= [W(y - X\beta), (X - \rho WX)] \\ &= [W\varepsilon, (X - \rho WX)] \,. \end{aligned} \tag{9}$$

Note that the expected value of the first column of the n by k+1 matrix in (9) is a vector of zero. Given this and the maintained assumptions, it is shown in the appendix that the first column of H is also a vector of zeroes. It follows that H does not have full column rank. The violation of Amemiya's rank condition implies that his proof of consistency does not apply to the nonlinear 2SLS estimator corresponding to (4). It also suggests that there may be a fundamental "identification problem" in the sense that the objective function $R_n(\bar{\theta}) = R_n(\bar{\rho}, \bar{\beta})$ becomes flat in the direction of $\bar{\rho}$ as n tends toward infinity. That is, it suggests that in the limit the minimum of $R_n(\bar{\rho}, \bar{\beta})$ is not associated with a unique value of $\bar{\rho}$. That this is indeed the case for $\bar{\beta} = \beta$ is now demonstrated.

The nonlinear 2SLS estimator can be viewed as a special case of an M-estimator. A basic condition maintained in the general literature on M-estimators is that the parameters be identifiably unique (see, e.g., Gallant and White 1988; Pötscher and Prucha 1991, 1997). For the problem at hand, this translates into the requirement that the limiting objective function

$$\bar{R}(\bar{\rho}, \bar{\beta}) = \operatorname*{plim}_{n\to\infty} R_n(\bar{\rho}, \bar{\beta})$$

has a unique minimum at the true parameter value, i.e., $\bar{R}(\bar{\rho}, \bar{\beta}) > \bar{R}(\rho, \beta)$ for all $(\bar{\rho}, \bar{\beta}) \neq (\rho, \beta)$. Now observe that, for any given value of $\bar{\rho}$,

---

[4] In somewhat more detail, consider for a moment the classical case of a linear model, say $y = f(\theta) + u$ with $f(\theta) = S\theta$, where S is the regressor matrix. In this case, the minimizer of (7), i.e., the 2SLS estimator, can be expressed explicitly (in terms of the usual formula) as $\hat{\theta} = [S'Z(Z'Z)^{-1}Z'S]^{-1}S'Z(Z'Z)^{-1}Z'y$. Furthermore, observe that in this case, $\partial f(\bar{\theta})/\partial\bar{\theta} = S$. Thus, in the linear case, Amemiya's condition reduces to the standard requirement that $\operatorname*{plim}_{n\to\infty} n^{-1}Z'S$ has full column rank.

$$\begin{aligned} y - f(\bar{\rho}, \beta) &= y - \bar{\rho} W y - (X - \bar{\rho} W X)\beta \\ &= (I - \bar{\rho} W)\varepsilon \\ &= (I - \bar{\rho} W)(I - \rho W)^{-1} u \ . \end{aligned} \tag{10}$$

Note that $E[y - f(\bar{\rho}, \beta)] = 0$, for all values of $\bar{\rho}$. Given this and the maintained assumptions, it is demonstrated in the appendix that

$$\bar{R}(\bar{\rho}, \beta) = \operatorname*{plim}_{n \to \infty} R_n(\bar{\rho}, \beta) = 0 \tag{11}$$

for all values of $\bar{\rho}$. That is, as conjectured above, the limiting objective function of the nonlinear 2SLS estimator is indeed flat in the direction of $\bar{\rho}$, and thus the identifiably uniqueness condition does not hold. Again, this indicates that, in general, the nonlinear 2SLS estimator $\hat{\theta} = (\hat{\rho}, \hat{\beta}')'$ will be inconsistent. This inconsistency is demonstrated in the appendix for a special case of the model in (4).

Completely analogous observations hold for the linear 2SLS estimator corresponding to the model in (5). While a formal demonstration of the inconsistency of the linear 2SLS estimator is not given, the result should be evident from the discussion above. For example, the model in (5) is an over-parameterization of the model in (4) and therefore contains less information. Since the parameters of (4) are not identifiably unique, it is intuitively clear that the parameters of (5) are also not identifiably unique. Finally, it should be evident that corresponding results also hold for cases in which the error terms of a linear regression model are spatially autoregressive of order $q > 1$. For example, Amemiya's condition corresponding to (8) would then not hold because the first q columns of the matrix involved would be columns of zeroes.

CONCLUSION

It has been shown in this paper that typically specified linear spatial regression models with spatially autoregressive errors cannot be consistently estimated by 2SLS procedures based on a Cochrane-Orcutt transformation of the model. In a sense, this is unfortunate because these procedures are computationally convenient and feasible, and hence their use is "tempting." Also noted here, as well as elsewhere in the literature, is that the parameters of the spatial Durbin form of the model, obtained from a Cochrane-Orcutt transformation, cannot be consistently estimated by OLS.

On a more constructive note, Kelejian and Prucha (1995) suggest a three-step procedure for estimating the parameters of the model in (1) and (2). Their procedure is computationally feasible even with large samples; for example, there are more than 3000 counties in the U.S. In Kelejian and Prucha's first stage, β is estimated by OLS from (1) and the residuals are obtained. These residuals are then used in their second stage to estimate ρ, as say $\hat{\rho}$, by a generalized moments technique. In their third stage, β is estimated by feasible generalized least squares based on the estimator $\hat{\rho}$. Kelejian and Prucha demonstrate that this

feasible estimator of β is asymptotically equivalent to the true generalized least squares estimator, which is based on ρ. They also demonstrate under an explicit set of assumptions that their feasible generalized least squares estimator is asymptotically normal. In doing this, they do not assume that the error terms are normally distributed.

The model in (1) and (2) is also frequently estimated by maximum likelihood assuming that the error terms are normally distributed. One reason for the importance of the three-step procedure in Kelejian and Prucha (1995) is that the maximum-likelihood estimator may not be computationally feasible in large samples unless the weights matrix satisfies special (simplifying) conditions such as sparseness, symmetry, and so on.

REFERENCES

Amemiya, T. 1985. *Advanced econometrics.* Cambridge: Harvard University Press.

Anselin, L. 1988. *Spatial econometrics: Methods and models.* Boston: Kluwer Academic Publishers.

Anselin, L. 1990. Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics* 20: 141–63.

Anselin, L., and H. H. Kelejian. 1997. Testing for spatial error autocorrelation in the presence of endogenous regressors. *International Regional Science Review* 20: 153–82.

Billingsley, P. 1979. *Probability and measure.* New York: Wiley.

Blommestein, H. 1983. Specification and estimation of spatial economic modelling. *Regional Science and Urban Economics* 13: 251–70.

Burridge, P. 1981. Testing for a common factor in a spatial autoregressive model. *Environment and Planning A* 13: 795–800.

Case, A. 1991. Spatial patterns in household demand. *Econometrica* 59: 953–66.

Cliff, A., and J. Ord. 1973. *Spatial autocorrelation.* London: Pion.

Cliff, A., and J. Ord. 1981. *Spatial process: Models and applications.* London: Pion.

Cochrane, D., and G. H. Orcutt. 1949. Application of least squares regressions to relationships containing autocorrelated error terms. *Journal of the American Statistical Association* 44: 32–61.

Dubin, R. 1988. Estimation of regression coefficients in the presence of spatially autocorrelated error terms. *Review of Economics and Statistics* 70: 466–74.

Durbin, J. 1960. Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society B* 22: 139–53.

Gallant, R., and H. White. 1988. *A unified theory of estimation and inference for nonlinear dynamic models.* Oxford: Oxford University Press.

Judge, G. G., W. E. Griffiths, R. C. Carter Hill, H. Lütkepohl, and T.-C. Lee. 1985. *The theory and practice of econometrics.* New York: Wiley.

Kelejian, H. H., and I. R. Prucha. 1995. *A generalized moments estimator for the autoregressive parameter in a spatial model.* College Park: University of Maryland, Department of Economics Working Paper 95-03.

Pötscher, B. M., and I. R. Prucha. 1991. Basic structure of asymptotic theory in dynamic nonlinear models, I: Consistency and approximation concepts. *Econometric Reviews* 10: 125–216.

Pötscher, B. M., and I. R. Prucha. 1997. *Dynamic nonlinear econometric models: Asymptotic theory.* New York: Springer Verlag.

Serfling, R. J. 1980. *Approximation theorems of mathematical statistics.* New York: Wiley.

Whittle, P. 1954. On stationary processes in the plane. *Biometrica* 41: 434–49.

APPENDIX

Proofs Relating to Probability Limits

To demonstrate a preliminary result, let Γ be an n by n nonstochastic matrix whose row and column sums are uniformly bounded in absolute value by, say, c_γ. Let $A = n^{-1}Z'\Gamma Z$; then the elements of A are bounded in absolute value for all $n > 1$.

To see this, let a_{ij} be the (i, j)-th element of A, and let c_z be the bound for the elements of Z, i.e., $|z_{ij}| \leq c_z$ (Assumption 5). Then

$$\begin{aligned} |a_{ij}| &\leq n^{-1}\left|\sum_{t=1}^{n}\sum_{s=1}^{n} z_{ti}\, z_{sj}\, \gamma_{ts}\right| \leq n^{-1}\sum_{t=1}^{n}\sum_{s=1}^{n} |z_{ti}||z_{sj}||\gamma_{ts}| \\ &\leq n^{-1}c_z^2 \sum_{t=1}^{n}\sum_{s=1}^{n} |\gamma_{ts}| \leq c_z^2\, c_\gamma . \end{aligned} \tag{A.1}$$

To show that the first column of H is indeed a column of zeroes, Chebyshev's inequality is employed. Let φ_n be the first column of H. Then, in light of (8) and (9), $\varphi_n = n^{-1}Z'W\varepsilon$. Note first that the mean vector of φ_n is $E(\varphi_n) = 0$; also note that its variance-covariance matrix is

$$E(\varphi_n\varphi_n') = n^{-2}Z'(W\Omega_\varepsilon W')Z,$$

where Ω_ε is defined in (3). Assumption 3 implies that row and column sums of W and Ω_ε are uniformly bounded in absolute value, and so therefore are the row and column sums of $W\Omega_\varepsilon W'$; see footnote 1. It then follows from the preliminary result relating to (A.1) that $E(\varphi_n\varphi_n') \to 0$ and hence, via Chebyshev's inequality, that $\operatorname{plim}\varphi_n = 0$.

Next, to prove that equation (11) holds, let $\psi_n = n^{-1}Z'[y - f(\bar\rho, \beta)]$. Then by (10), $\psi_n = n^{-1}Z'(I - \bar\rho W)(I - \rho W)^{-1}u$. It follows that $E(\psi_n) = 0$ and

$$E(\psi_n\psi_n') = \sigma^2 n^{-2}Z'\Gamma_* Z$$

where $\Gamma_* = (I - \bar\rho W)(I - \rho W)^{-1}(I - \rho W')^{-1}(I - \bar\rho W')$. By Assumption 3, the row and column sums of W, and hence of $(I - \bar\rho W)$, and those of $(I - \rho W)^{-1}(I - \rho W')^{-1}$ are bounded uniformly in absolute value. It then follows that the row and column sums of Γ_* are also bounded uniformly in absolute value. Thus, again from the preliminary result relating to (A.1), $E(\psi_n\psi_n') \to 0$, and hence by Chebyshev's inequality, $\operatorname{plim}\psi_n = 0$. Observing that

$$R_n(\bar\rho, \beta) = n^{-1}[y - f(\bar\rho, \beta)]'Z(Z'Z)^{-1}Z'[y - f(\bar\rho, \beta)],$$

the result in (11) now follows trivially from this and Assumption 6.

Inconsistency Results for a Special Case

Consider the special case of (4) in which $\rho = 0$ and β is known. In this case, if Z is a vector, the 2SLS estimator $\hat{\rho}$ would be

$$\hat{\rho} = [Z'W(y - X\beta)]^{-1}Z'(y - X\beta)$$

$$= \frac{n^{-1/2}Z'u}{n^{-1/2}Z'Wu}.$$

Now define $\xi_n = (\xi_{n1}, \xi_{n2})'$, with $\xi_{n1} = n^{-1/2}Z'u$ and $\xi_{n2} = n^{-1/2}Z'Wu$. Recalling that $P = (Z, W'Z)$, you have $\xi_n = n^{-1/2}P'u$. Given the assumptions maintained in this paper, and the central limit theorem for triangular arrays presented in Kelejian and Prucha (1995, Theorem A),[5] it follows that

$$\xi_n \xrightarrow{D} N(0, \sigma^2 Q_p).$$

Using the continuous mapping theorem (see, e.g., Serfling 1980: 24), it then follows that

$$\hat{\rho} = \frac{\xi_{n1}}{\xi_{n2}} \xrightarrow{D} \frac{\xi_1}{\xi_2}$$

where $\xi = (\xi_1, \xi_2)' \sim N(0, \sigma^2 Q_p)$. Recall that Q_p is nonsingular by Assumption 6, and hence ξ_1 and ξ_2 are not perfectly correlated. Consequently, $\Pr(|\xi_1/\xi_2| > \delta) > 0$ for $\delta > 0$ and thus $\operatorname{plim}_{n\to\infty} \hat{\rho} \neq 0$.

[5]This central limit theorem follows readily from a corollary to the Lindeberg-Feller central limit theorem for triangular arrays. The corollary itself is given in, e.g., Billingsley (1979: 319, Problem 27.6). The need for a central limit theorem for triangular arrays arises because, in general, the elements of $Z'W$ depend on the sample size.