Regional Science and Urban Economics 55 (2015) 68–79

Contents lists available at ScienceDirect

Regional Science and Urban Economics

journal homepage: www.elsevier.com/locate/regec

GMM estimation of SAR models with endogenous regressors☆

Xiaodong Liu [a,*], Paulo Saraiva [b]

[a] Department of Economics, University of Colorado Boulder, United States
[b] Econ One Research, United States

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form 24 August 2015
Accepted 19 September 2015
Available online 30 September 2015

JEL classification:
C31
C36
R15

Keywords:
Spatial models
Endogeneity
Simultaneous equations
Moment conditions
Efficiency

ABSTRACT

In this paper, we extend the GMM estimator in Lee (2007) to estimate SAR models with endogenous regressors. We propose a new set of quadratic moment conditions exploiting the correlation of the spatially lagged dependent variable with the disturbance term of the main regression equation and with the endogenous regressor. The proposed GMM estimator is more efficient than IV-based linear estimators in the literature, and computationally simpler than the ML estimator. With carefully constructed quadratic moment equations, the GMM estimator can be asymptotically as efficient as the ML estimator under normality. Monte Carlo experiments show that the proposed GMM estimator performs well in finite samples.

1. Introduction

In recent years, spatial econometric models play a vital role in empirical research on regional and urban economics. By expanding the notion of space from geographic space to "economic" space and "social" space, these models can be used to study cross-sectional interactions in much wider applications including education (e.g. Lin, 2010; Sacerdote, 2011; Carrell et al., 2013), crime (e.g. Patacchini and Zenou, 2012; Lindquist and Zenou, 2014), industrial organization (e.g. König et al., 2014), finance (e.g. Denbee et al., 2014), etc.

Among spatial econometric models, the spatial autoregressive (SAR) model introduced by Cliff and Ord (1973, 1981) has received the most attention. In this model, the cross-sectional dependence is modeled as the weighted average outcome of neighboring units, typically referred to as the spatially lagged dependent variable. As the spatially lagged dependent variable is endogenous, likelihood- and moment-based methods have been proposed to estimate the SAR model (e.g. Kelejian and Prucha, 1998; Lee, 2004, 2007; Lee and Liu, 2010). In particular, for the SAR model with exogenous regressors, Lee (2007) proposes a generalized method of moments (GMM) estimator that combines linear moment conditions, with the (estimated) mean of the spatially lagged dependent variable as the instrumental variable (IV), and quadratic moment conditions based on the covariance structure of the spatially lagged dependent variable and the model disturbance term. The GMM estimator improves estimation efficiency of IV-based linear estimators in Kelejian and Prucha (1998) and is computationally simple relative to the maximum likelihood (ML) estimator in Lee (2004). Furthermore, Lin and Lee (2010) show that a sub-class of the GMM estimators is consistent in the presence of an unknown form of heteroskedasticity in model disturbances, and thus more robust relative to the ML estimator.

For SAR models with endogenous regressors, Liu (2012) and Liu and Lee (2013) consider, respectively, the limited information maximum likelihood (LIML) and two stage least squares (2SLS) estimators, in the presence of many potential IVs. Liu and Lee (2013) also propose a criterion based on the approximate mean square error of the 2SLS estimator to select the optimal set of IVs. The SAR model with endogenous regressors can be considered as an equation in a system of simultaneous equations. For the full information estimation of the system, Kelejian and Prucha (2004) propose a three stage least squares (3SLS) estimator and, in a recent paper, Yang and Lee (2014) consider the quasi-maximum likelihood (QML) approach. The QML estimator is asymptotically more efficient than the 3SLS estimator under normality but can be computationally difficult to implement. The existing estimators for the SAR model with endogenous regressors are summarized in Table 1.

In this paper, we extend the GMM estimator in Lee (2007) to estimate SAR models with endogenous regressors. We propose a new

☆ We would like to thank the Co-Editor and an anonymous referee for their helpful comments and suggestions. The remaining errors are our own.

* Corresponding author at. Department of Economics, University of Colorado Boulder, UCB 256, Boulder, Colorado, 80309, United States.
E-mail addresses: xiaodong.liu@colorado.edu (X. Liu), psaraiva@econone.com (P. Saraiva).

http://dx.doi.org/10.1016/j.regsciurbeco.2015.09.002
0166-0462/

Table 1
Existing estimators for SAR models with endogenous regressors.

	Single-equation estimator	System estimator
IV-based linear estimator	Liu and Lee (2013)	Kelejian and Prucha (2004)
likelihood-based estimator	Liu (2012)	Yang and Lee (2014)

set of quadratic moment equations exploiting (i) the covariance structure of the spatially lagged dependent variable and the disturbance term of the main regression equation and (ii) the covariance structure of the spatially lagged dependent variable and the endogenous regressor. We establish the identification, consistency and asymptotic normality of the proposed GMM estimator. The GMM estimator is more efficient than the 2SLS and 3SLS estimators, and computationally simpler than the ML estimator. With carefully constructed quadratic moment equations, the GMM estimator can be asymptotically as efficient as the ML estimator under normality. We also conduct a limited Monte Carlo experiment to show that the proposed GMM estimator performs well in finite samples.

The rest of the paper is organized as follows. In Section 2, we introduce the SAR model with endogenous regressors. In Section 3, we define the GMM estimator and discuss the identification of model parameters. In Section 4, we study the asymptotic properties of the GMM estimator and discuss the optimal moment conditions to use. Section 5 reports Monte Carlo experiment results. Section 6 briefly concludes. The proofs are collected in the appendix.

Throughout the paper, we adopt the following notation. For an $n \times n$ matrix $\mathbf{A} = [a_{ij}]_{i,j=1,\cdots,n}$, let $\mathbf{A}^{(s)} = \mathbf{A} + \mathbf{A}'$, $\text{vec}_D(\mathbf{A}) = (a_{11}, \cdots, a_{nn})'$, and $\text{diag}(\mathbf{A}) = \text{diag}(a_{11}, \cdots, a_{nn})$. The row (or column) sums of $\mathbf{A}$ are uniformly bounded in absolute value if $\max_{i=1,\cdots,n}\sum_{j=1}^{n}|a_{ij}|$ (or $\max_{j=1,\cdots,n}\sum_{i=1}^{n}|a_{ij}|$) is bounded.

2. Model

Consider a SAR model with an endogenous regressor[1] given by

$$\mathbf{y}_1 = \lambda_0\mathbf{W}\mathbf{y}_1 + \phi_0\mathbf{y}_2 + \mathbf{X}_1\beta_0 + \mathbf{u}_1, \tag{1}$$

where $\mathbf{y}_1$ is an $n \times 1$ vector of observations on the dependent variable, $\mathbf{W}$ is an $n \times n$ nonstochastic spatial weights matrix with a zero diagonal, $\mathbf{y}_2$ is an $n \times 1$ vector of observations on an endogenous regressor, $\mathbf{X}_1$ is an $n \times K_1$ matrix of observations on K_1 nonstochastic exogenous regressors, and $\mathbf{u}_1$ is an $n \times 1$ vector of i.i.d. innovations.[2] $\mathbf{W}\mathbf{y}_1$ is usually referred to as the spatially lagged dependent variable. Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_2$ is an $n \times K_2$ matrix of observations on K_2 excluded nonstochastic exogenous variables. The reduced form of the endogenous regressor $\mathbf{y}_2$ is assumed to be

$$\mathbf{y}_2 = \mathbf{X}\gamma_0 + \mathbf{u}_2, \tag{2}$$

where $\mathbf{u}_2$ is an $n \times 1$ vector of i.i.d. innovations. Let $\theta_0 = (\delta_0{}', \gamma_0{}')'$, with $\delta_0 = (\lambda_0, \phi_0, \beta_0{}')'$, denote the vector of true parameter values in the data generating process (DGP). The following regularity conditions are common in the literature of SAR models (see, e.g., Lee, 2007; Kelejian and Prucha, 2010).

Assumption 1. *Let $u_{1,i}$ and $u_{2,i}$ denote, respectively, the i-th elements of* $\mathbf{u}_1$ *and* $\mathbf{u}_2$.

(i) $(u_{1,i}, u_{2,i})' \sim i.i.d.\ (0, \mathbf{\Sigma})$, *where*

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

(ii) $\text{E}|u_{k,i}u_{l,i}u_{r,i}u_{s,i}|^{1+\eta}$ *is bounded for* $k, l, r, s = 1, 2$ *and some small constant* $\eta > 0$.

Assumption 2. (i) *The elements of* $\mathbf{X}$ *are uniformly bounded constants.* (ii) $\mathbf{X}$ *has full column rank* $K_X = K_1 + K_2$. (iii) $\lim_{n\to\infty} n^{-1}\mathbf{X}'\mathbf{X}$ *exists and is nonsingular.*

Assumption 3. (i) *All diagonal elements of the spatial weights matrix* $\mathbf{W}$ *are zero.* (ii) $\lambda_0 \in (-\underline{\lambda}, \bar{\lambda})$ *with* $0 < \underline{\lambda}, \bar{\lambda} \le c_\lambda < \infty$. (iii) $\mathbf{S}(\lambda) = \mathbf{I}_n - \lambda\mathbf{W}$ *is nonsingular for all* $\lambda \in (-\underline{\lambda}, \bar{\lambda})$. (iv) *The row and column sums of* $\mathbf{W}$ *and* $\mathbf{S}(\lambda_0)^{-1}$ *are uniformly bounded in absolute value.*

Assumption 4. θ_0 *is in the interior of a compact parameter space* Θ.

3. GMM estimation

3.1. Estimator

Let $\mathbf{S} = \mathbf{S}(\lambda_0) = \mathbf{I}_n - \lambda_0\mathbf{W}$ and $\mathbf{G} = \mathbf{W}\mathbf{S}^{-1}$. Under Assumption 3, model (1) has a reduced form

$$\mathbf{y}_1 = \mathbf{S}^{-1}\mathbf{X}_1\beta_0 + \phi_0\mathbf{S}^{-1}\mathbf{X}\gamma_0 + \mathbf{S}^{-1}\mathbf{u}_1 + \phi_0\mathbf{S}^{-1}\mathbf{u}_2, \tag{3}$$

which implies that

$$\mathbf{W}\mathbf{y}_1 = \mathbf{G}\mathbf{X}_1\beta_0 + \phi_0\mathbf{G}\mathbf{X}\gamma_0 + \mathbf{G}\mathbf{u}_1 + \phi_0\mathbf{G}\mathbf{u}_2. \tag{4}$$

As $\mathbf{W}\mathbf{y}_1$ and $\mathbf{y}_2$ are endogenous, consistent estimation of Eq. (1) requires IVs for $\mathbf{W}\mathbf{y}_1$ and $\mathbf{y}_2$. From Eq. (4), the deterministic part of $\mathbf{W}\mathbf{y}_1$ is a linear combination of the columns in $\mathbf{G}\mathbf{X} = [\mathbf{G}\mathbf{X}_1, \mathbf{G}\mathbf{X}_2]$. Therefore, $\mathbf{G}\mathbf{X}$ can be used as an IV matrix for $\mathbf{W}\mathbf{y}_1$.[3] From Eq. (2), $\mathbf{X}$ can be used as an IV matrix for $\mathbf{y}_2$. In general, let $\mathbf{Q}$ be an $n \times K_Q$ matrix of IVs such that $\text{E}(\mathbf{Q}'\mathbf{u}_1) = \text{E}(\mathbf{Q}'\mathbf{u}_2) = 0$. Let $\mathbf{u}_1(\delta) = \mathbf{S}(\lambda)\mathbf{y}_1 - \phi\mathbf{y}_2 - \mathbf{X}_1\beta$ and $\mathbf{u}_2(\gamma) = \mathbf{y}_2 - \mathbf{X}\gamma$, where $\delta = (\lambda, \phi, \beta')'$. The linear moment function for the GMM estimation is given by

$$\mathbf{g}_1(\theta) = (\mathbf{I}_2 \otimes \mathbf{Q})'\mathbf{u}(\theta),$$

where $\otimes$ denotes the Kronecker product, $\mathbf{u}(\theta) = [\mathbf{u}_1(\delta)', \mathbf{u}_2(\gamma)']'$, and $\theta = (\delta', \gamma')'$.[4]

Besides the linear moment functions, Lee (2007) proposes to use quadratic moment functions based on the covariance structure of the spatially lagged dependent variable and model disturbances to improve estimation efficiency. We generalize this idea to SAR models with endogenous regressors. Substitution of Eq. (2) into Eq. (1) leads to a "pseudo" reduced form

$$\mathbf{y}_1 = \lambda_0\mathbf{W}\mathbf{y}_1 + \phi_0\mathbf{X}\gamma_0 + \mathbf{X}_1\beta_0 + \mathbf{u}_1 + \phi_0\mathbf{u}_2. \tag{5}$$

By exploiting the covariance structure of the spatially lagged dependent variable $\mathbf{W}\mathbf{y}_1$ and the disturbances of Eq. (5), we propose the following quadratic moment functions

$$\mathbf{g}_2(\theta) = [\mathbf{g}_{2,11}(\delta)', \mathbf{g}_{2,12}(\theta)', \mathbf{g}_{2,21}(\theta)', \mathbf{g}_{2,22}(\gamma)']'$$

[1] In this paper, we focus on the model with a single endogenous regressor for exposition purpose. The model and proposed estimator can be easily generalized to accommodate any fixed number of endogenous regressors.

[2] $\mathbf{y}_1, \mathbf{y}_2, \mathbf{u}_1, \mathbf{u}_2, \mathbf{X}, \mathbf{W}$ are allowed to depend on the sample size n, i.e., to formulate triangular arrays as in Kelejian and Prucha (2010). Nevertheless, we suppress the subscript n to simplify the notation.

[3] The IV matrix $\mathbf{G}\mathbf{X}$ is not feasible as $\mathbf{G}$ involves the unknown parameter λ_0. Under Assumption 3, $\mathbf{G}\mathbf{X} = \mathbf{W}\mathbf{X} + \lambda_0\mathbf{W}^2\mathbf{X} + \lambda_0^2\mathbf{W}^3\mathbf{X} + \cdots$. Therefore, we can use the leading order terms $\mathbf{W}\mathbf{X}$, $\mathbf{W}^2\mathbf{X}$, $\mathbf{W}^3\mathbf{X}$ of the series expansion as feasible IVs for $\mathbf{W}\mathbf{y}$.

[4] In practice, we could use two different IV matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ to construct linear moment functions $\mathbf{Q}_1{}'\mathbf{u}_1(\delta)$ and $\mathbf{Q}_2{}'\mathbf{u}_2(\gamma)$. The GMM estimator with $\mathbf{g}_1(\theta)$ is (asymptotically) no less efficient than that with $\mathbf{Q}_1{}'\mathbf{u}_1(\delta)$ and $\mathbf{Q}_2{}'\mathbf{u}_2(\gamma)$ if $\mathbf{Q}$ includes all linearly independent columns of $\mathbf{Q}_1$ and $\mathbf{Q}_2$.