GMM estimation of the autoregressive parameter in a spatial autoregressive error model using regression residuals

Matthias Arnold *

Institut für Wirtschafts- und Sozialstatistik, Universität Dortmund, Dortmund, Germany

31st July 2007

Abstract

This paper suggests an improved GMM estimator for the autoregressive parameter of a spatial autoregressive error model by taking into account that unobservable regression disturbances are different from observable regression residuals. Although this difference decreases in large samples, it is important in small samples. Monte Carlo simulations show that the bias can be reduced by $65 - 80\%$ compared to a GMM estimator that neglects the difference between disturbances and residuals. The mean squared error is smaller, too.

Keywords. GMM estimation, spatial autoregression, regression residuals.

*Matthias Arnold, Institut für Wirtschafts- und Sozialstatistik, Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany, email: arnold@statistik.uni-dortmund.de, Fon: +49/231/7553869, Fax: +49/231/7555284. JEL subject classifications: C13, C21. The financial support by the Deutsche Forschungsgemeinschaft (SFB 475, Komplexitätsreduktion in multivariaten Datenstrukturen) is gratefully acknowledged.

1 Introduction and Summary

Disturbances of regression models are typically not observable, so inference on the disturbances must rely on the regression residuals. It is well known that under general conditions, the residuals converge to the disturbances when the sample size increases, see e.g. Rao and Toutenburg (1995). However, the statistical properties of the disturbances and the residuals are different in finite samples.

This paper considers a linear regression model where the disturbances are generated by a spatial autoregressive model introduced by Cliff and Ord (1973), and where the parameter of main interest is the spatial autoregressive parameter.

Since the calculation of the maximum likelihood estimator can be computationally expensive, Kelejian and Prucha (1999) suggest a generalized method of moments (GMM) estimator, which uses three theoretical moments of the disturbances and equates them to the corresponding empirical moments of the residuals. This estimator has been applied to data of industrial specialization by Tingvall (2004), to microlevel data by Bell and Bockstael (2000) and to agricultural data by Schlenker et al. (2006) and Anselin et al. (2004). It has also been extended in several ways, for example to panel data by Druska and Horrace (2004) and to systems of simultaneous equations by Kelejian and Prucha (2004).

We suggest a variation of the estimator that is motivated by the following argument: If the empirical moments must rely on the residuals, the theoretical moments should be calculated in terms of the residuals, too. We show that both estimators coincide as the sample size increases. The computational costs are of the same order of magnitude for both estimators. A small Monte Carlo study shows that the bias can be reduced by $65 - 80\%$. The mean squared error decreases, too.

In the following, we restrict ourselves to linear regression in order to keep notation as simple as possible. Nevertheless, the main idea also applies to nonlinear regression.

2 The Model and the Main Result

We consider a linear regression model

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

where y is the $(n \times 1)$-vector of observations on the dependent variable, X is the nonstochastic $(n \times k)$-matrix on the explanatory variables and β is the $(k \times 1)$-vector of unknown model parameters. We assume that u, the $(n \times 1)$-vector of disturbances, is generated by a spatial autoregressive model,

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

where W $(n \times n)$ is a so called weighting matrix of known constants, ρ is a scalar parameter and ε is an $(n \times 1)$-vector of innovations. We maintain the following assumptions.

Assumption 1 *(a) All diagonal elements of W are zero. (b) The row sums of W are equal to one, $\sum_{j=1}^{n} w_{ij} = 1 \ \forall \ i = 1, \ldots, n$. (c) $|\rho| < 1$.*

Assumption 2 *The innovations $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed with zero mean and variance σ^2, where the variance is bounded by some positive constant b, $0 < \sigma^2 < b < \infty$. Additionally, $E(\varepsilon_i^4) < \infty$.*

Assumption 1 ensures that the matrix $I - \rho W$ is nonsingular so that $u = (I - \rho W)^{-1}\varepsilon$. Thus,

$$\mathrm{Cov}(u) = \sigma^2 (I - \rho W)^{-1} (I - \rho W^T)^{-1}, \tag{3}$$

where W^T denotes the transpose of the matrix W.

Since u is not observable, estimation of ρ and σ^2 must rely on $\hat{u}$, the vector of regression residuals. For the case of OLS-regression, $\hat{u}$ is given by

$\hat{u} = y - X\hat{\beta} = Mu$, where $M = I - X(X^TX)^{-1}X^T$, and $\hat{\beta} = (X^TX)^{-1}X^Ty$ is the OLS-estimator of β.

In this situation, Kelejian and Prucha (1999) suggest a GMM estimator for ρ and σ^2 that uses three moments of ε, namely

$$\mathrm{E}\left(\frac{1}{n}\varepsilon^T\varepsilon\right) = \sigma^2, \;\; \mathrm{E}\left(\frac{1}{n}\varepsilon^TW^TW\varepsilon\right) = \frac{\sigma^2}{n}\mathrm{tr}(W^TW), \;\; \mathrm{E}\left(\frac{1}{n}\varepsilon^TW^T\varepsilon\right) = 0. \tag{4}$$

With the help of equation (2), the sample counterpart of (4) can be written as

$$G(\rho, \rho^2, \sigma^2)^T - g = v(\rho, \sigma^2),$$

where

$$G = \begin{pmatrix} \frac{2}{n}\hat{u}^TW\hat{u} & -\frac{1}{n}\hat{u}^TW^TW\hat{u} & 1 \\ \frac{2}{n}\hat{u}^TW^TWW\hat{u} & -\frac{1}{n}\hat{u}^TW^TW^TWW\hat{u} & \frac{1}{n}\mathrm{tr}(W^TW) \\ \frac{1}{n}\hat{u}^T[W + W^T]W\hat{u} & -\frac{1}{n}\hat{u}^TW^TWW\hat{u} & 0 \end{pmatrix}$$

and

$$g = \left(\frac{1}{n}\hat{u}^T\hat{u}, \; \frac{1}{n}\hat{u}^TW^TW\hat{u}, \; \frac{1}{n}\hat{u}^TW\hat{u}\right)^T.$$

The nonlinear least squares estimator of Kelejian and Prucha (1999) for ρ and σ^2 is defined as

$$(\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS}) = \mathrm{argmin}\{v(\rho, \sigma^2)^Tv(\rho, \sigma^2) : \rho \in [-a, a], \sigma^2 \in [0, b]\}, \tag{5}$$

where $a \geq 1$ and $b < \infty$.

Now the main idea is the following: If the unobservable disturbances u have to be replaced by the regression residuals $\hat{u}$, one should calculate the moment conditions (4) also in terms of $\hat{\varepsilon} = M\varepsilon = Mu - \rho MWu$ instead of ε. Consequently, we recommend an estimator that is based on the moments of $M\varepsilon$ corresponding to (4):

$$\begin{aligned} \mathrm{E}\left(\frac{1}{n}(M\varepsilon)^T M\varepsilon\right) &= \frac{\sigma^2}{n}\mathrm{tr}(M), & (6)\\ \mathrm{E}\left(\frac{1}{n}(WM\varepsilon)^T WM\varepsilon\right) &= \frac{\sigma^2}{n}\mathrm{tr}(MW^TW), & (7)\\ \mathrm{E}\left(\frac{1}{n}(WM\varepsilon)^T M\varepsilon\right) &= \frac{\sigma^2}{n}\mathrm{tr}(WM), & (8) \end{aligned}$$

where we made use of the fact that M is an orthogonal projector. If we multiply (2) by M and WM, respectively, we get

$$\begin{aligned} M\varepsilon &= Mu - \rho MWu, & (9)\\ WM\varepsilon &= WMu - \rho WMWu. & (10) \end{aligned}$$

Plugging equations (9) and (10) into the moment conditions (6)-(8) yields

$$\begin{aligned} \frac{1}{n}\mathrm{E}(u^TMu) - \frac{2\rho}{n}\mathrm{E}(u^TMWu) & \\ +\frac{\rho^2}{n}\mathrm{E}(u^TW^TMWu) &= \frac{\sigma^2}{n}\mathrm{tr}(M),\\ \frac{1}{n}\mathrm{E}(u^TMW^TWMu) - \frac{2\rho}{n}\mathrm{E}(u^TW^TWMWMu) & \\ +\frac{\rho^2}{n}\mathrm{E}(u^TW^TMW^TWMWu) &= \frac{\sigma^2}{n}\mathrm{tr}(MW^TW),\\ \frac{1}{n}\mathrm{E}(u^TMW^TMu) - \frac{\rho}{n}\mathrm{E}(u^TM[W+W^T]MWu) & \\ +\frac{\rho^2}{n}\mathrm{E}(u^TW^TMWMWu) &= \frac{\sigma^2}{n}\mathrm{tr}(WM). \end{aligned}$$

Finally, for every $(n \times n)$-matrix A, the theoretical moments $\mathrm{E}(u^TAu)$ are replaced by their sample counterparts $\hat{u}^TA\hat{u}$. Since $Mu = \hat{u}$ and $\mathrm{tr}(M) = \frac{n-k}{n}$, the sample analogon to (6) - (8) can be written as

$$H\left(\rho, \rho^2, \sigma^2\right)^T - h = w(\rho, \sigma^2),$$

where

$$H = \begin{pmatrix} \frac{2}{n}\hat{u}^TW\hat{u} & -\frac{1}{n}\hat{u}^TW^TMW\hat{u} & \frac{n-k}{n}\\ \frac{2}{n}\hat{u}^TW^TWMW\hat{u} & -\frac{1}{n}\hat{u}^TW^TMW^TWMW\hat{u} & \frac{1}{n}\mathrm{tr}(MW^TW)\\ \frac{1}{n}\hat{u}^T[W+W^T]MW\hat{u} & -\frac{1}{n}\hat{u}^TW^TMWMW\hat{u} & \frac{1}{n}\mathrm{tr}(WM) \end{pmatrix}$$

and $h = g$. Our nonlinear least squares estimator for ρ and σ^2 is defined as

$$(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS}) = \text{argmin}\{w(\rho,\sigma^2)^T w(\rho,\sigma^2) : \rho \in [-a,a], \sigma^2 \in [0,b]\}, \quad (11)$$

where $a \geq 1$ and $b < \infty$. We maintain the following typical assumptions for the regressor matrix X.

Assumption 3 *The elements of X are nonstochastic and bounded in absolute value by $0 < c_X < \infty$. Further, X has full column rank, and the matrix $Q_X = \lim_{n\to\infty} \frac{1}{n} X^T X$ is finite and nonsingular.*

Theorem 2.1 states the asymptotic equivalence of $(\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS})$ and $(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS})$.

Theorem 2.1 *Under assumptions 1-3, for $n \to \infty$*

$$(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS}) - (\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS}) \xrightarrow{P} 0.$$

Proof. Because of assumption 3, for large n the elements of $X(X^T X)^{-1} X^T$ are bounded by the corresponding elements of $\frac{kc_X^2}{n} Q_X^{-1} \xrightarrow{n\to\infty} 0$ so that $M = I - X(X^T X)^{-1} X^T \xrightarrow{n\to\infty} I$ and thus $H \xrightarrow{P} G$ as $n \to \infty$. Since $g = h$, $w(\rho,\sigma^2) \xrightarrow{P} v(\rho,\sigma^2)$, so that the minimization problems (11) and (5) coincide for $n \to \infty$ because $w(\rho,\sigma^2)$ and $v(\rho,\sigma^2)$ are continuous functions of ρ and σ^2. □

As a consequence of theorem 2.1, $(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS})$ shares the asymptotic properties of $(\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS})$ given in theorems 1 and 2 of Kelejian and Prucha (1999): $(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS})$ is a consistent estimator for (ρ, σ^2), the feasible GLS estimator $\hat{\beta}^{FG}$ is a consistent estimator for β and the asymptotic covariance matrix of $\hat{\beta}^{FG}$ can be estimated consistently, too.

The next section compares the finite sample performance of $(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS})$ and $(\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS})$ by way of Monte Carlo simulation.

3 Monte Carlo simulation

We compare the finite sample properties of $(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2^*}_{NLS})$ and $(\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS})$ for $n = 20, 100, 400$, $\rho = -0.5, 0.5$ and $\sigma^2 = 1$. The matrix W is specified such that each element of u_i is directly related to the three elements immediately after and immediately before it. For the first three and the last three elements of u, we imply a circular setting such that for example u_1 is directly related to $u_2, u_3, u_4, u_{n-3}, u_{n-2}$ and u_{n-1}. The row sums of W are standardized to one. Thus, in each row of W, six elements are equal to $\frac{1}{6}$ and the other elements are equal to zero. With respect to the regression model (1), we decided for

$$X = \begin{pmatrix} 1 & 1 & 1 \\ \vdots & \vdots & 0 \\ \vdots & \vdots & \vdots \\ \vdots & 1 & \vdots \\ \vdots & 0 & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

the model matrix of a regression on an intercept and two binary regressors. For each combination of n and ρ, we generated $m = 10,000$ realizations of the disturbance vector u corresponding to the spatial autoregressive model (2), where the components of ε are i.i.d. $N(0, 1)$. The left part of Table 1 shows the simulated biases $\frac{1}{m}\sum_{i=1}^{m}(\widehat{\rho} - \rho)$, variances $\frac{1}{m}\sum_{i=1}^{m}(\widehat{\rho} - \bar{\widehat{\rho}})^2$ and MSEs $\frac{1}{m}\sum_{i=1}^{m}(\widehat{\rho} - \rho)^2$ of both estimators for ρ. The right part of the table contains the corresponding numbers for σ^2.

For the combinations of n and ρ considered here, the bias of $\widehat{\rho}^*_{NLS}$ is $65-80\%$ smaller than the bias of $\widehat{\rho}_{NLS}$. The variance of $\widehat{\rho}^*_{NLS}$ is also smaller than the variance of $\widehat{\rho}_{NLS}$, but in contrast to the situation for the bias, this effect seems to vanish as n increases. As a consequence, the MSE can be

n	ρ		Bias	Var	MSE		Bias	Var	MSE
20	-0.5	$\widehat{\rho}_{NLS}$	-0.5791	0.7007	1.0361	$\widehat{\sigma}^2_{NLS}$	-0.2725	0.0814	0.1557
		$\widehat{\rho}^*_{NLS}$	-0.1239	0.5770	0.5923	$\widehat{\sigma}^{2*}_{NLS}$	-0.0898	0.1282	0.1363
20	0.5	$\widehat{\rho}_{NLS}$	-0.6620	0.5238	0.9621	$\widehat{\sigma}^2_{NLS}$	-0.2281	0.0839	0.1359
		$\widehat{\rho}^*_{NLS}$	-0.1615	0.5040	0.5306	$\widehat{\sigma}^{2*}_{NLS}$	-0.0745	0.1174	0.1229
100	-0.5	$\widehat{\rho}_{NLS}$	-0.1005	0.0522	0.0623	$\widehat{\sigma}^2_{NLS}$	-0.0588	0.0203	0.0237
		$\widehat{\rho}^*_{NLS}$	-0.0293	0.0506	0.0514	$\widehat{\sigma}^{2*}_{NLS}$	-0.0180	0.0218	0.0221
100	0.5	$\widehat{\rho}_{NLS}$	-0.0718	0.0203	0.0255	$\widehat{\sigma}^2_{NLS}$	-0.0324	0.0197	0.0207
		$\widehat{\rho}^*_{NLS}$	-0.0253	0.0186	0.0192	$\widehat{\sigma}^{2*}_{NLS}$	-0.0099	0.0206	0.0207
400	-0.5	$\widehat{\rho}_{NLS}$	-0.0251	0.0117	0.0123	$\widehat{\sigma}^2_{NLS}$	-0.0160	0.0052	0.0055
		$\widehat{\rho}^*_{NLS}$	-0.0075	0.0115	0.0116	$\widehat{\sigma}^{2*}_{NLS}$	-0.0057	0.0053	0.0053
400	0.5	$\widehat{\rho}_{NLS}$	-0.0155	0.0034	0.0036	$\widehat{\sigma}^2_{NLS}$	-0.0086	0.0052	0.0053
		$\widehat{\rho}^*_{NLS}$	-0.0054	0.0033	0.0033	$\widehat{\sigma}^{2*}_{NLS}$	-0.0033	0.0053	0.0053

Table 1: Bias, variance and MSE of the estimators

reduced by around 45% for $n = 20$, 20% for $n = 100$ and 5% for $n = 400$ if we use $\widehat{\rho}^*_{NLS}$ instead of $\widehat{\rho}_{NLS}$ to estimate ρ.

With respect to the estimators for σ^2, the right part of Table 1 shows the same positive effect for the bias. Contrary to that, the variance of $\widehat{\sigma}^{2*}_{NLS}$ is larger than the variance of $\widehat{\sigma}^2_{NLS}$ in small samples. Again, this effect vanishes as n increases. We conjecture that this difference in the variances is caused by the bottom right element of the matrices G and H: In contrast to H, this element is zero for G so that in the elements of the vector v, σ^2 appears only twice whereas in the elements of the vector w, σ^2 appears three times. This could be the reason for the different variances. Despite the larger variance, the MSE of $\widehat{\sigma}^{2*}_{NLS}$ is smaller than the MSE of $\widehat{\sigma}^2_{NLS}$. For $n = 20$, this gain is about 10%. For larger samples, the MSEs are almost identical for both estimators. We conclude that the drawback for the variance is overcompensated by the gain for the bias.

The simulation also revealed that the squared differences between the estimators converge to zero more quickly than the MSEs of the estimators.

For example, $\frac{1}{m}\sum_{i=1}^{m}(\widehat{\rho}^{*}_{NLS} - \widehat{\rho}_{NLS})^2 = 0.0001$ for $n = 400$ and $\rho = 0.5$ whereas the corresponding MSEs are 0.0033 and 0.0036, respectively. This result is in line with Theorem 2.1.

4 Discussion

The main idea of this paper is to take into consideration the fact that the behavior of observable regression residuals is different from the behavior of the corresponding unobservable disturbances. This idea is applied to a GMM estimator for the spatial autocorrelation parameter in a linear regression model, but it also applies in other situations.

We considered the linear regression model (1) to keep notation as simple as possible. However, the idea generalizes to situations where the regression model is nonlinear. In this case, the matrix X can be replaced by the matrix of the first partial derivatives of the y_i on the β_j evaluated in $\beta = \hat{\beta}_{NLS}$, the nonlinear least squares estimator for β. The reason for that is that the nonlinear least squares regression residuals are orthogonal to the columns of the matrix of the first partial derivatives.

Furthermore, it would be interesting to simulate the small sample properties of $(\widehat{\rho}^{*}_{NLS}, \widehat{\sigma}^{2*}_{NLS})$ for other model matrices X and for nonnormal distributions of ε. One could compare $(\widehat{\rho}^{*}_{NLS}, \widehat{\sigma}^{2*}_{NLS})$ to the quasi maximum likelihood estimator which maximizes the normal likelihood. It should be noted here that the improvement of the GMM estimator discussed in this paper does not easily carry over to the (quasi) maximum likelihood estimator. In contrast to the covariance matrix of the disturbances, the covariance matrix of the regression residuals is singular since M is singular. As a consequence, the multivariate normal density of u cannot easily be rewritten in terms of $\hat{u}$.

Finally, it would be interesting to investigate the properties of significance tests and confidence regions for the parameter vector β in (1). To perform such procedures, an estimator for the disturbance covariance matrix (3) is needed. One way to construct such an estimator is to just plug in the estimators for ρ and σ^2 in (3). Since $(\widehat{\rho}^*_{NLS}, \widehat{\sigma}^{2*}_{NLS})$ has a smaller MSE and a considerably smaller bias than $(\widehat{\rho}_{NLS}, \widehat{\sigma}^2_{NLS})$, we guess that the distortion of significance tests on β could be reduced by our new estimator.

References

L. Anselin, R. Bongiovanni, and J. Lowenberg-DeBoer. A spatial econometric approach to the economics of site-specific nitrogen management in corn production. *American Journal of Agricultural Economics*, 86(3):675–687, 2004.

K. P. Bell and N. E. Bockstael. Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *The Review of Economics and Statistics*, 82(1):72–82, 2000.

A.D. Cliff and J.K. Ord. *Spatial Autocorrelation.* Pion, London, 1973.

V. Druska and W. C. Horrace. Generalized moments estimation for spatial panel data: Indonesian rice farming. *American Journal of Agricultural Economics*, 86(1):185–198, 2004.

H.H. Kelejian and I.R. Prucha. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40:509–533, 1999.

H.H. Kelejian and I.R. Prucha. Estimation of simultaneous systems of spatially interrelated cross sectional equations. *Journal of Econometrics*, 118: 27–50, 2004.

C.R. Rao and H. Toutenburg. *Linear Models: Least Squares and Alternatives.* Springer, New York, 1995.

W. Schlenker, W. M. Hanemann, and A. C. Fisher. The impact of global warming on u.s. agriculture: An econometric analysis of optimal growing conditions. *Review of Economics and Statistics*, 88(1):113–125, 2006.

P. G. Tingvall. The dynamics of european industrial structure. *Review of World Economics*, 140(4):665–687, 2004.